AARHUS UNIVERSITY

Cover sheet

This is the accepted manuscript (post-print version) of the article.

The content in the accepted manuscript version is identical to the final published version, although typography and layout may differ.

How to cite this publication

Please cite the final published version:

Wulff, J. (2019). Generalized two-part fractional regression with cmp. *Stata Journal*, 19(2), 375–389.
<https://doi.org/10.1177/1536867X19854017>

Publication metadata

Title:	Generalized two-part fractional regression with cmp
Author(s):	Jesper N. Wulff
Journal:	Stata Journal, 19(2), 375–389
DOI/Link:	10.1177/1536867X19854017
Document version:	Accepted manuscript (post-print)

General Rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

If the document is published under a Creative Commons license, this applies instead of the general rights.

Generalized two-part fractional regression with `cmp`

Jesper N. Wulff
Aarhus University
Aarhus, Denmark
jwulff@econ.au.dk

Abstract. Researchers who model fractional dependent variables often need to consider whether their data were generated by a two-part process. Two-part models are ideal for modeling two-part processes as they allow us to model the participation and magnitude decision separately. While user-written commands currently facilitate estimation of two-part models, no specialized command exists for fitting two-part models with process dependency. In this paper, I describe generalized two-part fractional regression, which allows for dependency between models parts. I show how this model can be estimated using the user-written `cmp` command (Roodman 2011, *Stata Journal* 11: 159-20). I use a data example on the financial leverage of firms to illustrate how `cmp` can be used to fit generalized two-part fractional regression. Further, I show how to obtain predicted values of the fractional dependent variable and marginal effects useful for model interpretation. Finally, I show how to compute model fit statistics and perform the RESET test, which are useful for model evaluation.

Keywords: `stxxxx`, generalized two-part fractional regression, process dependence, fractional probit, `cmp`

1 Introduction

In many different disciplines, researchers need to estimate regression models where the dependent variable is in the form of a fraction, percentage, or proportion. In finance, a commonly examined fractional dependent variable (FDV) is the financial leverage ratio of firms, i.e. the amount of debt a firm issues relative to its amount of capital. Empirical research suggests that the financial leverage decision of firms is best described as a two-step process: First, the firm decides whether or not to issue debt and then it decides how much debt to issue (Ramalho and da Silva 2009). However, the process determining which firms choose to issue debt is non-random as firms self-select into a leveraged position. Thus, we need a model that can not only separate the effects on the debt-vs-no-debt from the effects on the amount-of-debt decision, but can also take into account the non-random selection that leads to some firms issuing debt. The generalized two-part fractional regression model (GTP-FRM) is such a model.

Before we dig into the GTP-FRM, we need a short introduction on modeling FDVs. Modeling an FDV requires a fractional regression model (FRM). If we use the Quasi-Maximum Likelihood Estimator (QMLE) and the logit link, the model is known as the fractional logit (FL) or fractional probit (FP) if we use the probit (Papke and Wooldridge

1996). An FRM is preferable because it ensures predictions within the unit interval and only requires correct specification of the conditional mean. Estimation of FRMs with various link functions is straight-forward in Stata using the `glm` command and has become even more accessible with the Stata 14 addition of the `fracreg` command.

In many cases, an FDV may contain a substantial amount of values at one or both boundaries. The values at e.g. the zero boundary may be governed by a different process than the values between the boundaries. For instance, the covariates that affect a firm's decision to issue debt are likely to be different from those affecting how much debt to issue (Cook et al. 2008). For such purposes, Ramalho and da Silva (2009) proposed a two-part fractional model (TP-FRM). The TP-FRM allows for specification of a binary model for the participation decision ($y = 0$ vs. $y > 0$), e.g. debt-vs-no-debt and an FRM for the magnitude decision (the magnitude of y when $y > 0$), e.g. how much debt to issue. Using this model, we allow the effects of a covariate on the participation decision to be different than on the magnitude decision. In Stata, TP-FRMs can be estimated using the user-written `frm` (Ramalho et al. 2011) or `tpm` (Belotti et al. 2015) commands.

In some cases, we may need even more flexibility than what is offered by the TP-FRM. This happens when the participation and magnitude decisions are dependent. For instance, continuing the example from above, there may be a selection bias in the types of firms that choose to issue debt. This problem is analogous to the well-known sample selection problem. Recently, Schwiebert and Wagner (2015) proposed the GTP-FRM as a means to model fractional two-part processes with dependence. The GTP-FRM formulation is advantageous because it nests the TP-FRM as a special case when the two processes are independent.

