Paper 1304-2014

Modeling Fractional Outcomes with SAS ®

WenSui Liu, Fifth Third Bancorp

Jason Xin, SAS

Abstract

For practitioners, OLS (Ordinary Least Squares) regression with Gaussian distributional assumption has been the top choice to model fractional outcomes in many business problems. However, it is conceptually flawed to assume Gaussian distribution for a response variable in the [0, 1] range. In this paper, several modeling techniques for fractional outcomes with their implementations in SAS should be discussed through a data analysis exercise in modeling financial leverage ratios of businesses. The purpose of this paper is to provide a relatively comprehensive survey of how to model fractional outcomes to the SAS user community and interested statistical practitioners.

Keywords

Fractional outcomes, Tobit model, NLS (Non-linear Least Squares) regression, Fractional Logit model, Beta regression, Simplex regression, Vuong statistic.

1. Introduction

In the financial service industry, we often observed business necessities to model fractional outcomes in the range of [0, 1]. For instance, in the context of credit risk, LGD (Loss Given Default) measures the proportion of losses not recovered from a default borrower during the collection process, which is observed in the closed interval [0, 1]. Another example is the corporate financial leverage ratio represented by the long-term debt as a proportion of the summation for both the long-term debt and the equity.

Although research interests in statistical models for fractional outcomes have remained strong in the past years, there is still no general agreement on either the distributional assumption or the modeling practice. An interesting but somewhat ironic reality is that the simple OLS regression with Gaussian distributional assumption has remained the most popular method to model fractional outcomes due to the simplicity. However, the approach with OLS regression suffers from a couple of conceptual flaws. First and the most evidential of all, fractional outcomes in the [0, 1] interval are not defined on the whole real line and therefore shouldn’t be considered normally distributed. Moreover, a distinctive statistical nature of fractional outcomes is that the variance is not independent of the mean. For instance, the variance shrinks as the mean approaches boundary points of [0, 1], which is nothing but a form of Heteroscedasticity.

In addition to the aforementioned naïve OLS regression, another class of transformed OLS regression based upon the logistic normal distribution is also overwhelmingly popular. In this approach, while boundary points at 0 or 1 can be handled heuristically, e.g. adding or subtracting a small value such that ***Y` = [Y × (n - 1) + / - 0.5] / n***, any value in the open interval (0, 1) would be transformed by the Logit function such that

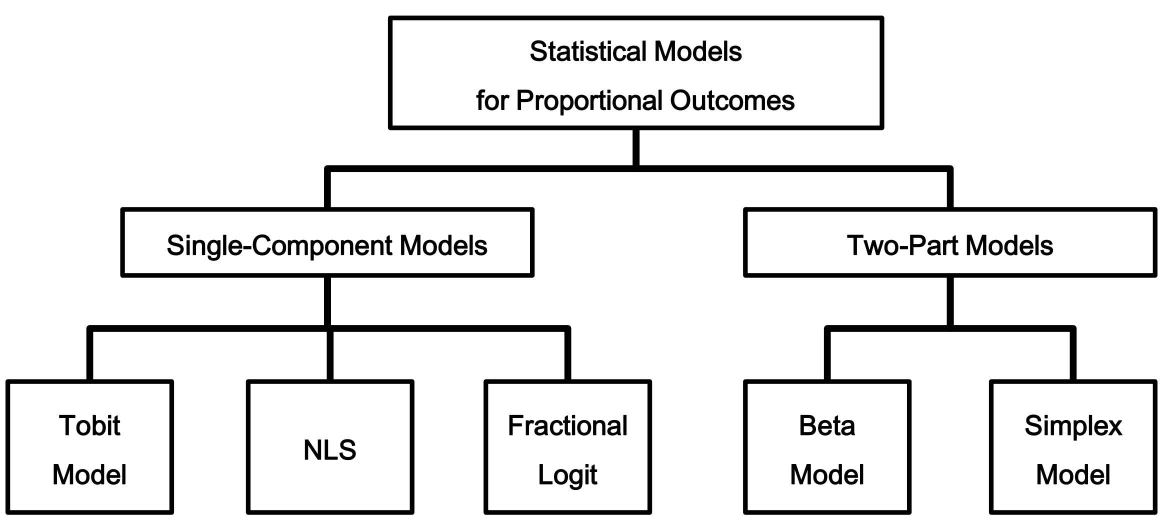
$$\boldsymbol{LOG}(\boldsymbol{Y}/(\mathbf{1}-\boldsymbol{Y})) = \boldsymbol{X}\grave{}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \textit{where the error term } \boldsymbol{\varepsilon} \sim \boldsymbol{Normal}(\mathbf{0}, \boldsymbol{\sigma}\hat{}\mathbf{2}\ )$$

After the Logit transformation, whilst ***Y*** is still strictly bounded by (0, 1), ***LOG (Y / (1 – Y))*** is however well defined on the whole real line. More attractively, most model development techniques and statistical diagnostics can be ported directly from the simple OLS regression with no or little adjustment.

Albeit simple, OLS regression with the Logit transformation is not free of either conceptual or practical difficulties. A major concern is that, in order to ensure ***LOG (Y / (1 – Y)) ~ Normal (X`β,*** $\boldsymbol{\sigma^2}$***)*** and therefore ***ε ~ Normal (0,*** $\boldsymbol{\sigma^2}$***)***, the variable ***Y*** must, by theory, follow the additive logistic normal distribution, which defers to statistical diagnostics. For instance, it is important to check whether the error term ***ε*** follows a standard normal distribution in the post-model diagnostics with Shapiro-Wilk or Jarque-Bera test. In addition, since the model response is ***LOG (Y / (1 – Y))*** instead of ***Y***, the interpretation on model results might not be straightforward. Extra efforts are often necessary to recover model effects on ***E (Y | X)*** from ***E (LOG (Y / (1 – Y)) | X)***.

Given limitations of OLS regressions discussed above, alternative approaches are called for. In the paper, five different modeling approaches, which can be loosely classified into two broad categories, are surveyed and discussed. The first category encompasses single-component modeling approaches that are able to generically handle fractional outcomes in the close interval of [0, 1], including Tobit, NLS (Nonlinear Least Squares), and Fractional Logit models. The second category covers two-part modeling approaches with one model, e.g. a Logit model, separating between boundary points and the open interval of (0, 1) and the other governing all values in the (0, 1) interval by a Beta or Simplex model. A schematic view of all five approaches is given below.

