Notes on Fractional Regression and Related Estimation Procedures

OLS Problems with "Bounded" Dependent Variables

OLS regression assumes the dependent variable is interval level and is measured on an unbounded scale (in either direction). Thus, strictly speaking, OLS regression is not appropriate for modeling bounded interval-level variables.

Fractional regression is a well-developed alternative for modeling bounded dependent variables.

It is similar to ordered logit regression in many respects but is more flexible in that the dependent variable can be measured as continuous over a defined bounded range.

While using OLS regression to model bounded dependent variables is a common practice, it can lead to a variety of problems.

Problems – Implications for Violating OLS Assumptions

Error terms are heteroskedastic

Error terms are non-normal

Error terms display symptoms of non-linearity and non-additivity

Results of "standard" significant tests are not trustworthy

Predictions can be inaccurate and even impossible (i.e., outside of logical bounds) because the assumptions that estimated effects are linear and additive is incorrect.

Problems – Rules of Thumb for Concerns

The problems with OLS will usually be moderate or even negligible when predictions for Y are in the range 0.2-0.8 (assuming Y is rescaled to 0-1, continuous)

The problems with OLS will potentially be serious when predictions for Y approach boundaries; that is, when predictions fall in the ranges 0.0-0.2 or 0.8-1.0 (assuming Y is rescaled to 0-1, continuous)

Fractional (Logit) Regression

Fractional regression is similar to ologit regression in the sense that it can be used to model a variable that takes values within a bounded range. A key difference is that Y can be measured as continuous and does not need to be converted to categories.

Regular linear regression (OLS) could be used to model Y, but it will not keep predictions in bounds and OLS assumptions will be violated making statistical tests untrustworthy.

Beta regression is an alternative to fractional regression for modeling a continuous variable within a bounded range. But it cannot deal with cases that fall on the boundaries of Y. Additionally,

beta regression is a more complicated procedure because it simultaneously models both the mean of the distribution and the dispersion (variance) in the distribution.

Fractional regression has an especially attractive quality. When OLS regression is "okay", fractional regression will near-exactly reproduce the OLS results for predictions and significance tests. Thus, fractional regression will not lead to misleading findings (in comparison to OLS) when modeling bounded dependent variables. In contrast, OLS regression can lead to misleading findings.

The following figure provides a more careful review of these options.

Comparison of Options for Modeling Bounded Variables

Characteristic	Ordered Logit Regression	Fractional Regression	Beta Regression	OLS Linear Regression
Stata Procedure	ologit	fracreg (or glm)	betareg	regress
Lower bound	Rank value for lowest category of Y	Lowest logically possible score for Y (rescaled to 0)	Lowest logically possible score for Y (rescaled to ~0)	Y's as observed (or rescaled so logical minimum is 0)
Upper bound	Rank value for highest category of Y	Highest logically possible score for Y (rescaled to 1)	Highest logically possible score for y (rescaled to ~1)	Y's as observed (or rescaled so logical maximum is 1)
Intermediate Scores (between lower and upper boundaries)	Discrete rank values for intermediate categories on Y	Values over the bounded range *including* end points	Values over the bounded range *excluding* end points	Values over the bounded range *including* end points
Scaling	Integers	Continuous	Continuous	Continuous
Measurement level	Ordinal	Interval	Interval	Interval
Modeling goal	Predict relative frequency distribution for categories of Y over values of X's	Predict mean of Y over values of X's	Predict mean of Y *and* dispersion of Y over values of X's	Predict mean of Y over values f X's
Estimation Method	Maximum Likelihood	Quasi-Maximum Likelihood	Maximum Likelihood	OLS or Maximum Likelihood
Properties of estimates	consistent, normal, asymptotically efficient	consistent, normal	consistent, normal, asymptotically efficient	consistent, normal, efficient
Pro's	Can be used with crude measures	Can be used when end point values are common	Good match for binomial models	Familiar, easy to implement
Con's	May introduce measurement	May require computationally demanding	Model of mean can be biased if model of dispersion is	OLS assumptions are violated, can generate out-of-

error in Y, difficult to interpret	methods for statistical tests	incorrect; cannot handle end points	bounds predictions

Fractional regression is estimated using the Stata fracreg command. It also can be estimated using the GLM regression framework (using the Stata glm command). In fact, the Stata fracreg procedure can be characterized as being a convenient way to fit a particular quasi-maximum likelihood estimator from the family of generalized linear models (GLM's).