Currently, the GTP-FRM does not have a dedicated Stata command. In this paper, I demonstrate how Stata users can estimate GTP-FRMs using the conditional mixed-process framework implemented by the `cmp` (Roodman 2011) command. In section 2, I briefly describe the GTP-FRM. In the form of a short tutorial, section 3 shows how `cmp` can be used to estimate the GTP-FRM and demonstrates how to compute predictions, marginal effects, information criteria and RESET test statistics. In section 4, I provide a short conclusion.

2 Brief review of generalized two-part fractional regression

In many practical applications, FDVs naturally give rise to a large amount of zeroes. Some firms do not have any exports (0% of total sales attributed to foreign sales), some individuals do not smoke (0% of their income spent on cigarettes), and some firms do not issue any debt (0% leveraged capital). When encountering such FDVs, researchers have to decide how to qualitatively interpret the zeroes (Ramalho et al. 2011). The reason is that the zeros may actually be best described by a different mechanism than the positive values. Consider, for instance, the financial leverage decisions of firms.

While a financing life-cycle approach would argue that older firms are more likely to issue debt, pecking-order theory would suggest that older firm prefer a lower proportion of debt due to a large amount of accumulated retained earnings (Ramalho and da Silva 2009). In such cases, a TP-FRM is ideal as it allows us to model the participation and magnitude decision separately.

If a two-part process constitutes the best description of our data, we need to consider whether the two decisions are dependent. If unobserved factors affecting the decision to issue debt are correlated with factors that influence the proportion of debt sought by firms, the TP-FRM estimates will be biased (Wooldridge 2010). For instance, research suggests that a moderate degree CEO narcissism is associated with a substantial increase in corporate risk taking (Aabo and Eriksen 2018). If CEO narcissism is related to firm profitability and not accounted for, the estimate of firm profitability on the proportion of debt is likely to be biased. The GTP-FRM models the correlation between the two decisions, thus attempting to both estimate and adjust for the dependency between the processes. This makes the GTP-FRM ideal for FDVs best described by dependent two-part processes.

2.1 Model formulation

Following Schwiebert and Wagner (2015), we start by specifying the process determining the participation decision, where we assume an FDV, y , with values in the unit interval:

$$s = 1(z'\beta_1 + u > 0) \quad (1)$$

where s is an indicator variable that takes the value 1 if the value of the outcome, y , is non-zero; z is a vector of covariates affecting the participation decision, β_1 is a vector of coefficients and u is the error term. To model the participation decision in (1), we use a probit model specification:

$$\Pr(s = 1|z) = \Phi(z'\beta_1) \quad (2)$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. We specify the conditional mean of the magnitude decision in the following way:

$$E(y|x, z, s = 0) = 0 \quad (3a)$$

$$E(y|x, z, s = 1) = \frac{\Phi_2(x'\beta_2, z'\beta_1; \rho)}{\Phi(x'\beta_2)} \quad (3b)$$

where x is a vector of covariates affecting the magnitude decision; β_2 is a coefficient vector with respect to x and $\Phi_2(\cdot; \rho)$ denotes the bivariate standard normal distribution with ρ representing the correlation between the participation and magnitude decision. Equation (3b) is the fractional probit specification, which is used to model non-zero values of y . If the two processes in equation (3b) are independent, i.e. $\rho = 0$, the GTP-FRM reduces to the simpler TP-FRM, where $E(y|x, z, s = 1) = \Phi(x'\beta_2)$. However, if $\rho > 0$, the TP-FRM is misspecified. The larger the dependence between the two processes, the larger the bias we would expect by using a TP-FRM.

Based on the above formulation, it is clear that the GTP-FRM is a more complicated model than the TP-FRM. In contrast to the TP-FRM, the GTP-FRM lets both x and z affect the magnitude decision: While x has a direct effect, the effect of z is indirect through s (Schwiebert and Wagner 2015). This extra complexity does not come for free. Indeed, the GTP-FRM needs an exclusion restriction in order to be identified, i.e. a variable affecting the participation decision without directly affecting the amount decision. It is important to note that even though we may be able to estimate a GTP-FRM without an exclusion restriction in practice, we should not trust its estimates (Sartori 2003).