Figure 1.1, Schematic Diagram of Statistical Models for Fractional Outcomes



To better illustrate how to employ these models in the practice, we will show a use case of modeling the financial leverage ratio defined in the [0, 1) interval with the point mass at 0 implying zero debt in the corporate capital structure. All information, including both the response and predictors, is given in the table below.

Table 1.1, Data Description

Variables	Names	Descriptions
Y	Leverage ratio	ratio between long-term debt and the summation of long-term debt and equity
X1	Non-debt tax shields	ratio between depreciation and earnings before interest, taxes, and depreciation
X2	Collateral	sum of tangible assets and inventories, divided by total assets
X3	Size	natural logarithm of sales
X4	Profitability	ratio between earnings before interest and taxes and total assets
X5	Expected growth	percentage change in total assets
X6	Age	years since foundation
X7	Liquidity	sum of cash and marketable securities, divided by current assets

2. Data Analysis

First of all, a preliminary data analysis provides the summary statistics of all variables, as below.

Table 2.1, Summary Statistics for Full Sample

Full Sample = 4,421					
Variables	Min	Median	Max	Average	Variance
Leverage ratio	0.0000	0.0000	0.9984	0.0908	0.0376
Non-debt tax shields	0.0000	0.5666	102.1495	0.8245	8.3182
Collateral	0.0000	0.2876	0.9953	0.3174	0.0516
Size	7.7381	13.5396	18.5866	13.5109	2.8646
Profitability	0.0000	0.1203	1.5902	0.1446	0.0123
Expected growth	-81.2476	6.1643	681.3542	13.6196	1333.5500
Age	6.0000	17.0000	210.0000	20.3664	211.3824
Liquidity	0.0000	0.1085	1.0002	0.2028	0.0544

The median of the response variable is equal to 0, implying that the majority of values are point mass at 0. Given the sighting, it might be helpful to take a second look without boundary points at 0 included. After the exclusion, only 25% of the original sample remains, suggesting that a two-part model might be appropriate.

Table 2.2, Summary Statistics for Sample without Boundary Points

Sample without Boundary Cases = 1,116					
Variables	Min	Median	Max	Average	Variance
Leverage ratio	0.0001	0.3304	0.9984	0.3598	0.0521
Non-debt tax shields	0.0000	0.6179	22.6650	0.7792	1.2978
Collateral	0.0004	0.3724	0.9583	0.3794	0.0485
Size	11.0652	14.7983	18.5866	14.6759	1.8242
Profitability	0.0021	0.1071	0.5606	0.1218	0.0055
Expected growth	-52.2755	6.9420	207.5058	12.6273	670.0033
Age	6.0000	19.0000	163.0000	23.2070	267.3015
Liquidity	0.0000	0.0578	0.9522	0.1188	0.0240

Before the model estimation, it is important to have a general understanding about the predictiveness of each attribute by checking the Information Value (IV) and K-S statistic (KS). A rule of thumb is that predictors with IV < 0.03 are considered weak.

```
----------------------------------------------------------------------
| RANK   | VARIABLE RANKED BY IV          |  KS     | INFO. VALUE |
----------------------------------------------------------------------
|  001   | X3                             | 30.5942 |  0.6568     |
|  002   | X7                             | 19.7778 |  0.2236     |
|  003   | X4                             | 13.4758 |  0.0957     |
|  004   | X2                             |  9.4490 |  0.0389     |
|  005   | X1                             |  4.5119 |  0.0108     |
|  006   | X6                             |  4.5101 |  0.0082     |
|  007   | X5                             |  3.5278 |  0.0065     |
----------------------------------------------------------------------
```

From the above output, three attributes, ***X1*** (non-debt tax shields), ***X5*** (expected growth), and ***X6*** (age), are deemed unpredictive.

A further bivariate analysis might provide a deeper understanding about the relationship between each predictor and the response. The output below shows that large-size (***X3***) businesses with higher collaterals (***X2***) might be more likely to raise debts. On the other hand, a business with higher liquidity (***X7***) and profitability (***X4***) might be less likely to borrow.

```
X1
  BIN#    LOWER LIMIT    UPPER LIMIT     #FREQ    DISTRIBUTION      AVERAGE Y    INFO. VALUE        KS
---------------------------------------------------------------------------------------------------------
   001         0.0000         0.3493      1105        24.9943%        8.0476%     0.00414902    3.1343
   002         0.3494         0.5666      1105        24.9943%        8.6281%     0.00077704    4.5119
   003         0.5666         0.7891      1106        25.0170%        9.2821%     0.00014369    3.9094
   004         0.7894       102.1495      1105        24.9943%       10.3749%     0.00575742    0.0000
---------------------------------------------------------------------------------------------------------
         # TOTAL = 4421, AVERAGE Y = 0.090832, MAX. KS = 4.5119, INFO. VALUE = 0.0108.
---------------------------------------------------------------------------------------------------------
```

X2

BIN#	LOWER LIMIT	UPPER LIMIT	#FREQ	DISTRIBUTION	AVERAGE Y	INFO. VALUE	KS
001	0.0000	0.1241	1105	24.9943%	7.4923%	0.01010749	4.8152
002	0.1241	0.2875	1105	24.9943%	7.5522%	0.00932720	9.4490
003	0.2876	0.4724	1106	25.0170%	9.9575%	0.00268999	6.8004
004	0.4724	0.9953	1105	24.9943%	11.3301%	0.01673299	0.0000

TOTAL = 4421, AVERAGE Y = 0.090832, MAX. KS = 9.4490, INFO. VALUE = 0.0389.