Fractional regression can be applied to continuous variables that fall in a bounded range (including the end points of the range).

Fractional response data may occur when the outcome of interest is measured as a fraction, for example, Gini coefficient values, segregation index scores, racial composition scores for neighborhoods, sex composition of occupations and industries, values on 0-100 "feeling thermometers", ratings on Likert scales, and many others.

For modeling purposes, the variable is rescaled to the range 0-1 by subtracting the lowest logically possible score from all values and dividing by the difference between the lowest and highest possible scores. (For example, Likert scores on a continuum from 1 to 9 would be rescaled by subtracting 1 and dividing by 8.)

Fractional regression predicts the mean of the dependent variable y conditional on covariates x (i.e., μx). Because y is in the range 0-1, it is necessary to ensure that predictions of the mean (μx) fall in this range. This is accomplished by using a logit (or probit) model for μx. Specifying the logit option imposes the requirement that predictions of the mean must fall on a logistic curve. Stated another way, the model is linear and additive for the logit of the mean.[1] This keeps the predictions of the mean "in bounds".

The fractional regression model is fit by the method of quasi-maximum likelihood. One advantage of this method is that it does not require the researcher to make strong assumptions about all aspects of the data generating model to obtain consistent parameter estimates. The only crucial assumption is that the model of the conditional mean is correct (i.e., include relevant variables, fit nonlinearities, etc.). This, of course, is crucial to all estimation methods.

The quasi-maximum likelihood estimation (QMLE) method is a more flexible (less restrictive) version of maximum likelihood estimation (MLE) methods. QMLE maximizes a function that is similar to the log likelihood function of MLE but it requires fewer assumptions about the specification of the model (e.g., the form of the distribution of errors around the estimate of the predictions). QMLE estimates have the desirable properties of being consistent and asymptotically normal, but they are less efficient than comparable MLE estimates.

In some cases, the disadvantage in efficiency of QMLE estimates of model parameters (e.g., b's) may be modest and standard approaches to statistical inference for maximum likelihood estimates can be used. More generally one should take a conservative approach and assume standard tests are too optimistic (standard errors are too small and generate too many false positives for statistical significance). Stata's fracreg procedure adopts this approach and estimates "robust standard errors" by default. The fracreg procedure also provides the option to use bootstrap

---

[1] Note this is ***NOT*** equivalent to modeling logit scores for values of Y. That approach predicts the mean of the logits of Y, which is not the same as predicting the logit of the mean of Y measured in its original scores.

methods to estimate standard errors. These methods can be computationally demanding, but provide estimates of standard errors that require fewer assumptions about the underlying model.

Stata's glm procedure also can be used to estimate fractional regression. The following two Stata commands implement identical estimations and report identical results. The first command uses the "fracreg" procedure and is easier to use. The second command uses the "glm" procedure and is harder to use, but allows for using more complex model specifications.

```
fracreg logit y x1 x2 , vce(robust)

glm y x1 x2 , family(binomial) link(logit) vce(robust)
```

The "vce(robust)" option specifies a method of estimating standard errors that does not make strong assumptions about the distribution of the error term and typically yields larger (i.e., more conservative) standard errors. The option does not need to be specified with fracreg because it is invoked by default. Alternatively, one can invoke the computationally more intensive bootstrap option by replacing "vce(robust)" with "vce(bootstrap,reps(500))"

Papke and Wooldridge (1996) introduced fractional regression and quasilikelihood estimation for applications in economics. Wooldridge (2010) provides a more recent technical discussion. Crowell and Fossett (2018) provides an example of an application in sociological research.

Example Analyses

This section presents example analyses using the same data that were discussed in the "notes" document for ordered logit regression. The examples here compare OLS regression and fractional regression for analyzing opposition to residential integration as measured by the bounded, interval-level variable OPPINT. The OPPINT variable is created from the four-category ordinal variable OPPINT4. Ordinal measurement is converted into interval measurement by rescaling the scores of 1-4 for OPPINT4 to fall continuously over the range 0-1.

