Paper FS-05-2013

Statistical Models for Proportional Outcomes

WenSui Liu, Fifth Third Bancorp, Cincinnati, OH

Kelly Zhao, Fifth Third Bancorp, Cincinnati, OH

Abstract

For many practitioners, ordinary least square (OLS) regression with Gaussian distributional assumption might be the top choice to model proportional 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 methodologies for proportional 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 proportional outcomes to the SAS user community and interested statistical practitioners in various industries.

Keywords

Proportional outcomes, Tobit model, Non-linear least squares (NLS) regression, Fractional Logit model, Beta regression, Simplex regression.

Introduction

In the financial service industry, we often observed business necessities to model proportional outcomes in the range of [0, 1]. For instance, in the context of credit risk, loss given default (LGD) 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 both the long-term debt and the equity.

To the best of my knowledge, although research interests in statistical models for proportional outcomes have remained strong in the past years, there is still no unanimous consensus on either the distributional assumption or the modeling practice. An interesting but somewhat ironic observation is that the simple OLS regression with Gaussian assumption has been the prevailing method to model proportional outcomes by most practitioners due to the simplicity. However, this approach suffers from a couple of conceptual flaws. First and the most evidential of all, proportional outcomes in the interval [0, 1] are not defined on the real line and therefore shouldn't be assumed normally distributed. Secondly, a profound statistical nature of proportion outcomes is that the variance is not independent of the mean. For instance, the variance shrinks when the mean approaches boundary points of [0, 1], which is a typical representation of the so-called Heteroscedasticity.

In addition to the aforementioned OLS regression approach, another class of OLS regression based upon the logistic normal assumption is also overwhelmingly popular among practitioners. In this approach, while boundary points at 0 or 1 can be handled heuristically, any outcome value in the open interval (0, 1) would be transformed by a Logit function such that

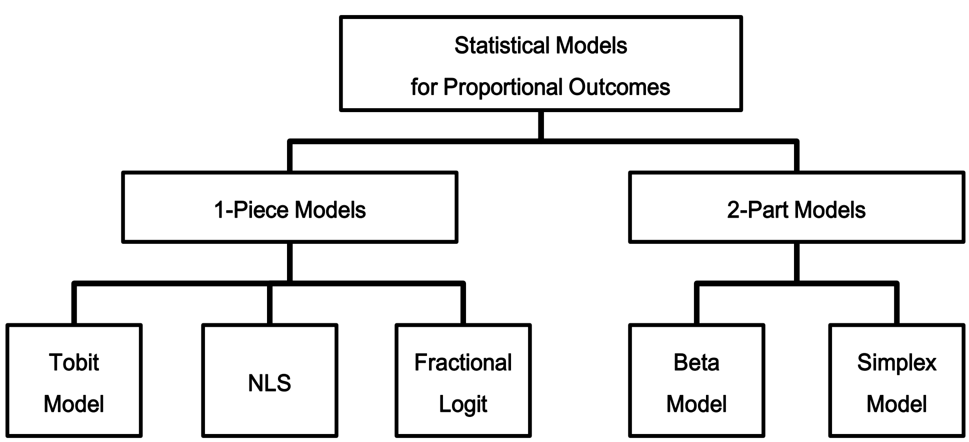
$$\boldsymbol{LOG}\left({}^{\boldsymbol{Y}}\!/_{\boldsymbol{1}-\boldsymbol{Y}}\right) = \boldsymbol{X}`\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ where the error term } \boldsymbol{\varepsilon} \sim \boldsymbol{Normal}(\boldsymbol{0}, \boldsymbol{\sigma^2})$$

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

Albeit simple, the OLS-based model with Logit transformation is not free of either conceptual or practical difficulties. A key concern is that, in order to ensure ***LOG (Y / (1 – Y)) ~ Normal (X`β, σ²)*** and therefore ***ε ~ Normal (0, σ²)***, the outcome variable ***Y*** should theoretically follow the additive logistic normal distribution, which might be questionable and is subject to statistical tests. For instance, it is important to check if 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 necessary to recover marginal effects on ***E (Y | X)*** from ***E (LOG (Y / (1 – Y)) | X)***.