X3

BIN#	LOWER LIMIT	UPPER LIMIT	#FREQ	DISTRIBUTION	AVERAGE Y	INFO. VALUE	KS
001	7.7381	11.2836	442	9.9977%	0.5141%	0.30726072	10.3742
002	11.2871	12.0539	442	9.9977%	2.9520%	0.08827258	17.7969
003	12.0548	12.5705	442	9.9977%	4.6285%	0.03894078	23.1901
004	12.5721	13.0748	442	9.9977%	5.3080%	0.02641217	27.7605
005	13.0772	13.5393	442	9.9977%	6.7426%	0.00916396	30.5942
006	13.5396	14.0091	443	10.0204%	10.7623%	0.00383568	28.5568
007	14.0106	14.5012	442	9.9977%	12.1215%	0.01186405	24.8785
008	14.5018	15.0207	442	9.9977%	12.8358%	0.01762509	20.3354
009	15.0218	15.6997	442	9.9977%	16.8331%	0.06624132	10.9530
010	15.7019	18.5866	442	9.9977%	18.1305%	0.08718412	0.0000

TOTAL = 4421, AVERAGE Y = 0.090832, MAX. KS = 30.5942, INFO. VALUE = 0.6568.

X4

BIN#	LOWER LIMIT	UPPER LIMIT	#FREQ	DISTRIBUTION	AVERAGE Y	INFO. VALUE	KS
001	0.0000	0.0875	1473	33.3183%	11.4561%	0.02475130	9.5735
002	0.0875	0.1606	1474	33.3409%	10.0498%	0.00436317	13.4758
003	0.1606	1.5902	1474	33.3409%	5.7454%	0.06658190	0.0000

TOTAL = 4421, AVERAGE Y = 0.090832, MAX. KS = 13.4758, INFO. VALUE = 0.0957.

X5

BIN#	LOWER LIMIT	UPPER LIMIT	#FREQ	DISTRIBUTION	AVERAGE Y	INFO. VALUE	KS
001	-81.2476	-0.2018	1473	33.3183%	8.2088%	0.00390846	3.5278
002	-0.1945	15.2839	1474	33.3409%	9.2366%	0.00011419	2.9084
003	15.3126	681.3542	1474	33.3409%	9.8036%	0.00245116	0.0000

TOTAL = 4421, AVERAGE Y = 0.090832, MAX. KS = 3.5278, INFO. VALUE = 0.0065.

X6

BIN#	LOWER LIMIT	UPPER LIMIT	#FREQ	DISTRIBUTION	AVERAGE Y	INFO. VALUE	KS
001	6.0000	16.0000	2176	49.2196%	8.3265%	0.00429694	4.5101
002	17.0000	210.0000	2245	50.7804%	9.8167%	0.00386761	0.0000

TOTAL = 4421, AVERAGE Y = 0.090832, MAX. KS = 4.5101, INFO. VALUE = 0.0082.

X7

BIN#	LOWER LIMIT	UPPER LIMIT	#FREQ	DISTRIBUTION	AVERAGE Y	INFO. VALUE	KS
001	0.0000	0.0161	631	14.2728%	14.0435%	0.04216450	8.5729
002	0.0161	0.0417	632	14.2954%	12.8438%	0.02530139	15.0827
003	0.0418	0.0796	632	14.2954%	11.7955%	0.01368992	19.7778
004	0.0797	0.1436	631	14.2728%	8.4866%	0.00076815	18.7466
005	0.1437	0.2518	632	14.2954%	7.1569%	0.00864727	15.4120

```
   006           0.2532           0.4509        632       14.2954%          5.9821%       0.02422189    10.0437
   007           0.4510           1.0002        631       14.2728%          3.2720%       0.10877110     0.0000
--------------------------------------------------------------------------------------------------------------
          # TOTAL = 4421, AVERAGE Y = 0.090832, MAX. KS = 19.7778, INFO. VALUE = 0.2236.
--------------------------------------------------------------------------------------------------------------
```

3. Single-Component Models

In this section, three modeling approaches, Tobit, NLS (Nonlinear Least Squares), and Fractional Logit models, that can generically handle fractional outcomes with boundary points at 0 / 1 are discussed. Although these models differ significantly from each other from statistical aspects, they all share the business assumption that both zero debt and positive debt decisions are determined by the same mechanism.

3.1 Tobit Model

Based upon the censored normal distribution, Tobit model has been commonly used in modeling outcomes with boundaries and is generalizable to fractional outcomes in the [0, 1] interval. Specifically, Tobit model assumes that there is a latent variable ***Y**** such that

$$\boldsymbol{Y} = \begin{cases} \boldsymbol{0} & for\ \boldsymbol{Y}^* \le \boldsymbol{0} \\ \boldsymbol{X`\beta} + \boldsymbol{\varepsilon}\ for\ \boldsymbol{1} > \boldsymbol{Y}^* > 0 \\ \boldsymbol{1} & for\ \boldsymbol{Y}^* \ge \boldsymbol{1} \end{cases}, where\ the\ error\ term\ \boldsymbol{\varepsilon} \sim \boldsymbol{Normal}(\boldsymbol{0}, \boldsymbol{\sigma^2})$$

Therefore, the response ***Y*** bounded by [0, 1] might be considered the observable part of a normally distributed variable ***Y* ~ Normal (X`β, σ²)*** defined on the real line. However, a fundamental argument against the censoring assumption is that the reason for unobservable values out of the [0, 1] interval is not a result of the censorship but due to the fact that any value out of [0, 1] is not theoretically defined. Hence, the censored normal distribution might not be the most appropriate assumption for fractional outcomes. Moreover, since Tobit model is still based on the normal distribution and the probability function of any value in the (0, 1) interval is identical to the one of OLS regression, it defers to statistical assumptions applicable to OLS regression, e.g. homoscedasticity, that are often violated in fractional outcomes.

In SAS, the most convenient way to estimate Tobit model is with QLIM procedure in SAS / ETS module. In order to clearly illustrate the log likelihood function of Tobit model, we'd like to choose NLMIXED procedure in SAS / STAT module. The maximum likelihood estimator for Tobit model assumes that errors are normal and homoscedastic and would be otherwise inconsistent. Consequently, the simultaneous estimation of a variance model might be needed to account for the heteroscedasticity as the following.

$$\boldsymbol{E}(\boldsymbol{\varepsilon^2}) = \boldsymbol{\sigma^2} \times \big(\boldsymbol{1} + \boldsymbol{EXP}(\boldsymbol{Z`G})\big)$$