The rescaling procedure assumes the interval-level values of the latent variable giving rise to the "mild" no and yes responses cover a larger share of the bounded range than is the case for the "firm" no and yes responses. Thus "firm no's" are distributed over 0.00-0.15, "mild no's are distributed over 0.15-0.50, "mild yes's" are distributed over 0.50-0.85, and "firm yes's are distributed over 0.85-1.00. (The rationale for applying this rescaling procedure could be debated, but it is not relevant for the illustration of the methods.)

Previously we analyzed OPPINT4 using mlogit and ologit. Now we use OLS regression and fractional regression. First two very simple models that are the most favorable for OLS regression. One is the "No-X" model and the other is a "simple difference of means" model. The No-X model has no predictors. The "difference of means model has only one X which is a dichotomy.

In these two special-case models OLS and fractional regression will give the same predictions. The main difference between them is the validity of the statistical test in the difference of means model. The OLS assumptions are less valid. So even in the situation that is most favorable to OLS regression, fractional regression may be preferred.

Problems with OLS become more severe under the following conditions:

(1) there are many X variables,

(2) the X variables are continuous, and

(3) the X variables have strong effects and generate predictions near boundaries.

The problems that arise under these conditions include the following:

(1) the OLS assumption that effects are additive is inappropriate,

(2) the OLS assumption that effects are linear is inappropriate,

(3) due to (1) and (2) OLS can make invalid predictions,

(4) the OLS assumptions that errors of prediction are normally distributed with equal variance over all levels and combinations of predictors (X's) is not met,

(5) due to (4) OLS significance tests cannot be trusted.

Now lets consider the simplest model – no predictors (X's).

The No-X model has only a constant. For OLS, the constant represents the mean of Y (OPPINT measured 0-1). For fractional regression, the constant is the logit for the mean of Y (remember, this is not the mean of the logits for Y calculated for individual cases). The mean of Y is obtained from the inverse logit transformation (as described above).

The tabulation of the predictions from the models given below shows they are identical.

```
. reg oppint

      Source |       SS           df       MS      Number of obs   =     8,235
-------------+----------------------------------   F(0, 8234)      =      0.00
       Model |           0         0           .   Prob > F        =         .
    Residual |  829.075143     8,234  .100689233   R-squared       =    0.0000
-------------+----------------------------------   Adj R-squared   =    0.0000
       Total |  829.075143     8,234  .100689233   Root MSE        =    .31732

------------------------------------------------------------------------------
      oppint |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |    .342631   .0034967    97.99   0.000     .3357766    .3494855
------------------------------------------------------------------------------

. predict ols0_p
```

```
. fracreg logit oppint

Iteration 0:   log pseudolikelihood =  -6272.352
   ... (iterations omitted to save space)
Iteration 3:   log pseudolikelihood = -5293.1717

Fractional logistic regression                  Number of obs     =      8,235
Log pseudolikelihood = -5293.1717               Pseudo R2         =     0.0000

------------------------------------------------------------------------------
             |               Robust
      oppint |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -.6515912   .0155247   -41.97   0.000    -.6820191   -.6211633
------------------------------------------------------------------------------

. predict fr0_p
(option cm assumed)

.
. table south , c(mean ols0_p mean fr0_p) format(%8.3f)

--------------------------------------
Resides   |
in South  | mean(ols0_p)   mean(fr0_p)
----------+---------------------------
     0=No |        0.343         0.343
    1=Yes |        0.343         0.343
--------------------------------------
```