2.2 Marginal effects

As for the regular binary probit model, we should abstain from interpreting the model coefficients directly. Instead, we should rely on marginal effects preferably accompanied by graphical illustrations of predicted values of the FDV (Wulff 2015). As shown by Schwiebert and Wagner (2015), the predicted values of the FRV given values of the covariates are given by

$$E(y|x, z) = E(y|x, z, s = 1) \Pr(s = 1|z) \quad (4a)$$

$$= \Phi_2(x'\beta_2, z'\beta_1; \rho) \quad (4b)$$

Note as for other two-part models, it is possible to obtain other types of predictions based on the GTP-FRM. For instance, we can compute the expected value of y conditional on $y > 0$, i.e. the term $E(y|x, z, s = 1)$.

Based on equation 4b, the marginal effect of x_k in the GTP-FRM is given by

$$\frac{\partial E(y|x, z)}{\partial x_k} = \frac{\partial \Phi_2(x'\beta_2, z'\beta_1; \rho)}{\partial x_k} \quad (5)$$

As I illustrate later, marginal effects and predicted values of the FDV can be obtained using `margins` after estimating the model using `cmp`. Especially, we will need to use equation 4a to obtain the predicted values of the FRV and the corresponding marginal effects.

2.3 Model fit measures

When comparing various fractional model specifications, we can rely on the information criteria AIC and BIC . Comparing measures based on pseudo-likelihoods from different models can be tricky. Papke and Wooldridge (1996) argue for defining R^2 in terms of the actual and predicted values. Like R^2 , information criteria can also be defined in terms of the residual sum of squares (RSS). Thus, I suggest using the following definition of

AIC and *BIC* when comparing FRMs, TP-FRMs and GTP-FRMs:

$$AIC = \log \left(\frac{RSS}{n} \right) + \frac{2K}{n} \quad (6a)$$

$$BIC = \log \left(\frac{RSS}{n} \right) + \frac{K \log(n)}{n} \quad (6b)$$

where n is the sample size and K is the total number of parameters from each part of the GTP-FRM. Defining information criteria in this way has two major advantages when working with FDVs. First, they are comparable across any model for the conditional mean and for any estimation method (Papke and Wooldridge 1996). For instance, we can compare our GTP-FRM to its Tobit equivalent - the exponential Type II Tobit model for corner solution responses (Wulff and Villadsen 2018). Second, we do not have to worry about unintentionally comparing information criteria across models with different (pseudo) likelihood functions.

2.4 RESET test

For the GTP-FRM, we can employ Ramsey's (1969) RESET test as a simple functional form diagnostic. Essentially, the test can be used to check for missing non-linearities in the GTP-FRM or any other index model (Pagan and Vella 1989). For each model part, we obtain the index predictions. These are added in squared and cubic form to the relevant model part after which we can perform a joint hypothesis test of the coefficients of the two extra terms (Ramalho and Ramalho 2012). For the fractional part, it is important that we use the robust version of the test. However, as robust standard errors are computed by default by `fracreg probit`, this is of no concern to the user when following the implementation I suggest below.

3 Stata implementation

The FRM, TP-FRM and the GTP-FRM can be estimated using Quasi-Maximum Likelihood Estimation (QMLE). As noted in the introduction, the FP is already implemented in the `fracreg probit` routine. This implementation and the regular binary probit model are both available in the `cmp` framework. Thus, we can use the power of `cmp` not only to estimate GTP-FRMs but to also compute marginal effects and predictions.

3.1 Data example

To illustrate the use of `cmp`, I rely on data from the study on financial leverage decisions by Ramalho and da Silva (2009). The data are available for download in a `.txt` file-format here and can be loaded into Stata using `import delimited`. The dataset contains information on several firm characteristics. For this example, I will rely on `leverage` (long-term debt to long-term capital assets), `size` (natural logarithm of sales), and `tangibility` (sum of tangible assets and inventories, divided by total

assets). A complete description of the data is available in Ramalho and da Silva (2009).