In other words, there are two components in the Tobit model specification, a mean model and a variance counterpart, as demonstrated below.

```
proc nlmixed data = data.deve tech = trureg alpha = 0.01;
  parms b0 = 0  b1 = 0  b2 = 0  b3 = 0  b4 = 0  b5 = 0  b6 = 0  b7 = 0
        _s = 1  c1 = 0  c2 = 0  c3 = 0  c4 = 0  c5 = 0  c6 = 0  c7 = 0;
  xb = b0 + b1 * x1 + b2 * x2 + b3 * x3 + b4 * x4 + b5 * x5 + b6 * x6 + b7 * x7;
  xc = c1 * x1 + c2 * x2 + c3 * x3 + c4 * x4 + c5 * x5 + c6 * x6 + c7 * x7;
  s = (_s ** 2 * (1 + exp(xc))) ** 0.5;
  if y > 0 and y < 1 then lh = pdf('normal', y, xb, s);
  else if y <= 0 then lh = cdf('normal', 0, xb, s);
  else if y >= 1 then lh = 1 - cdf('normal', 1, xb, s);
  ll = log(lh);
  model y ~ general(ll);
run;
/*
             Fit Statistics
-2 Log Likelihood                 2347.3
AIC (smaller is better)           2379.3
AICC (smaller is better)          2379.5
BIC (smaller is better)           2473.4

                        Parameter Estimates
                       Standard
Parameter   Estimate      Error     DF   t Value   Pr > |t|
b0           -2.2379     0.1551   2641    -14.43     <.0001
b1          -0.01309    0.01286   2641     -1.02     0.3085
b2            0.4974    0.07421   2641      6.70     <.0001
b3            0.1415    0.01075   2641     13.16     <.0001
b4           -0.6824     0.2227   2641     -3.06     0.0022
b5          -0.00008   0.000530   2641     -0.16     0.8753
b6          -0.00075   0.000917   2641     -0.82     0.4122
b7           -0.6039     0.1231   2641     -4.91     <.0001
_s            0.3657    0.03059   2641     11.95     <.0001
c1           0.01383    0.06904   2641      0.20     0.8412
c2           -2.3440     0.6898   2641     -3.40     0.0007
c3           0.04668    0.02472   2641      1.89     0.0591
c4            0.1218     1.2744   2641      0.10     0.9238
c5          0.001200   0.002851   2641      0.42     0.6739
c6          -0.02245    0.01166   2641     -1.93     0.0543
c7            1.5452     0.4685   2641      3.30     0.0010
*/
```

As shown in the output, ***X2*** and ***X7*** are statistically significant in both models, implying the dependence between the conditional variance and the conditional mean.

3.2 NLS Regression Model

NLS regression is another alternative to model outcomes in the [0, 1] interval by assuming

$$\boldsymbol{Y} = \frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{EXP}(-\boldsymbol{X}`\boldsymbol{\beta})} + \boldsymbol{\varepsilon}, where\ the\ error\ term\ \boldsymbol{\varepsilon} \sim \boldsymbol{Normal}(\mathbf{0}, \boldsymbol{\sigma}^2)$$

Therefore, the conditional mean of ***Y*** can be represented as ***1 / [1 + EXP (-X`β)]***. Similar to OLS or Tobit regression, NLS regression also defers to the homoscedastic assumption. As a result, a variance model is also needed to account for the heteroscedasticity.

$$\boldsymbol{E}(\boldsymbol{\varepsilon}^2) = \boldsymbol{\sigma}^2 \times \left(\mathbf{1} + \boldsymbol{EXP}(\boldsymbol{Z}`\boldsymbol{G})\right)$$

The SAS implementation of NLS regression with NLMIXED procedure is shown below.

```
proc nlmixed data = data.deve tech = trureg;
  parms b0 = 0    b1 = 0  b2 = 0  b3 = 0  b4 = 0  b5 = 0  b6 = 0  b7 = 0
        _s = 0.1  c1 = 0  c2 = 0  c3 = 0  c4 = 0  c5 = 0  c6 = 0  c7 = 0;
  xb = b0 + b1 * x1 + b2 * x2 + b3 * x3 + b4 * x4 + b5 * x5 + b6 * x6 + b7 * x7;
  xc = c1 * x1 + c2 * x2 + c3 * x3 + c4 * x4 + c5 * x5 + c6 * x6 + c7 * x7;
  mu = 1 / (1 + exp(-xb));
  s  = (_s ** 2 * (1 + exp(xc))) ** 0.5;
  lh = pdf('normal', y, mu, s);
  ll = log(lh);
  model y ~ general(ll);
run;
/*
             Fit Statistics
-2 Log Likelihood                  -2167
AIC (smaller is better)            -2135
AICC (smaller is better)           -2135
BIC (smaller is better)            -2041

                        Parameter Estimates
                       Standard
Parameter   Estimate      Error     DF   t Value   Pr > |t|
b0           -7.4915     0.4692   2641    -15.97     <.0001
b1          -0.04652    0.03268   2641     -1.42     0.1547
b2            0.8447     0.2123   2641      3.98     <.0001
b3            0.4098    0.03316   2641     12.36     <.0001
b4           -3.3437     0.6233   2641     -5.36     <.0001
b5          0.001015   0.001341   2641      0.76     0.4489
b6          -0.00914   0.002853   2641     -3.20     0.0014
b7           -1.1170     0.2911   2641     -3.84     0.0001
_s           0.01499   0.002028   2641      7.39     <.0001
c1          -0.05461    0.01310   2641     -4.17     <.0001
c2            0.4066     0.1347   2641      3.02     0.0026
c3            0.4229    0.02041   2641     20.72     <.0001
c4           -3.6905     0.3188   2641    -11.58     <.0001
c5          0.001291   0.000842   2641      1.53     0.1255
c6          -0.01644   0.002053   2641     -8.01     <.0001
c7           -1.0388     0.1332   2641     -7.80     <.0001
*/
```

In the above output, most predictors are statistically significant in both the mean and the variance models, indicating a strong likelihood of heteroscedasticity.

3.3 Fractional Logit Model

Different from two models discussed above with specific distributional assumptions, fractional Logit model (Papke and Wooldridge, 1996) is a quasi-likelihood method that does not assume any distribution but only requires the conditional mean to be correctly specified for consistent parameter estimates. Under the assumption ***E (Y|X) = G (X`β) = 1 / [1 + EXP (-X`β)]***, fractional Logit model has the identical likelihood function

$$\boldsymbol{F(Y) = G(X`\beta)^{Y} \times \left(1 - G(X`\beta)\right)^{1-Y}} \ for\ \boldsymbol{1 \geq Y \geq 0}$$

Based upon the above formulation, parameters can be estimated in the same manner as in the binary logistic regression by maximizing the log likelihood function.