Given all limitations of OLS regression discussed above, five alternative modeling approaches for proportional outcomes, which are loosely fallen into two broad categories, should be surveyed in the paper. The first category governs one-piece modeling approaches that are able to generically handle proportional 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 component, e.g. a Logit model, separating between boundary points and the open interval of (0, 1) and the other component governing all values in the (0, 1) interval by a Beta or Simplex model.

Figure 1, Schematic Diagram of Statistical Models for Fractional Outcomes



To better illustrate how to employ these five models in the practice, we would apply them to a use case of modeling the financial leverage ratio defined in the interval of [0, 1) 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, 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

Data Analysis

The preliminary data analysis might be the simplest and somehow tedious work in the pre-modeling stage. However, it is by all means the most critical component and is able to provide a more granular view about the data. First of all, we might take a look at the summary statistics of all variables.

Table 2, 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

Since the median of our response variable is equal to 0, it is evidential that the majority of outcome values are point mass at 0. Given this special statistical nature of the response variable, it might be helpful to take a second look at the data without boundary points at 0 in outcomes. After excluding cases with Y = 0, there are only 25% of the whole samples left, implying a potential necessity of two-part models.

Table 3, 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

In order to have a true picture about the performance of five different models, we split the full sample of 4,421 cases into two pieces, ~60% for the model development and ~40% for the post-model performance testing.

Table 4, Sample Separations

# of Cases	Full Sample	Deve. Sample	Test Sample
Y = 0	3,305	1,965	1,340
0 < Y < 1	1,116	676	440
Total	4,421	2,641	1,780

Before proceeding with the model development, we might need to have an idea about the predictiveness and the strength of each model attribute by checking Information Value and K-S statistic, as shown below. Empirically, variables with IV < 0.03 are usually considered unpredictive.

```
---------------------------------------------------------------------
| RANK   | VARIABLE RANKED BY IV            |  KS      | INFO. VALUE |
---------------------------------------------------------------------
|  001   | X3                               | 29.6582  |  0.6490     |
|  002   | X7                               | 18.0106  |  0.1995     |
|  003   | X4                               | 13.9611  |  0.1314     |
|  004   | X2                               | 10.7026  |  0.0470     |
|  005   | X5                               | 4.2203   |  0.0099     |
|  006   | X6                               | 4.0867   |  0.0083     |
|  007   | X1                               | 3.4650   |  0.0048     |
---------------------------------------------------------------------
```

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

For the other four variables with IV ≥ 0.03, the bivariate analysis might help us gain a deeper understanding about their relationships with the outcome variable of interest. For instance, it is clearly shown in the output below 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.

X3

BIN#	LOWER LIMIT	UPPER LIMIT	#FREQ	DISTRIBUTION	AVERAGE Y	INFO. VALUE	KS
001	7.7381	11.3302	293	11.0943%	0.6980%	0.30101685	11.2883
002	11.3313	12.1320	294	11.1321%	3.1150%	0.09286426	19.3900
003	12.1328	12.6855	293	11.0943%	5.2841%	0.03088467	24.5797
004	12.6924	13.2757	294	11.1321%	5.3806%	0.02925218	29.6582
005	13.2765	13.8196	293	11.0943%	9.6427%	0.00032449	29.0517
006	13.8201	14.3690	294	11.1321%	10.8879%	0.00428652	26.7815
007	14.3703	14.8901	293	11.0943%	11.7722%	0.00952048	23.3430
008	14.8925	15.6010	294	11.1321%	17.1834%	0.07665603	12.6723
009	15.6033	18.5045	293	11.0943%	18.7160%	0.10422852	0.0000

TOTAL = 2641, AVERAGE Y = 0.091866, MAX. KS = 29.6582, INFO. VALUE = 0.6490.