. summarize leverage size tangibility					
Variable	Obs	Mean	Std. Dev.	Min	Max
leverage	1,295	.1483878	.1988493	0	.9779415
size	1,295	15.81369	1.385884	11.73562	22.26972
tangibility	1,295	.3770338	.1967606	.0015893	.9774859

I start by generating a non-zero indicator variable indicating the firms that have issued debt. To simplify the syntax, I assign the regressors to a global macro. Note that I exclude the `size` variable from the list as we are going to use it as an exclusion restriction below. Valid exclusion restrictions are notoriously hard to come by and this example is not different. By using `size` as an exclusion restriction, we are assuming that firm size is directly related to the participation decision but only related to the amount decision through the participation decision. In other words, firm size does directly affect the proportion of debt sought by firms. While this assumption may be questionable, it will do for the purpose of this illustration.

```
. gen s = leverage > 0 // non-zero indicator variable
. global regressors ndts tangibility profitability growth age liquidity ///
> manufacturing construction trade communication
```

3.2 Model estimation strategy

Some readers might have noticed how the GTP-FRM looks conceptually similar to the Heckman (1976) sample selection model. In fact, the GTP-FRM is also applicable to fractional missing data problems (Schwiebert and Wagner 2015). In the current setting, however, we do not have a sample selection issue. Instead, we have a fractional response where the FDV is always observed, yet generated by two different dependent processes.

Conveniently, we can exploit the similarities to the sample selection model by using the same approach to estimate the GTP-FRM. We do this by estimating systems of equations with errors that are jointly normally distributed. If our FDV had been binary, this estimation could have been performed by using the `heckprobit` command. As `cmp` allows for QMLE with a probit link, we can use `cmp` to estimate our parameters from the participation and magnitude equations simultaneously while obtaining an estimate of and accounting for ρ . In this way, `cmp` allows the participation and decision equation to vary by observation, and enables consistent and efficient estimation. Thus, the estimation procedure becomes similar to that of the `heckprobit` command.

3.3 Syntax

The syntax for `cmp` is thoroughly described in Roodman (2011) and the `cmp` help file. Thus, in line with the focus of this paper, I focus on its application to GTP-FRMs.

The `cmp setup` subcommand defines global macros that we use in the command line.

In the command line, I specify the two equations using `size` as an exclusion restriction in the participation part. In `indicators`, I specify the fractional probit using `cmp_frac` and the regular probit using `cmp_probit`. Note that I multiply `cmp_frac` by the non-zero indicator variable, `s`. I use `quietly` to suppress most model output.

```
. cmp setup
$cmp_out      = 0
$cmp_missing  = .
$cmp_cont     = 1
$cmp_left     = 2
$cmp_right    = 3
$cmp_probit   = 4
$cmp_oprobit  = 5
$cmp_mprobit  = 6
$cmp_int      = 7
$cmp_trunc    = 8 (deprecated)
$cmp_roprobit = 9
$cmp_frac     = 10
```

```
. cmp (leverage = $regressors) (s = size $regressors), ///
> ind(s*$cmp_frac $cmp_probit) quietly
```

Note: fractional probit models imply `vce(robust)`.

Fitting individual models as starting point for full model fit.

Note: For programming reasons, these initial estimates may deviate from your specification.
For exact fits of each equation alone, run `cmp` separately on each.

Fitting full model.

```
Mixed-process regression          Number of obs    =      1,295
                                Wald chi2(21)       =      216.81
Log pseudolikelihood = -1190.6853  Prob > chi2      =      0.0000
```

	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
leverage						
ndts	-.0536411	.0265636	-2.02	0.043	-.1057048	-.0015774
tangibility	.6278901	.1894899	3.31	0.001	.2564967	.9992835
profitability	-2.502263	.3480411	-7.19	0.000	-3.184411	-1.820115
growth	.0038992	.0014633	2.66	0.008	.0010312	.0067671
age	-.0007824	.0011114	-0.70	0.481	-.0029606	.0013959
liquidity	-.6594569	.2718519	-2.43	0.015	-1.192277	-.126637
manufacturing	-.1415292	.1083216	-1.31	0.191	-.3538357	.0707773
construction	.0186647	.1369112	0.14	0.892	-.2496764	.2870058
trade	-.329909	.2714891	-1.22	0.224	-.8620179	.2021999
communication	.0456521	.14377	0.32	0.751	-.2361319	.327436
_cons	-.6295222	.1755718	-3.59	0.000	-.9736366	-.2854078
s						
size	.2159975	.0327738	6.59	0.000	.1517621	.280233
ndts	-.0470524	.0243417	-1.93	0.053	-.0947612	.0006564
tangibility	1.526855	.2211802	6.90	0.000	1.09335	1.96036
profitability	-2.266795	.4896678	-4.63	0.000	-3.226526	-1.307064
growth	.0051865	.0018226	2.85	0.004	.0016143	.0087587
age	-.0009803	.0019149	-0.51	0.609	-.0047334	.0027728
liquidity	-1.504778	.2610957	-5.76	0.000	-2.016517	-.9930402
manufacturing	-.1583662	.1738511	-0.91	0.362	-.4991081	.1823756
construction	-.2397432	.2089277	-1.15	0.251	-.649234	.1697475
trade	-.7439219	.3545159	-2.10	0.036	-1.43876	-.0490834
communication	-.2948638	.2442249	-1.21	0.227	-.7735359	.1838083