In SAS, the most convenient way to implement a fractional Logit model is with GLIMMIX procedure in SAS / STAT module. In addition, we can also use NLMIXED procedure by explicitly specifying the likelihood function as below.

```
proc nlmixed data = data.deve tech = trureg;
  parms b0 = 0  b1 = 0  b2 = 0  b3 = 0  b4 = 0  b5 = 0  b6 = 0  b7 = 0;
  xb = b0 + b1 * x1 + b2 * x2 + b3 * x3 + b4 * x4 + b5 * x5 + b6 * x6 + b7 * x7;
  mu = 1 / (1 + exp(-xb));
  lh = (mu ** y) * ((1 - mu) ** (1 - y));
  ll = log(lh);
  model y ~ general(ll);
run;
/*
             Fit Statistics
-2 Log Likelihood                 1483.7
AIC (smaller is better)           1499.7
AICC (smaller is better)          1499.7
BIC (smaller is better)           1546.7

                        Parameter Estimates
                       Standard
Parameter   Estimate      Error     DF   t Value   Pr > |t|
b0           -7.3467     0.7437   2641     -9.88     <.0001
b1          -0.05820    0.06035   2641     -0.96     0.3349
b2            0.8480     0.3276   2641      2.59     0.0097
b3            0.3996    0.05151   2641      7.76     <.0001
b4           -3.4801     1.0181   2641     -3.42     0.0006
b5          0.000910   0.002027   2641      0.45     0.6534
b6          -0.00859   0.005018   2641     -1.71     0.0871
b7           -1.0455     0.4403   2641     -2.37     0.0176
*/
```

It is worth mentioning that fractional Logit model can be easily transformed to a weighted logistic regression with binary outcomes (shown below), which will yield identical parameter estimates and statistical inferences. As a result, most model development techniques and statistical diagnostics used in the logistic regression might also be applicable to fractional Logit model.

```
data deve;
  set data.deve (in = a) data.deve (in = b);
  if a then do;
    y2 = 1;
    wt = y;
  end;
  if b then do;
    y2 = 0;
    wt = 1 - y;
  end;
run;

proc logistic data = deve desc;
  model y2 = x1 - x7;
  weight wt;
run;
/*
                             Intercept
              Intercept            and
Criterion          Only     Covariates
```

AIC		1622.697	1499.668		
SC		1628.804	1548.523		
-2 Log L		1620.697	1483.668		

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-7.3469	0.7437	97.6017	<.0001
x1	1	-0.0581	0.0603	0.9276	0.3355
x2	1	0.8478	0.3276	6.6991	0.0096
x3	1	0.3996	0.0515	60.1804	<.0001
x4	1	-3.4794	1.0180	11.6819	0.0006
x5	1	0.000910	0.00203	0.2017	0.6533
x6	1	-0.00859	0.00502	2.9288	0.0870
x7	1	-1.0455	0.4403	5.6386	0.0176

*/

4. Two-Part Composite Models

The previous data exploration shows that ~75% businesses in the sample carried no debt at all. Therefore, it might be appealing to employ zero-inflated fractional models, a Logit model separating zeroes from positive fractional outcomes and then a subsequent sub-model governing all values in the interval (0, 1) conditional on nonzero outcomes. A general form of the conditional mean for zero-inflated fractional models can be represented by

$$\boldsymbol{E(Y|X) = E(Y|X, Y = 0) \times Pr(Y = 0|X) + E\big(Y\big|X, Y \in (0, 1)\big) \times Pr(Y \in (0, 1)|X)}$$

$$\Rightarrow \quad \boldsymbol{E(Y|X) = E\big(Y\big|X, Y \in (0, 1)\big) \times Pr(Y \in (0, 1)|X)}$$

In this paper, Beta and Simplex regressions will be used to model nonzero fractional outcomes. From the interpretation standpoint, two-part models could imply that the financial leverage of a business might be a two-stage decision process. First of all, the business should decide if it is going to take the debt or not. Given the condition that the business is planning to take the debt, then it will further decide how much to borrow.

4.1 Beta Model

Beta regression is a flexible modeling facility based upon the two-parameter Beta distribution and can be employed to model any continuous variable bounded by two known endpoints, e.g. 0 and 1 in this case. Assuming that $\boldsymbol{Y}$ follows a standard Beta distribution defined in the interval (0, 1) with two shape parameters $\boldsymbol{\omega}$ and $\boldsymbol{\tau}$, the density function can be specified as

$$\boldsymbol{F(Y) = \frac{Gamma(\omega + \tau)}{\big(Gamma(\omega) \times Gamma(\tau)\big)} \times Y^{\omega - 1} \times (1 - Y)^{\tau - 1}}$$

In the above formulation, while $\boldsymbol{\omega}$ is pulling the density toward 0, $\boldsymbol{\tau}$ is pushing the density toward 1. Without the loss of generality, $\boldsymbol{\omega}$ and $\boldsymbol{\tau}$ can be re-parameterized and translated into two other parameters, location parameter $\boldsymbol{\mu}$ and dispersion parameter $\boldsymbol{\varphi}$ with $\boldsymbol{\omega = \mu \times \varphi}$ and $\boldsymbol{\tau = \varphi \times (1 - \mu)}$, of which $\boldsymbol{\mu}$ is the expected mean and $\boldsymbol{\varphi}$ governs the variance such that

$$\boldsymbol{\sigma^2 = \frac{\mu \times (1 - \mu)}{(1 + \varphi)}}$$

Within the framework of Generalized Linear Models (GLM), $\boldsymbol{\mu}$ and $\boldsymbol{\varphi}$ can be modeled separately with a location model for $\boldsymbol{\mu}$ and a dispersion model for $\boldsymbol{\varphi}$ using two different or identical sets of covariates $\boldsymbol{X}$ and $\boldsymbol{Z}$. Since the expected mean $\boldsymbol{\mu}$ is bounded by 0 and 1, a natural choice of the link function is the Logit function such that $\boldsymbol{LOG\ [\mu / (1 - \mu)] = X`\beta}$. With the strictly positive nature of $\boldsymbol{\varphi}$, the Log function seems appropriate such that $\boldsymbol{LOG\ (\varphi) = Z`\gamma}$.