X7

BIN#	LOWER LIMIT	UPPER LIMIT	#FREQ	DISTRIBUTION	AVERAGE Y	INFO. VALUE	KS
001	0.0000	0.0161	377	14.2749%	13.7083%	0.03491945	7.7370
002	0.0161	0.0422	377	14.2749%	12.2633%	0.01702171	13.0015
003	0.0424	0.0798	378	14.3128%	12.1063%	0.01546113	18.0106
004	0.0802	0.1473	377	14.2749%	8.3757%	0.00140556	16.6230
005	0.1473	0.2610	378	14.3128%	7.6741%	0.00509632	14.0283
006	0.2613	0.4593	377	14.2749%	6.9672%	0.01141868	10.2307
007	0.4611	1.0002	377	14.2749%	3.2075%	0.11417533	0.0000

TOTAL = 2641, AVERAGE Y = 0.091866, MAX. KS = 18.0106, INFO. VALUE = 0.1995.

X4

BIN#	LOWER LIMIT	UPPER LIMIT	#FREQ	DISTRIBUTION	AVERAGE Y	INFO. VALUE	KS
001	0.0000	0.0628	528	19.9924%	11.6035%	0.01508943	5.7918
002	0.0628	0.1007	528	19.9924%	11.5788%	0.01479740	11.5245
003	0.1007	0.1423	529	20.0303%	10.2015%	0.00282718	13.9611
004	0.1425	0.2090	528	19.9924%	8.2715%	0.00252082	11.7682
005	0.2090	1.5902	528	19.9924%	4.2758%	0.09619720	0.0000

TOTAL = 2641, AVERAGE Y = 0.091866, MAX. KS = 13.9611, INFO. VALUE = 0.1314.

X2

BIN#	LOWER LIMIT	UPPER LIMIT	#FREQ	DISTRIBUTION	AVERAGE Y	INFO. VALUE	KS
001	0.0000	0.1249	660	24.9905%	7.3259%	0.01374506	5.5737
002	0.1251	0.2846	660	24.9905%	7.4744%	0.01153704	10.7026
003	0.2849	0.4670	661	25.0284%	10.6583%	0.00728202	6.2874
004	0.4671	0.9953	660	24.9905%	11.2856%	0.01440832	0.0000

TOTAL = 2641, AVERAGE Y = 0.091866, MAX. KS = 10.7026, INFO. VALUE = 0.0470.

One-Piece Models

In this section, three models, namely Tobit, NLS (nonlinear least squares), and Fractional Logit models, that can generically handle proportional outcomes with boundary points would be discussed. Although these three models are different significantly from each other from statistical aspects, they all share the assumption that both zero debt and positive debt decisions are determined by the same mechanism.

1. Tobit Model

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

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

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

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

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

Thus, there are two components in the Tobit model specification, both a mean and a variance sub-models. Due to the computational complexity of two-component joint models with NLMIXED, it is always a good strategy to start with a simpler model estimating the conditional mean only and then extend to the variance component, as shown below.

```
ods output parameterestimates = _parms;
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 = 1;
  xb = b0 + b1 * x1 + b2 * x2 + b3 * x3 + b4 * x4 +
       b5 * x5 + b6 * x6 + b7 * x7;
  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;

proc sql noprint;
  select parameter||" = "||compress(put(estimate, 18.4), ' ')
  into :parms separated by ' ' from _parms;
quit;

proc nlmixed data = data.deve tech = trureg;
  parms &parms  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|    Alpha
b0           -2.2379    0.1548   2641    -14.46     <.0001     0.05   ***
b1          -0.01309   0.01276   2641     -1.03     0.3049     0.05
b2            0.4974   0.07353   2641      6.76     <.0001     0.05   ***
b3            0.1415   0.01072   2641     13.20     <.0001     0.05   ***
b4           -0.6824    0.2178   2641     -3.13     0.0017     0.05   ***
b5          -0.00008  0.000528   2641     -0.16     0.8749     0.05
b6          -0.00075  0.000918   2641     -0.82     0.4126     0.05
b7           -0.6039    0.1231   2641     -4.90     <.0001     0.05   ***
_s            0.3657   0.03066   2641     11.93     <.0001     0.05   ***
c1           0.01383   0.06872   2641      0.20     0.8405     0.05
c2           -2.3440    0.6881   2641     -3.41     0.0007     0.05   ***
c3           0.04668   0.02469   2641      1.89     0.0588     0.05   *
c4            0.1219    1.2489   2641      0.10     0.9223     0.05
c5          0.001200  0.002845   2641      0.42     0.6732     0.05
c6          -0.02245   0.01167   2641     -1.92     0.0546     0.05   *
c7            1.5452    0.4678   2641      3.30     0.0010     0.05   ***
*/
```