The Simple Difference of Means model has a constant and one coefficient for the dummy variable (south). For OLS, the constant represents the mean of Y (OPPINT measured 0-1) for the Non-South (when south=0) and the coefficient for south captures the South vs Non-South difference of means on Y.

```
. reg oppint south

      Source |       SS           df       MS      Number of obs   =     8,235
-------------+----------------------------------   F(1, 8233)      =    172.81
       Model |  17.0447313         1  17.0447313   Prob > F        =    0.0000
    Residual |  812.030412     8,233  .098631169   R-squared       =    0.0206
-------------+----------------------------------   Adj R-squared   =    0.0204
       Total |  829.075143     8,234  .100689233   Root MSE        =    .31406

------------------------------------------------------------------------------
      oppint |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       south |   .0984312   .0074876    13.15   0.000     .0837535    .1131088
       _cons |   .3121873   .0041642    74.97   0.000     .3040245    .3203501
------------------------------------------------------------------------------

. predict ols1_p
(option xb assumed; fitted values)
```

For fractional regression, the constant is the logit for the mean of Y for the Non-South (when south==0). The coefficient for South is the difference between the logit of the mean of Y for the Non-South compared to the logit of the mean of Y for the South.

```
. fracreg logit oppint south

Iteration 0:   log pseudolikelihood = -6257.3719
   ... (iterations omitted to save space)
Iteration 3:   log pseudolikelihood = -5255.8657

Fractional logistic regression                  Number of obs     =      8,235
                                                Wald chi2(1)      =     166.80
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -5255.8657               Pseudo R2         =     0.0070

------------------------------------------------------------------------------
             |               Robust
      oppint |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       south |    .428504   .0331781    12.92   0.000     .3634762    .4935317
       _cons |  -.7899132   .0188224   -41.97   0.000    -.8268045   -.7530219
------------------------------------------------------------------------------

. predict fr1_p
```

The predictions for the mean of Y by region are obtained by applying the inverse logit transformation (as described above) to the predictions for the logit of the mean of Y for each region.

As before, the tabulation of the predictions from the models shows they are identical.

This is the most complex case where this result can be expected. This result is not expected when X is continuous or when there are two X's (it can be approximated under certain conditions; for example, if the two X's have small effects and/or the continuous X varies in a narrow range).

```
. table south , c(mean ols1_p mean fr1_p) format(%8.3f)

--------------------------------------
Resides   |
in South  | mean(ols1_p)   mean(fr1_p)
----------+---------------------------
     0=No |        0.312         0.312
    1=Yes |        0.411         0.411
--------------------------------------
```

The next results are for a "full model" that includes several predictors (X's) including some interaction variables that allow effect of education, year, and male to vary across South and Non-South.

```
. gen double xeduc6 = south * educ6   // creating interactions with south
. gen double xyear7 = south * year7
. gen double xmale  = south * male
```