_cons	-3.285265	.5689352	-5.77	0.000	-4.400358	-2.170173
/atanhrho_12	.6036901	.2445748	2.47	0.014	.1243323	1.083048
rho_12	.5396701	.1733439			.1236956	.7943266

```
. estimates store gt_frm
```

The first part of the output relates to the amount decision (**leverage**), while the second part of the output refers to the participation decision (**s**). The chosen estimation strategy provides a direct estimate of $\text{arctanh } \rho = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right)$. Automatically, a Wald test is performed rejecting the regular TP-FRM at $\alpha = 0.05$. Thus, a comparison with the TP-FRM is already incorporated in the procedure. Finally, we observe that the estimate of ρ is around 0.5. Thus, the Heckman type estimation makes it possible to estimate the magnitude of the dependency between the two processes.

For illustration, we can compare the GTP-FRM estimates to those of the TP-FRM and regular FRM using `esttab` (Jann 2005):

```
. qui fracreg probit leverage size $regressors
. estimates store frm
. qui probit leverage size $regressors
. estimates store tp_frm_participation
. qui fracreg probit leverage $regressors if leverage > 0
. estimates store tp_frm_magnitude
. esttab frm tp_frm_participation tp_frm_magnitude, ///
> nostar mlabel("FRM" "TP-FRM (1)" "TP-FRM (2)")
```

	(1) FRM	(2) TP-FRM (1)	(3) TP-FRM (2)
leverage			
size	0.0609 (3.34)	0.216 (7.58)	
ndts	-0.0584 (-2.31)	-0.0468 (-1.54)	-0.0397 (-1.47)
tangibility	0.849 (5.95)	1.521 (7.06)	0.210 (1.49)
profitabil-y	-2.470 (-7.14)	-2.250 (-4.55)	-1.818 (-5.41)
growth	0.00396 (3.22)	0.00517 (2.86)	0.00254 (1.76)
age	-0.000904 (-0.77)	-0.00100 (-0.53)	-0.00144 (-1.36)
liquidity	-0.975 (-4.77)	-1.509 (-5.90)	-0.236 (-1.15)
manufactur-g	-0.151 (-1.40)	-0.163 (-0.89)	-0.0946 (-0.94)
construction	-0.0386 (-0.28)	-0.243 (-1.12)	0.0985 (0.77)
trade	-0.419	-0.749	-0.0937

	(-1.55)	(-2.04)	(-0.36)
communicat-n	-0.0237 (-0.16)	-0.304 (-1.23)	0.129 (0.97)
_cons	-1.764 (-5.53)	-3.273 (-6.51)	-0.291 (-2.20)
<hr/>			
N	1295	1295	661
<hr/>			
t statistics in parentheses			

In the `cmp` output, we can observe a positive and significant association between `tangibility` and the amount decision. However, the corresponding estimate from the TP-FRM is substantially smaller and non-significant. In the FRM column, we can observe the impact of ignoring the two-part process. In this model, the estimated coefficient on `tangibility` is larger than the corresponding estimate in the second part of the two-part models, as it "mixes" the coefficients from the two processes.

3.4 Predictions and marginal effects

To substantially interpret the results from the GTP-FRM, we can compute the (average) predicted proportions at selected values of `tangibility` using `margins` and plot them using `marginsplot`. In `expression` we specify the conditional expectation as shown in equation 4a. Here, the first part is the predicted proportion of debt conditional on having debt and the second part is the predicted probability of issuing debt.

```
. quiet margins, at(tangibility = (0 (.05) 1)) ///
> expression(predict(pr eq(leverage) cond(0 ., eq(s)))*predict(pr eq(s)))
. marginsplot, noci ytitle("Predicted leverage") scheme(sj)
Variables that uniquely identify margins: tangibility
```

Figure 1 shows the average predicted proportion of debt for various levels of firm profitability. At a tangibility around 0, the GTP-FRM predicts a debt proportion around 0.08. For firms with higher ratio of tangible to total assets, the model predicts a dramatically higher debt proportion. For instance, for firms with roughly equal parts of tangible and intangible assets the GTP-FRM predicts an average debt proportion of around 0.16.