SAS, as of today, does not provide an out-of-box procedure to estimate the two-parameter Beta model. GLIMMIX procedure can only estimate a simple form of Beta regression without the dispersion model. However, NLMIXED procedure supports the straightforward Beta model estimation if one can explicitly specify the log likelihood function.

```
proc nlmixed data = data.deve tech = trureg;
  parms a0 = 0  a1 = 0  a2 = 0  a3 = 0  a4 = 0  a5 = 0  a6 = 0  a7 = 0
        b0 = 0  b1 = 0  b2 = 0  b3 = 0  b4 = 0  b5 = 0  b6 = 0  b7 = 0
        c0 = 1  c1 = 0  c2 = 0  c3 = 0  c4 = 0  c5 = 0  c6 = 0  c7 = 0;
  xa = a0 + a1 * x1 + a2 * x2 + a3 * x3 + a4 * x4 + a5 * x5 + a6 * x6 + a7 * x7;
  xb = b0 + b1 * x1 + b2 * x2 + b3 * x3 + b4 * x4 + b5 * x5 + b6 * x6 + b7 * x7;
  xc = c0 + c1 * x1 + c2 * x2 + c3 * x3 + c4 * x4 + c5 * x5 + c6 * x6 + c7 * x7;
  mu_xa = 1 / (1 + exp(-xa));
  mu_xb = 1 / (1 + exp(-xb));
  phi = exp(xc);
  w = mu_xb * phi;
  t = (1 - mu_xb) * phi;
  if y = 0 then lh = 1 - mu_xa;
  else lh = mu_xa * (gamma(w + t) / (gamma(w) * gamma(t)) * (y ** (w - 1)) * ((1 - y) ** (t - 1)));
  ll = log(lh);
  model y ~ general(ll);
run;
/*
             Fit Statistics
-2 Log Likelihood                2131.2
AIC (smaller is better)          2179.2
AICC (smaller is better)         2179.7
BIC (smaller is better)          2320.3

                       Parameter Estimates
                       Standard
Parameter   Estimate      Error     DF   t Value    Pr > |t|
a0           -9.5003     0.5590   2641    -17.00     <.0001
a1          -0.03997    0.03456   2641     -1.16     0.2476
a2            1.5725     0.2360   2641      6.66     <.0001
a3            0.6185    0.03921   2641     15.77     <.0001
a4           -2.2842     0.6445   2641     -3.54     0.0004
a5          -0.00087   0.001656   2641     -0.52     0.6010
a6          -0.00530   0.003460   2641     -1.53     0.1256
a7           -1.5349     0.3096   2641     -4.96     <.0001
b0            1.6136     0.4473   2641      3.61     0.0003
b1          -0.02592    0.03277   2641     -0.79     0.4290
b2           -0.3756     0.1781   2641     -2.11     0.0351
b3           -0.1139    0.03017   2641     -3.77     0.0002
b4           -2.7927     0.5133   2641     -5.44     <.0001
b5          0.003064   0.001527   2641      2.01     0.0448
b6          -0.00439   0.002475   2641     -1.77     0.0764
b7            0.2253     0.2434   2641      0.93     0.3548
c0           -0.2832     0.5877   2641     -0.48     0.6300
```

```
c1          -0.00171    0.04219   2641     -0.04     0.9678
c2            0.6073     0.2311   2641      2.63     0.0086
c3           0.07857    0.03988   2641      1.97     0.0489
c4            2.2920     0.7207   2641      3.18     0.0015
c5          -0.00435   0.001643   2641     -2.65     0.0081
c6          0.001714   0.003388   2641      0.51     0.6130
c7          -0.09279     0.3357   2641     -0.28     0.7823
*/
```

As shown above, in a zero-inflated Beta model, there are three sets of parameters to be estimated, one for Logit model and the other two for Beta model. It is worth pointing out that instead of jointly estimating both Logit and Beta models simultaneously, one might estimate them independently with a logistic regression to separate zeroes from non-zeroes and then a Beta regression applied to all values in the (0, 1) interval.

4.2 Simplex Model

The last approach introduced for modeling fractional outcomes in the interval (0, 1), known as Simplex model, might be a “new kid in town” for most statisticians and is considered a special case of dispersion models (Jorgensen, 1997). Within the framework of dispersion models, Song (Song, 2009) showed that the probability function of any dispersion model can represented by a general form

$$\boldsymbol{F}(\boldsymbol{Y}) = \{\mathbf{2} \times \boldsymbol{\pi} \times \boldsymbol{\sigma}^{\mathbf{2}} \times \boldsymbol{V}(\boldsymbol{Y})\}^{-\mathbf{0.5}} \times \boldsymbol{EXP}\left\{\frac{-\mathbf{1}}{\mathbf{2} \times \boldsymbol{\sigma}^{\mathbf{2}}} \times \boldsymbol{D}(\boldsymbol{Y})\right\}$$

The variance function ***V(Y)*** and the deviance function ***D(Y)*** vary by distributional assumptions. For the Simplex distribution,

$$\boldsymbol{V}(\boldsymbol{Y}) = \boldsymbol{Y}^{\mathbf{3}} \times (\mathbf{1} - \boldsymbol{Y})^{\mathbf{3}}$$

$$\boldsymbol{D}(\boldsymbol{Y}) = \frac{(\boldsymbol{Y} - \boldsymbol{\mu})^{\mathbf{2}}}{\boldsymbol{Y} \times (\mathbf{1} - \boldsymbol{Y}) \times \boldsymbol{\mu}^{\mathbf{2}} \times (\mathbf{1} - \boldsymbol{\mu})^{\mathbf{2}}}$$

Similar to the Beta model, a Simplex model also consists of two components, a model estimating the expected mean $\boldsymbol{\mu}$ such that $\boldsymbol{0 < \mu < 1}$ and the other describing the pattern of a dispersion parameter $\boldsymbol{\sigma}$. Since $\boldsymbol{0 < \mu < 1}$, Logit link function can be used to specify the relationship between the expected mean $\boldsymbol{\mu}$ and covariates $\boldsymbol{X}$ such that $\boldsymbol{LOG\ [\mu\ /\ (1 - \mu)] = X`\beta}$. Because of the strict positivity of $\boldsymbol{\sigma^2}$, the model for dispersion parameter $\boldsymbol{\sigma}$ can be formulated as $\boldsymbol{LOG\ (\sigma^2) = Z`\gamma}$.