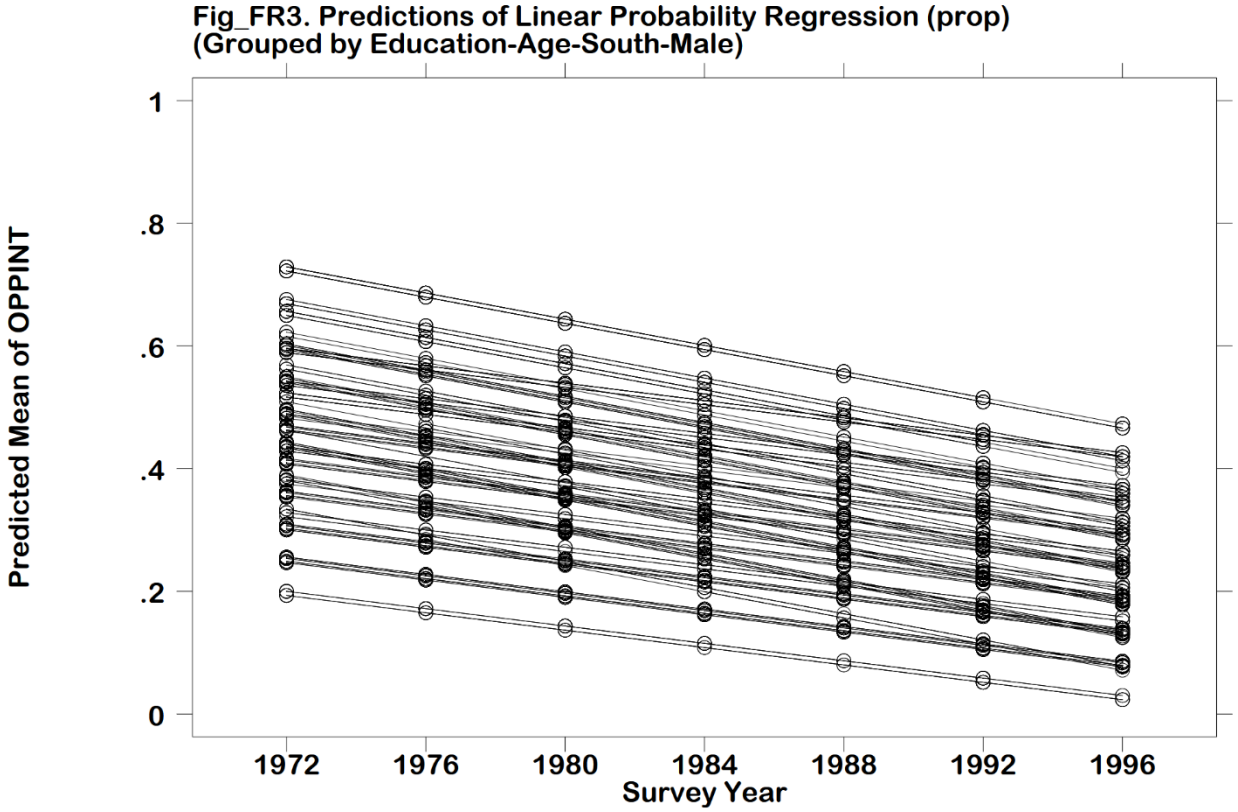
```
. reg oppint year7 xyear7 south male educ6 xeduc6 i.age3

      Source |       SS           df       MS      Number of obs   =     8,235
-------------+----------------------------------   F(8, 8226)      =    214.50
       Model |  143.048245         8  17.8810306   Prob > F        =    0.0000
    Residual |  685.738503     8,226  .083362327   R-squared       =    0.1726
-------------+----------------------------------   Adj R-squared   =    0.1718
       Total |  828.786748     8,234  .100654208   Root MSE        =    .28873

------------------------------------------------------------------------------
      oppint |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       year7 |  -.0282988   .0019899   -14.22   0.000    -.0321995   -.0243981
      xyear7 |  -.0143587    .003611    -3.98   0.000    -.0214372   -.0072803
       south |   .1316764   .0155079     8.49   0.000     .1012769    .1620758
        male |   .0064825   .0064131     1.01   0.312    -.0060887    .0190538
       educ6 |  -.0548333   .0030728   -17.84   0.000    -.0608567   -.0488099
      xeduc6 |   -.000167    .005127    -0.03   0.974    -.0102173    .0098833
             |
       age3 |
      30-59  |   .0568588   .0073587     7.73   0.000      .042434    .0712837
      60-99  |   .1286406   .0087472    14.71   0.000      .111494    .1457872
             |
       _cons |   .4635363   .0105831    43.80   0.000     .4427908    .4842819
------------------------------------------------------------------------------

. predict ols2_p


. fracreg logit oppint year7 xyear7 south male educ6 xeduc6 i.age3

Iteration 0:   log pseudolikelihood = -6146.3635
   ... (iterations omitted to save space)
Iteration 4:   log pseudolikelihood = -4970.9657

Fractional logistic regression                  Number of obs     =      8,235
                                                Wald chi2(8)      =    1579.91
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -4970.9657               Pseudo R2         =     0.0611

------------------------------------------------------------------------------
             |               Robust
      oppint |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       year7 |  -.1397249   .0096108   -14.54   0.000    -.1585617   -.1208882
      xyear7 |  -.0483953   .0174826    -2.77   0.006    -.0826605   -.0141301
       south |   .4937653   .0752815     6.56   0.000     .3462162    .6413143
        male |   .0271416    .030895     0.88   0.380    -.0334115    .0876946
       educ6 |    -.26878   .0148411   -18.11   0.000    -.2978679   -.2396921
      xeduc6 |   .0270013   .0242389     1.11   0.265     -.020506    .0745087
             |
       age3 |
      30-59  |   .2689436   .0357766     7.52   0.000     .1988226    .3390645
      60-99  |   .5824124   .0426284    13.66   0.000     .4988623    .6659625
             |
       _cons |  -.0831893   .0513144    -1.62   0.105    -.1837638    .0173852
------------------------------------------------------------------------------

. predict fr2_p
```

In this case, the OLS and FRACREG results appear to be fairly close. That is, they tend to agree on the sign (direction) of effects and pm reported statistical significance. The OLS t-tests are less valid than the fracreg Z-tests. But the OLS t tests and the fracreg Z-tests do not disagree in any major ways.