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

2. NLS Regression Model

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

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

Therefore, the conditional mean of ***Y*** can be represented as ***1 / [1 + EXP (-X`β)]***. Similar to OLS or Tobit regression, NLS is also subject to the homoscedastic assumption. As a result, a sub-model is also needed to account for the heteroscedasticity in a similar way to what has been done in the previous section.

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

Again, for the computational reason, a simpler NLS regression assuming the constant variance would be estimated first in order to obtain a set of reasonable starting values for parameter estimates, as shown below.

```
ods output parameterestimates = _parm1;
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;
```

```
  xb = b0 + b1 * x1 + b2 * x2 + b3 * x3 + b4 * x4 + b5 * x5 + b6 * x6 + b7 * x7;
  mu = 1 / (1 + exp(-xb));
  lh = pdf('normal', y, mu, _s);
  ll = log(lh);
  model y ~ general(ll);
run;

proc sql noprint;
  select parameter||" = "||compress(put(estimate, 18.4), ' ')
  into :parms separated by ' ' from _parm1;
quit;

proc nlmixed data = data.deve tech = trureg;
  parms &parms  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|    Alpha
b0           -7.4915     0.4692   2641    -15.97     <.0001     0.05   ***
b1          -0.04652    0.03268   2641     -1.42     0.1547     0.05
b2            0.8447     0.2125   2641      3.98     <.0001     0.05   ***
b3            0.4098    0.03315   2641     12.36     <.0001     0.05   ***
b4           -3.3437     0.6229   2641     -5.37     <.0001     0.05   ***
b5          0.001015   0.001341   2641      0.76     0.4489     0.05
b6          -0.00914   0.002853   2641     -3.20     0.0014     0.05   ***
b7           -1.1170     0.2910   2641     -3.84     0.0001     0.05   ***
_s          -0.01499   0.002022   2641     -7.41     <.0001     0.05   ***
c1          -0.05461    0.01310   2641     -4.17     <.0001     0.05   ***
c2            0.4066     0.1347   2641      3.02     0.0026     0.05   ***
c3            0.4229    0.02035   2641     20.78     <.0001     0.05   ***
c4           -3.6905     0.3187   2641    -11.58     <.0001     0.05   ***
c5          0.001291   0.000842   2641      1.53     0.1255     0.05
c6          -0.01644   0.002053   2641     -8.01     <.0001     0.05   ***
c7           -1.0388     0.1332   2641     -7.80     <.0001     0.05   ***
*/
```

In the above output, most predictors are statistically significant in both the mean and the variance sub-models, showing a strong evidence of heteroscedasticity.

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 $E(Y|X) = G(X`\beta)$

= 1 / [1 + EXP (-X`β)], Fractional Logit model has the identical likelihood function to the one for a Bernoulli distribution such that

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

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 Fractional Logit model is with GLIMMIX procedure. 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|    Alpha
b0           -7.3467     0.7437   2641     -9.88     <.0001     0.05   ***
b1          -0.05820    0.06035   2641     -0.96     0.3349     0.05
b2            0.8480     0.3276   2641      2.59     0.0097     0.05   ***
b3            0.3996    0.05151   2641      7.76     <.0001     0.05   ***
b4           -3.4801     1.0181   2641     -3.42     0.0006     0.05   ***
b5          0.000910   0.002027   2641      0.45     0.6534     0.05
b6          -0.00859   0.005018   2641     -1.71     0.0871     0.05   *
b7           -1.0455     0.4403   2641     -2.37     0.0176     0.05   **
*/
```

It is worth mentioning that Fractional Logit model can be easily transformed to Weighted Logistic regression with binary outcomes (shown below), which would yield almost identical parameter estimates and statistical inferences. As a result, most model development techniques and statistical diagnostics used in Logistic regression can 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
                               Standard          Wald
Parameter    DF    Estimate       Error    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
*/
```

Two-Part Composite Models

In the preliminary data analysis, it's been shown that ~75% businesses in the study carried no debt at all. Therefore, it might be appealing to employ zero-inflated fractional models, a Logit model separating zero outcomes from positive proportional 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

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

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

In this paper, Beta and Simplex model would be used to analyze nonzero proportional 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 would take the debt, then it should further decide how much to borrow.