Again, there is no out-of-box procedure in SAS to estimate the Simplex model. The probability function needs to be specified explicitly with NLMIXED procedure in order to estimate a Simplex model.

```
proc nlmixed data = data.deve tech = trureg;
  parms a0 = 0  a1 = 0  a2 = 0  a3 = 0  a4 = 0  a5 = 0  a6 = 0  a7 = 0
        b0 = 0  b1 = 0  b2 = 0  b3 = 0  b4 = 0  b5 = 0  b6 = 0  b7 = 0
        c0 = 9  c1 = 0  c2 = 0  c3 = 0  c4 = 0  c5 = 0  c6 = 0  c7 = 0;
  xa = a0 + a1 * x1 + a2 * x2 + a3 * x3 + a4 * x4 + a5 * x5 + a6 * x6 + a7 * x7;
  xb = b0 + b1 * x1 + b2 * x2 + b3 * x3 + b4 * x4 + b5 * x5 + b6 * x6 + b7 * x7;
```

```
  xc = c0 + c1 * x1 + c2 * x2 + c3 * x3 + c4 * x4 + c5 * x5 + c6 * x6 + c7 * x7;
  mu_xa = 1 / (1 + exp(-xa));
  mu_xb = 1 / (1 + exp(-xb));
  s2 = exp(xc);
  v = (y * (1 - y)) ** 3;
  if y = 0 then do;
    lh = 1 - mu_xa;
    ll = log(lh);
  end;
  else do;
    d = ((y - mu_xb) ** 2) / (y * (1 - y) * mu_xb ** 2 * (1 - mu_xb) ** 2);
    lh = mu_xa * (2 * constant('pi') * s2 * v) ** (-0.5) * exp(-(2 * s2) ** (-1) * d);
    ll = log(lh);
  end;
  model y ~ general(ll);
run;
/*
             Fit Statistics
-2 Log Likelihood                 2672.1
AIC (smaller is better)           2720.1
AICC (smaller is better)          2720.5
BIC (smaller is better)           2861.1

                        Parameter Estimates
                       Standard
Parameter   Estimate      Error     DF   t Value   Pr > |t|
a0           -9.5003     0.5590   2641    -17.00     <.0001
a1          -0.03997    0.03456   2641     -1.16     0.2476
a2            1.5725     0.2360   2641      6.66     <.0001
a3            0.6185    0.03921   2641     15.77     <.0001
a4           -2.2842     0.6445   2641     -3.54     0.0004
a5          -0.00087   0.001656   2641     -0.52     0.6010
a6          -0.00530   0.003460   2641     -1.53     0.1256
a7           -1.5349     0.3096   2641     -4.96     <.0001
b0           -0.5412     0.4689   2641     -1.15     0.2485
b1           0.03485    0.02576   2641      1.35     0.1763
b2           -1.3480     0.2006   2641     -6.72     <.0001
b3           0.01708    0.03098   2641      0.55     0.5814
b4           -2.0596     0.5731   2641     -3.59     0.0003
b5          0.004635   0.001683   2641      2.75     0.0059
b6          -0.00006   0.002652   2641     -0.02     0.9818
b7            0.7973     0.2945   2641      2.71     0.0068
c0            9.9250     0.5582   2641     17.78     <.0001
c1           -0.1034    0.04846   2641     -2.13     0.0329
c2            1.6217     0.2960   2641      5.48     <.0001
c3           -0.4550    0.03652   2641    -12.46     <.0001
c4           -4.1401     0.8523   2641     -4.86     <.0001
c5          0.007653   0.002079   2641      3.68     0.0002
c6          -0.00742   0.003526   2641     -2.11     0.0354
c7           -0.6699     0.4484   2641     -1.49     0.1353
*/
```

As demonstrated above, there are also three sets of parameters to be estimated in a zero-inflated Simplex model. Likewise, one might estimate a zero-inflated Simplex model either jointly or separately, both of which should yield identical results.

5. Model Evaluations

In previous sections, five models for fractional outcomes have been showcased with the financial leverage data. Upon the completion of model estimation, it is often of interests to check parameter estimates if they make both statistical and business senses. Since model effects of attributes and prediction accuracies are mainly determined by mean models, we would focus on parameter estimates of mean models only.

Table 5.1, Parameter Estimates of Five Models (mean models only)

Prameter Estimates	Single-Component Models			Two-Part Models		
	Tobit	NLS	Fractional	Logit	Beta	Simplex
β0	-2.2379	-7.4915	-7.3471	-9.5002	1.6136	-0.5412
β1	-0.0131	-0.0465	-0.0578	-0.0399	-0.0259	0.0349
β2	0.4974	0.8447	0.8475	1.5724	-0.3756	-1.3480
β3	0.1415	0.4098	0.3996	0.6184	-0.1139	0.0171
β4	-0.6824	-3.3437	-3.4783	-2.2838	-2.7927	-2.0596
β5	-0.0001	0.0010	0.0009	-0.0009	0.0031	0.0046
β6	-0.0008	-0.0091	-0.0086	-0.0053	-0.0044	-0.0001
β7	-0.6039	-1.1170	-1.0455	-1.5347	0.2253	0.7973

In table 5, all parameter estimates with p-values lower than 0.01 are highlighted. The negative relationship between ***X4*** (profitability) and the financial leverage is significant and consistent across all models. Interesting is that both ***X2*** (collateral) and ***X3*** (size) have consistent and significant positive impacts on the financial leverage in all single-component models. However, the story is different in two-part models. For instance, in the zero-inflated Beta model, while large-size firms might be more likely to borrow, there is a negative relationship between the size of a business and the leverage ratio given a decision made to raise the debt. Similarly in the zero-inflated Simplex model, although the business with a greater percent of collateral might be more likely to raise the debt, a significant negative relationship is observed between the collateral percentage and the leverage ratio conditional on the decision of borrowing. These are all worth further investigations.