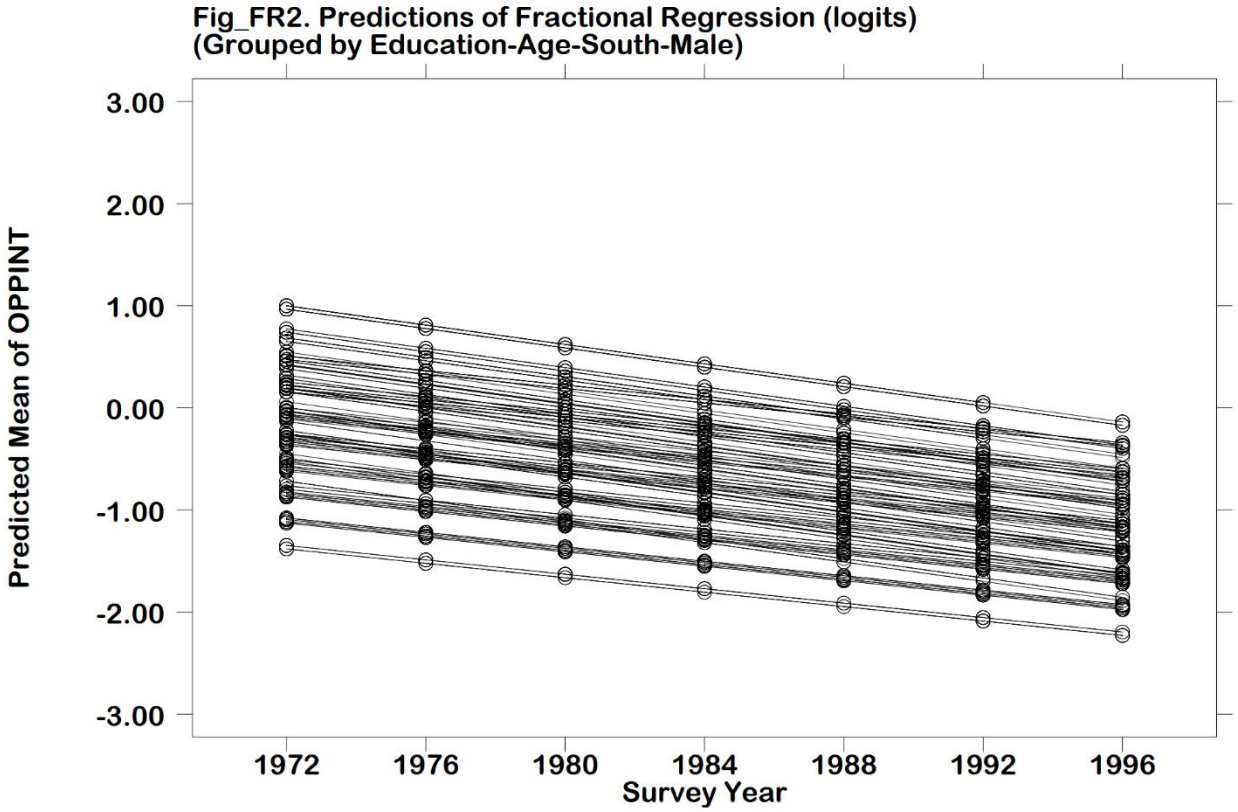
One advantage of OLS is that effects are easier to interpret. But the OLS effects are potentially misleading because a more accurate description of effects requires adopting a frame of reference (based on the settings of other X's).