For computation of the average marginal effect on the conditional mean of `leverage`, we can invoke the `dydx` option:

```
. margins, dydx(tangibility) ///
> expression(predict(pr eq(leverage) cond(0 ., eq(s)))*predict(pr eq(s)))
Average marginal effects      Number of obs      =      1,295
Model VCE      : Robust
Expression      : predict(pr eq(leverage) cond(0 ., eq(s)))*predict(pr eq(s))
dy/dx w.r.t.    : tangibility
```

	Delta-method				
	dy/dx	Std. Err.	z	P> z	[95% Conf. Interval]

Figure 1: Marginsplot with predicted proportions

tangibility	.1894122	.0332206	5.70	0.000	.124301	.2545234
-------------	----------	----------	------	-------	---------	----------

We can observe that a one unit increase in `tangibility` is associated with an average increase in the proportion of debt by about 0.19. This estimate is significant at conventional alpha levels.

By slightly modifying the syntax above we can quite easily obtain other types of predictions if we wish to. For instance, we can compute the average marginal effect on the predicted probability of issuing debt by only specifying the `predict(pr eq(s))` option:

```
. margins, dydx(tangibility) predict(pr eq(s))
```

Average marginal effects	Number of obs	=	1,295
Model VCE : Robust			
Expression : Pr(s), predict(pr eq(s))			
dy/dx w.r.t. : tangibility			

	Delta-method				
	dy/dx	Std. Err.	z	P> z	[95% Conf. Interval]
tangibility	.5376247	.0728696	7.38	0.000	.3948029 .6804464

A one unit increase in `tangibility` is associated with a 0.54 average increase in the probability of issuing debt. Finally, we can compute the average marginal effect on the proportion of debt conditional on the firm having issued debt:

```
. margins, dydx(tangibility) predict(e eq(leverage) cond(0 ., eq(s)))
```

Average marginal effects	Number of obs	=	1,295
--------------------------	---------------	---	-------

```

Model VCE      : Robust
Expression     : E(leverage), predict(e eq(leverage) cond(0 ., eq(s)))
dy/dx w.r.t.   : tangibility

```

	Delta-method					
	dy/dx	Std. Err.	z	P> z	[95% Conf. Interval]	
tangibility	.1197657	.1783612	0.67	0.502	-.2298159	.4693472

The estimated average marginal effect is much smaller than the one on the conditional mean and non-significant. This suggests that for firms that have issued debt tangibility matters little for how much debt they choose to issue.

3.5 Model fit

While the procedure illustrated above automatically yields a test comparing the GTP-FRM to the TP-FRM, we may want to compare the GTP-FRM to other specifications. Using equations 6a and 6b we can compare the GTP-FRM to any fractional specification we like. For instance, we can compare the GTP-FRM to the much less complex FRM specification:

```

. *Estimate GTP-FRM and get predictions
. quiet cmp (leverage = $regressors) (s = size $regressors), ///
> ind(s*$cmp_frac $cmp_probit) quietly
vce(robust)
. qui predict y_hat_first, pr eq(leverage) cond(0 ., eq(s))
. qui predict y_hat_second, pr eq(s)
. qui gen y_exp = y_hat_first*y_hat_second
. *Compute RSS and information criteria
. local K = e(rank)
. qui gen RS = (leverage - y_exp)^2
. qui su RS
. local RSS = r(mean)
. local n = r(N)
. scalar aic_cmp = log(`RSS'/`n') + 2*`K'/`n'
. scalar bic_cmp = log(`RSS'/`n') + `K'*log(`n')/`n'
. drop RS
.
. *Estimate FRM and get predictions
. qui fracreg probit leverage $regressors
. predict pred
(option cm assumed)
. local K = e(rank)
. qui gen RS = (leverage - pred)^2
. qui su RS
. local RSS = r(mean)
. local n = r(N)
. *Compute RSS and information criteria