1. Beta Model

Beta regression is a flexible modeling technique based upon the two-parameter beta distribution and can be employed to model any dependent variable that is continuous and bounded by two known endpoints, e.g. 0 and 1 in our context. Assumed 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

$$F(Y) = {Gamma(\omega + \tau)} \Big/ {\left(Gamma(\omega) \times Gamma(\tau)\right)} \times Y^{\omega-1} \times (1-Y)^{\tau-1}$$

In the above function, while ω is pulling the density toward 0, τ is pulling the density toward 1. Without the loss of generality, ω and τ can be re-parameterized and translated into two other parameters, namely location parameter μ and dispersion parameter φ such that $\omega = \mu \times \varphi$ and $\tau = \varphi \times (1 - \mu)$, where μ is the expected mean and φ governs the variance such that

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

Within the framework of GLM (generalized linear models), μ and φ can be modeled separately with a location sub-model for μ and the other dispersion sub-model for φ using two different or identical sets of covariates X and Z. Since the expected mean μ is bounded by 0 and 1, a natural choice of the link function for location sub-model is Logit function such that $LOG\ [\mu / (1 - \mu)] = X`\beta$. With the strictly positive nature of $\varphi$, Log function seems appropriate to serve our purpose such that $LOG\ (\varphi) = Z`\gamma$.

SAS does not provide an out-of-box procedure to estimate the two-parameter Beta model formulated as above. While GLIMMIX procedure is claimed to accommodate Beta modeling, it can only estimate a simple-form model without the dispersion sub-model. However, with the probability function of Beta distribution, it is straightforward to estimate the Beta model with NLMIXED procedure by explicitly specifying the log likelihood function as below.

```
ods output parameterestimates = _parm1;
proc nlmixed data = data.deve tech = trureg maxiter = 500;
  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;
  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;
  mu_xa = 1 / (1 + exp(-xa));
  mu_xb = 1 / (1 + exp(-xb));
  phi = exp(c0);
  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;

proc sql noprint;
  select parameter||" = "||compress(put(estimate, 18.4), ' ')
  into :parm1 separated by ' ' from _parm1;
quit;

proc nlmixed data = data.deve tech = trureg;
  parms &parm1  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|    Alpha
a0           -9.5003    0.5590   2641    -17.00     <.0001     0.05   ***
a1          -0.03997   0.03456   2641     -1.16     0.2476     0.05
a2            1.5725    0.2360   2641      6.66     <.0001     0.05   ***
a3            0.6185   0.03921   2641     15.77     <.0001     0.05   ***
a4           -2.2842    0.6445   2641     -3.54     0.0004     0.05   ***
a5          -0.00087  0.001656   2641     -0.52     0.6010     0.05
a6          -0.00530  0.003460   2641     -1.53     0.1256     0.05
a7           -1.5349    0.3096   2641     -4.96     <.0001     0.05   ***
b0            1.6136    0.4473   2641      3.61     0.0003     0.05   ***
b1          -0.02592   0.03277   2641     -0.79     0.4290     0.05
b2           -0.3756    0.1781   2641     -2.11     0.0351     0.05   **
b3           -0.1139   0.03017   2641     -3.77     0.0002     0.05   ***
b4           -2.7927    0.5133   2641     -5.44     <.0001     0.05   ***
b5          0.003064  0.001527   2641      2.01     0.0448     0.05   **
b6          -0.00439  0.002475   2641     -1.77     0.0764     0.05   *
b7            0.2253    0.2434   2641      0.93     0.3548     0.05
c0           -0.2832    0.5877   2641     -0.48     0.6300     0.05
c1          -0.00171   0.04219   2641     -0.04     0.9678     0.05
c2            0.6073    0.2311   2641      2.63     0.0086     0.05   ***
c3           0.07857   0.03988   2641      1.97     0.0489     0.05   **
c4            2.2920    0.7207   2641      3.18     0.0015     0.05   ***
c5          -0.00435  0.001643   2641     -2.65     0.0081     0.05   ***
c6          0.001714  0.003388   2641      0.51     0.6130     0.05
c7          -0.09279    0.3357   2641     -0.28     0.7823     0.05
*/
```