While we can get a general sense about the impact direction, e.g. positive or negative, of each predictor from parameter estimates, we cannot simply compare coefficients across various models under different distributional assumptions to gauge the impact magnitude of a predictor. To facilitate such assessment, marginal effects, which are partial derivatives of the conditional expectation with respect to different predictors, are also provided towards an apple-to-apple comparison of predictor impacts across models.

Table 5.2, Marginal Effects

Marginal Effects	Tobit	NLS	Fractional	ZI Beta	ZI Simplex
X1	-0.003299	-0.003684	-0.004594	-0.003741	-0.000057
X2	0.125318	0.066882	0.066937	0.066859	0.003452
X3	0.035657	0.032450	0.031544	0.028246	0.029917
X4	-0.171933	-0.264740	-0.274700	-0.289299	-0.214689
X5	-0.000021	0.000080	0.000072	0.000128	0.000201
X6	-0.000190	-0.000724	-0.000678	-0.000551	-0.000252
X7	-0.152173	-0.088443	-0.082527	-0.073400	-0.030425

In the above comparison of marginal effects, it is shown that ***X4*** (profitability) has the biggest impact on the leverage ratio across all models. It is also interesting to notice that although fractional Logit and zero-inflated Beta regressions differs significantly in both distributional assumptions and modeling practices, marginal effects of all predictors between two models are reasonably consistent.

To compare multiple models with different distributional assumptions, academic statisticians often prefer to use likelihood-based measures such as Vuong statistic. Proposed by Quang Vuong (1989), Vuong test considers a better model with the individual log likelihoods significantly higher than the ones of its rival and is calculated as

$$\boldsymbol{Vuong\ Statistic} = \frac{\boldsymbol{LR}(\boldsymbol{Model1}, \boldsymbol{Model2}) - \boldsymbol{C}}{\sqrt{\boldsymbol{N} \times \boldsymbol{V}}} \sim \boldsymbol{Normal}(\boldsymbol{0}, \boldsymbol{1})$$

LR(…) is the summation of individual log likelihood ratios between 2 models. ***C*** is a correction term for the difference in parameter numbers between 2 models. ***N*** is the number of records. ***V*** is the variance of individual log likelihood ratios between 2 models. Vuong statistic is distributed as a standard ***Normal (0, 1)***. The model 1 is better with Vuong statistic > 1.96 and the model 2 is better with Vuong statistic < -1.96. A SAS macro calculating Vuong statistic is given below.

```
%macro vuong(data = , ll1 = , q1 = , ll2 = , q2 = );
***************************************************;
* INPUT PARAMETERS:                               *;
* =============================================== *;
*  DATA : INPUT SAS DATASET                       *;
*  LL1  : LOG LIKELIHOOD OF MODEL 1               *;
*  Q1   : # OF PARAMETERS IN MODEL 1              *;
*  LL2  : LOG LIKELIHOOD OF MODEL 2               *;
*  Q2   : # OF PARAMETERS IN MODEL 2              *;
* =============================================== *;
* OUTPUT:                                         *;
*  PRINT OUT OF VUONG STATISTIC                   *;
***************************************************;

data _tmp1;
  set &data;
  where &ll1 - &ll2 ~= .;
  m = &ll1 - &ll2;
run;

proc sql;
select
  (sum(m) - 0.5 * (&q1 - &q2) * log(count(*))) / ((count(*) * var(m)) ** 0.5) as vuong_stat
from
  _tmp1;
quit;

%mend vuong;
```

For instance, if we compare Tobit (Model 1) and zero-inflated Beta (Model 2) regressions with the above macro, Vuong statistic is equal to -5.84, implying that zero-inflated Beta is significantly better than Tobit in our case.

For most practitioners, it might be more intuitive to use empirical measures such as Information Value or R^2 calculated from the separate hold-out data sample, as shown below.

Table 5.3, Model Performances

Model Performance on Hold-out Sample					
Measures	**Tobit**	**NLS**	**Fractional**	**ZI Beta**	**ZI Simplex**
R^2	0.0896	0.0957	0.0965	0.1075	0.0868
Info. Value	0.7370	0.8241	0.8678	0.8551	0.7672

It appears that the zero-inflated Beta model, followed by the fractional Logit model, yields the best performance in terms of R^2 on the hold-out validation sample. In addition, the two-part nature of a zero-inflated Beta model might offer a bit more intriguing insights for further discussions. On the other hand, rolling out a zero-inflated Beta model to real-world problems might present challenges beyond the model development. The fractional Logit model therefore is favored more often overall.

6. Conclusion

In this paper, five different modeling facilities for fractional outcomes have been explored and initial use case details with various SAS procedures were provided. The fractional Logit model is a serious contender due to simpler implementations and liberal assumptions required therein. Complex models such as zero-inflated Beta may be where major potential improvements lie.

References

1. Kieschnick, R. and McCullough, B (2003), regression analysis of variates observed on (0, 1): percentages, proportions and fractions, Statistical Modeling 2003, 3, 193 – 213

2. Song, P (2009), dispersion models in regression analysis, Pakistan Journal of Statistics, Vol. 25(4), 529 – 551

3. Barndorff-Nielsen, O. E. and Jorgensen, B (1991), Journal of Multivariate Analysis, 39, 106 – 116

4. Papke, L. and J.M. Wooldridge (1996), Econometric methods for fractional response variables with an application to 401(K) plan participation rates, Journal of Applied Econometrics, 11(6), 619 - 632

5. Ramalho, E.A., J.J.S. Ramalho, and J.M.R. Murteira (2011), Alternative estimating and testing empirical strategies for fractional regression models, Journal of Economic Surveys, 25(1), 19 - 68

6. Smithson, M. and Verkuilen, J. (2006), A Better Lemon-Squeezer? Maximum Likelihood Regression with Beta-Distributed Dependent Variables, Psychological Methods, Vol. 11, No. 1, 54 – 71

7. Vuong, Q. (1989), Likelihood ratio tests for model selection and non-nested hypotheses, Econometrica 57, 307 - 333

Contact Information

WenSui Liu, Portfolio Analysis & Forecasting Manager, VP

Fifth Third Bancorp, Cincinnati, OH

Email: wensui.liu@foxmail.com

Jia (Jason) Xin, Pre-Sales

SAS, Boston, MA

Email: xinj2@yahoo.com