```

```

. scalar aic_frm = log(`RSS`/`n`) + 2*`K`/`n`
. scalar bic_frm = log(`RSS`/`n`) + `K`*log(`n`)/`n`
.
. *Display AIC and BIC
. di _newline "GTP-FRM AIC = " aic_cmp _newline "GTP-FRM BIC = " bic_cmp ///
> _newline _newline "FRM AIC = " aic_frm _newline "FRM BIC = " bic_frm
GTP-FRM AIC = -10.479378
GTP-FRM BIC = -10.383632
FRM AIC = -10.49064
FRM BIC = -10.446756

```

The *AIC* and *BIC* values indicate that the GTP-FRM is not improving its fit enough to make up for its added complexity. Thus, it would seem that the FRM provides a better trade-off between parsimony and fit than the GTP-FRM. As explained above, we can use this procedure to compare any model for the conditional mean as long as we can obtain *RSS* on the fractional scale.

3.6 RESET test

For testing the conditional mean specification, we can implement the RESET test. This can be done for each part separately using the following procedure:

```

. *get index predictions
. qui predict y_hat_frac, index eq(leverage)
. qui predict y_hat_bin, index eq(s)
. *compute second and third order terms
. qui gen y_hat_frac2 = y_hat_frac^2
. qui gen y_hat_frac3 = y_hat_frac^3
. qui gen y_hat_bin2 = y_hat_bin^2
. qui gen y_hat_bin3 = y_hat_bin^3
. *RESET test fractional part
. qui cmp (leverage = $regressors y_hat_frac2 y_hat_frac3) ///
> (s = size $regressors), ind(s*$cmp_frac $cmp_probit) quietly
vce(robust)
. di "RESET GTP-FRM frac part"
RESET GTP-FRM frac part
. test [leverage]y_hat_frac2 [leverage]y_hat_frac3
( 1) [leverage]y_hat_frac2 = 0
( 2) [leverage]y_hat_frac3 = 0
      chi2( 2) =    0.07
Prob > chi2 =    0.9675
. *RESET test binary part
. qui cmp (leverage = $regressors) ///
> (s = size $regressors y_hat_bin2 y_hat_bin3), ///
> ind(s*$cmp_frac $cmp_probit) quietly
vce(robust)
. di "RESET GTP-FRM bin part"
RESET GTP-FRM bin part
. test [s]y_hat_bin2 [s]y_hat_bin3
( 1) [s]y_hat_bin2 = 0

```

```
( 2)  [s]y_hat_bin3 = 0
      chi2( 2) =    8.22
      Prob > chi2 =   0.0164
```

At an alpha of 1%, neither tests reject the H_0 that the model specification is correct. In contrast, the RESET test rejects the H_0 that the first part of the TP-FRM with a probit link is correctly specified:

```
. qui probit leverage size $regressors
. qui predict y_hat_tpfrm_one
. qui gen y_hat_tpfrm_one2 = y_hat_tpfrm_one ^2
. qui gen y_hat_tpfrm_one3 = y_hat_tpfrm_one ^3
.
. qui fracreg probit leverage $regressors if leverage > 0
. qui predict y_hat_tpfrm_two
. qui gen y_hat_tpfrm_two2 = y_hat_tpfrm_two ^2
. qui gen y_hat_tpfrm_two3 = y_hat_tpfrm_two ^3
.
. qui probit leverage $regressors y_hat_tpfrm_one2 y_hat_tpfrm_one3
. di "RESET TP-FRM binary part"
RESET TP-FRM binary part
. test y_hat_tpfrm_one2 y_hat_tpfrm_one3
( 1)  [leverage]y_hat_tpfrm_one2 = 0
( 2)  [leverage]y_hat_tpfrm_one3 = 0
      chi2( 2) =   58.93
      Prob > chi2 =   0.0000
.
. qui fracreg probit leverage $regressors y_hat_tpfrm_two2 y_hat_tpfrm_two3 ///
>      if leverage > 0
. di "RESET TP-FRM frac part"
RESET TP-FRM frac part
. test y_hat_tpfrm_two2 y_hat_tpfrm_two3
( 1)  [leverage]y_hat_tpfrm_two2 = 0
( 2)  [leverage]y_hat_tpfrm_two3 = 0
      chi2( 2) =    1.27
      Prob > chi2 =   0.5300
```

4 Conclusion

FDVs are often modeled by researchers across many different disciplines. When such variables are best described by a two-part process with dependence, researchers should apply the GTP-FRM. Currently, no dedicated Stata command exists to fit this kind of fractional model. In this paper, I show how GTP-FRMs can be fit with the user-written `cmp` command. Using a data example on the financial leverage of firms, I demonstrate how Stata users can estimate GTP-FRMs and compute predictions, marginal effects, information criteria and the RESET test statistic. Thus, the `cmp` command is a very useful command for fitting fractional responses generated by two dependent processes.