As shown above, since there are three sets of parameters to be estimated in the zero-inflated Beta model, it is a good practice to start with a simpler form assuming that the dispersion parameter $\boldsymbol{\varphi}$ is a constant and estimating two sets of parameters for mean models first, which works very well empirically.

2. Simplex Model

The last one introduced, which is called Simplex model, might be a “new kid in town” for most of statisticians and can be 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

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

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

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

$$\boldsymbol{D}(\boldsymbol{Y}) = {}^{(\boldsymbol{Y}-\boldsymbol{\mu})^{2}}\!\Big/_{\boldsymbol{Y} \times (\mathbf{1}-\boldsymbol{Y}) \times \boldsymbol{\mu}^{\mathbf{2}} \times (\mathbf{1}-\boldsymbol{\mu})^{\mathbf{2}}}$$

Similar to the Beta model, a simplex model also consists of two components, a sub-model estimating the expected mean $\boldsymbol{\mu}$ 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 ***X*** such that ***LOG [μ / (1 – μ)] = X`β***. Also because of the strict positivity of $\boldsymbol{\sigma^2}$, the sub-model for dispersion parameter $\boldsymbol{\sigma}$ can be formulated as ***LOG*** $\boldsymbol{(\sigma^2) = Z\grave{}\gamma}$.

Currently, 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 as given below.

```
ods output parameterestimates = _parm1;
proc nlmixed data = data.deve tech = trureg;
  parms a0 = 0  a1 = 0  a2 = 0  a3 = 0  a4 = 0  a5 = 0  a6 = 0  a7 = 0;
  xa = a0 + a1 * x1 + a2 * x2 + a3 * x3 + a4 * x4 + a5 * x5 + a6 * x6 + a7 * x7;
  mu_xa = 1 / (1 + exp(-xa));
  if y = 0 then y2 = 0;
  else y2 = 1;
  lh = (mu_xa ** y2) * ((1 - mu_xa) ** (1 - y2));
  ll = log(lh);
  model y ~ general(ll);
run;

proc sql noprint;
  select parameter||" = "||compress(put(estimate, 18.4), ' ')
  into :parm1 separated by ' ' from _parm1;
quit;

ods output parameterestimates = _parm2;
proc nlmixed data = data.deve tech = trureg;
  parms &parm1  b0 = 0  b1 = 0  b2 = 0  b3 = 0  b4 = 0  b5 = 0  b6 = 0  b7 = 0  c0 = 4;
  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;
  mu_xa = 1 / (1 + exp(-xa));
  mu_xb = 1 / (1 + exp(-xb));
  s2 = exp(c0);
  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);
    v = (y * (1 - y)) ** 3;
    lh = mu_xa * (2 * constant('pi') * s2 * v) ** (-0.5) * exp(-(2 * s2) ** (-1) * d);
    ll = log(lh);
  end;
  model y ~ general(ll);
run;

proc sql noprint;
  select parameter||" = "||compress(put(estimate, 18.4), ' ')
  into :parm2 separated by ' ' from _parm2;
```

```
quit;

proc nlmixed data = data.deve tech = trureg;
  parms &parm2  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);
  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);
    v = (y * (1 - y)) ** 3;
    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|    Alpha
a0           -9.5003     0.5590   2641    -17.00     <.0001     0.05   ***
a1          -0.03997    0.03456   2641     -1.16     0.2476     0.05
a2            1.5725     0.2360   2641      6.66     <.0001     0.05   ***
a3            0.6185    0.03921   2641     15.77     <.0001     0.05   ***
a4           -2.2842     0.6445   2641     -3.54     0.0004     0.05   ***
a5          -0.00087   0.001656   2641     -0.52     0.6010     0.05
a6          -0.00530   0.003460   2641     -1.53     0.1256     0.05
a7           -1.5349     0.3096   2641     -4.96     <.0001     0.05   ***
b0           -0.5412     0.4689   2641     -1.15     0.2485     0.05
b1           0.03485    0.02576   2641      1.35     0.1763     0.05
b2           -1.3480     0.2006   2641     -6.72     <.0001     0.05   ***
b3           0.01708    0.03098   2641      0.55     0.5814     0.05
b4           -2.0596     0.5731   2641     -3.59     0.0003     0.05   ***
b5          0.004635   0.001683   2641      2.75     0.0059     0.05   ***
b6          -0.00006   0.002652   2641     -0.02     0.9818     0.05
b7            0.7973     0.2945   2641      2.71     0.0068     0.05   ***
c0            9.9250     0.5582   2641     17.78     <.0001     0.05   ***
c1           -0.1034    0.04846   2641     -2.13     0.0329     0.05   **
c2            1.6217     0.2960   2641      5.48     <.0001     0.05   ***
c3           -0.4550    0.03652   2641    -12.46     <.0001     0.05   ***
c4           -4.1401     0.8523   2641     -4.86     <.0001     0.05   ***
c5          0.007653   0.002079   2641      3.68     0.0002     0.05   ***
c6          -0.00742   0.003526   2641     -2.11     0.0354     0.05   **
c7           -0.6699     0.4484   2641     -1.49     0.1353     0.05
*/
```