5 References

- Aabo, T., and N. B. Eriksen. 2018. Corporate risk and the hump-back of CEO narcissism. *Review of Behavioral Finance* 10(3): 252–273. <https://www.emeraldinsight.com/doi/10.1108/RBF-07-2017-0070>.
- Belotti, F., P. Deb, W. G. Manning, E. C. Norton, F. Belotti, P. Deb, W. Manning, and E. Norton. 2015. twopm: Two-part models. *Stata Journal* 15(1): 3–20. https://econpapers.repec.org/article/tsjstataj/v_3a15_3ay_3a2015_3ai_3a1_3ap_3a3-20.htm.
- Cook, D. O., R. Kieschnick, and B. D. McCullough. 2008. Regression analysis of proportions in finance with self selection. *Journal of Empirical Finance* 15(5): 860–867. <http://www.sciencedirect.com/science/article/pii/S0927539808000145>.
- Heckman, J. J. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables. *Annals of Economic and Social Measurement* 5(4): 475–492.
- Jann, B. 2005. Making regression tables from stored estimates. *Stata Journal* 5(3): 288–308.
- Pagan, A., and F. Vella. 1989. Diagnostic tests for models based on individual data: A survey. *Journal of Applied Econometrics* 4(S1): S29–S59. <http://doi.wiley.com/10.1002/jae.3950040504>.
- Papke, L. E., 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. <http://doi.wiley.com/10.1002/%28SICI%291099-1255%28199611%2911%3A6%3C619%3A%3AAID-JAE418%3E3.0.CO%3B2-1>.
- Ramalho, E. A., and J. J. Ramalho. 2012. Alternative versions of the RESET test for binary response index models: A comparative study. *Oxford Bulletin of Economics and Statistics* 74(1): 107–130. <http://doi.wiley.com/10.1111/j.1468-0084.2011.00654.x>.
- Ramalho, E. A., J. J. Ramalho, and J. M. Murteira. 2011. Alternative estimating and testing empirical strategies for fractional regression models. *Journal of Economic Surveys* 25(1): 19–68. <http://doi.wiley.com/10.1111/j.1467-6419.2009.00602.x>.
- Ramalho, J. J., and J. V. da Silva. 2009. A two-part fractional regression model for the financial leverage decisions of micro, small, medium and large firms. *Quantitative Finance* 9(5): 621–636. <http://www.tandfonline.com/doi/abs/10.1080/14697680802448777>.
- Ramsey, J. B. 1969. Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis. *Journal of the Royal Statistical Society. Series B (Methodological)* 31(2): 350–371. <https://www.jstor.org/stable/2984219>.
- Roodman, D. 2011. Fitting fully observed recursive mixed-process models with cmp. *Stata Journal* 11(2): 159–206. <http://ageconsearch.umn.edu/record/180074?ln=en>.

- Sartori, A. E. 2003. An Estimator for Some Binary-Outcome Selection Models Without Exclusion Restrictions. *Political Analysis* 11(02): 111–138. <https://www.jstor.org/stable/25791717>.
- Schwiebert, J., and J. Wagner. 2015. A Generalized Two-Part Model for Fractional Response Variables with Excess Zeros. In *Beiträge zur Jahrestagung des Vereins für Socialpolitik 2015: Ökonomische Entwicklung*, B04–V2. <http://hdl.handle.net/10419/113059>.
- Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*, vol. 2nd. Cambridge MA: The MIT Press.
- Wulff, J. N. 2015. Interpreting results from the multinomial logit model: Demonstrated by foreign market entry. *Organizational Research Methods* 18(2): 300–325.
- Wulff, J. N., and A. R. Villadsen. 2018. The use of tobit regression for modelling proportions and percentages in international business research. *Aarhus University Business Statistics Working Paper*.

About the author

Jesper N. Wulff is Assistant Professor at the Department of Economics and Business Economics, Aarhus University. His teaching and research interests include applied statistics and various topics within management science.