Model Evaluations

In previous sections, five models for proportional outcomes have been demonstrated 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, Parameter Estimates of Five Models (mean models only)

Prameter Estimates	1-Piece Model			2-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 estimates with p-values lower than 0.01 are highlighted. It is shown that the negative relationship between ***X4*** (profitability) and the financial leverage is significant and consistent across all five models. It is interesting to notice that both ***X2*** (collateral) and ***X3*** (size) have consistent and significant positive impacts on the financial leverage in all 1-piece models. However, the story differs in 2-part models. For instance, in the ZI (zero-inflated) Beta model, while large-size firms might be more likely to borrow, there is however a negative relationship between the size of a business and the leverage ratio given a decision made to raise the debt. Similarly in the ZI 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 percent and the leverage ratio conditional on the decision of borrowing. These are all interesting observations worth further investigations.

To compare multiple models with different distributional assumptions, academic statisticians might prefer to use likelihood-based approaches such as Vuong or nonparametric Clarke test (Clarke, 2007). However, from a practical perspective, it might be more intuitive to use the empirical measures such as Information Value or R^2 calculated from the separate hold-out data sample, as shown below.

Table 6, 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 is clear that ZI Beta yields the best performance in terms of R^2, followed by Fractional Logit model. Moreover, the 2-part nature of ZI Beta model might be able to provider more intriguing insights for further discussions. However, due to the difficult implementation, applying ZI Beta model to real-world problems might present more troubles than benefits for many practitioners. Therefore, Fractional Logit model might be often preferred in reality for the sake of simplicity.

Conclusion

In this paper, five different modeling strategies for proportional outcomes in the [0, 1] interval have been surveyed. An example in financial leverage has been used to illustrate implementations of various models in SAS. In real-world business problems, it is highly recommended that practitioners should start with Fractional Logit model due to simple implementations and liberal assumptions and then might look for further improvements from more complex models such as ZI Beta model.

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. Clarke, K. (2007), A Simple Distribution-Free Test for Nonnested Model Selection, Political Analysis, Vol. 15 Issue 3, 347 - 363

Contact Information

WenSui Liu, Portfolio Analysis & Forecasting Manager, VP

Fifth Third Bancorp, Cincinnati, OH

Email: wensui.liu@53.com

Kelly Zhao, Credit Risk Analytics Manager, VP

Fifth Third Bancorp, Cincinnati, OH

Email: xia.zhao@53.com