Another disadvantage of OLS is that its errors of prediction are not random with respect to the values of X and especially combinations of values on X. Its errors will be largest when X's are at extreme values and occur in combinations that put predictions near Y's upper or lower boundaries.

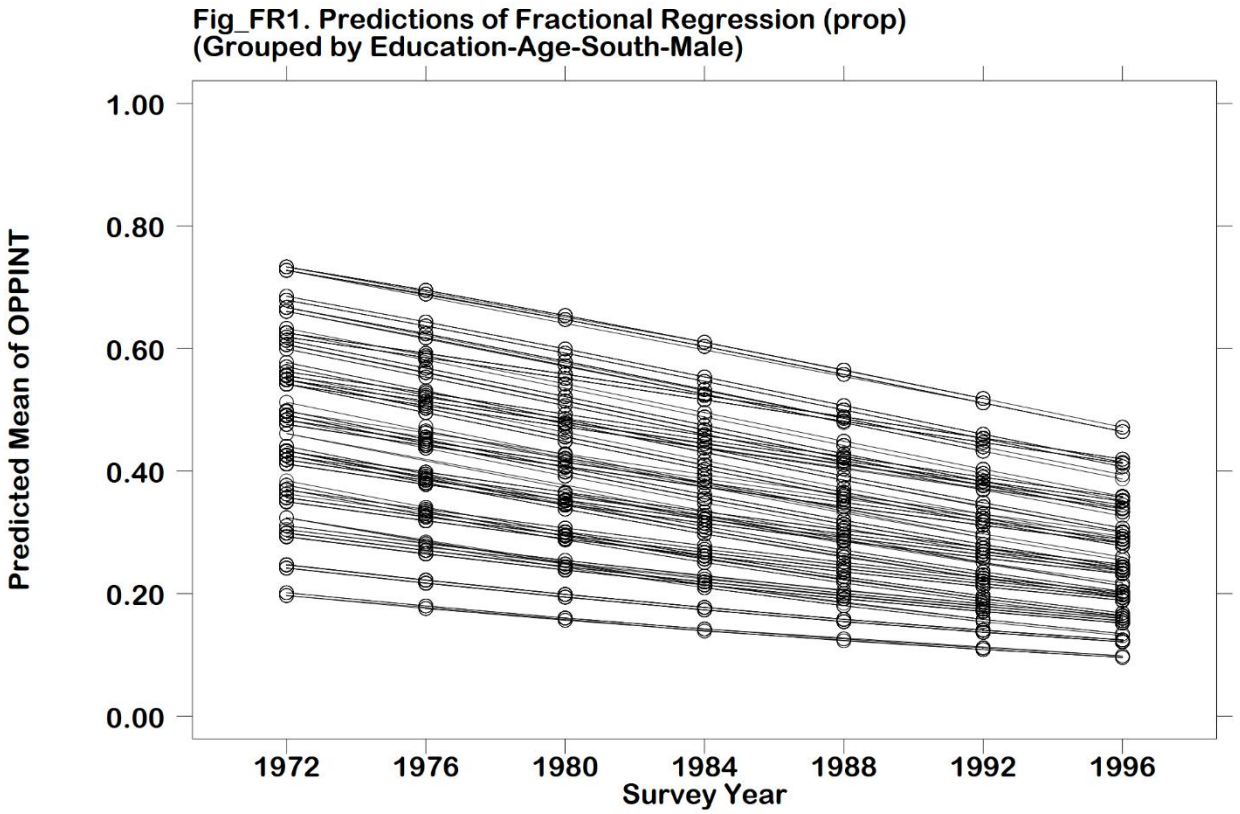
This can be seen in the following graph of OLS predictions which take some predictions close to the boundary. This may occur because the OLS model is linear and additive in the proportions.



The fractional logit regression is linear and additive in predicting logits. Thus, the graph below looks similar to the previous graph.



The important difference is that linear, additive predictions for logits translate into non-linear, non-additive predictions for proportions. That is evident in the following graph showing predictions of proportions from the fractional logit regression. Notice the top line and bottom lines have different slopes for the effect of year. In the earlier graphs those lines were parallel. As a result, fractional regression predictions do not get as close to the boundary as OLS predictions.

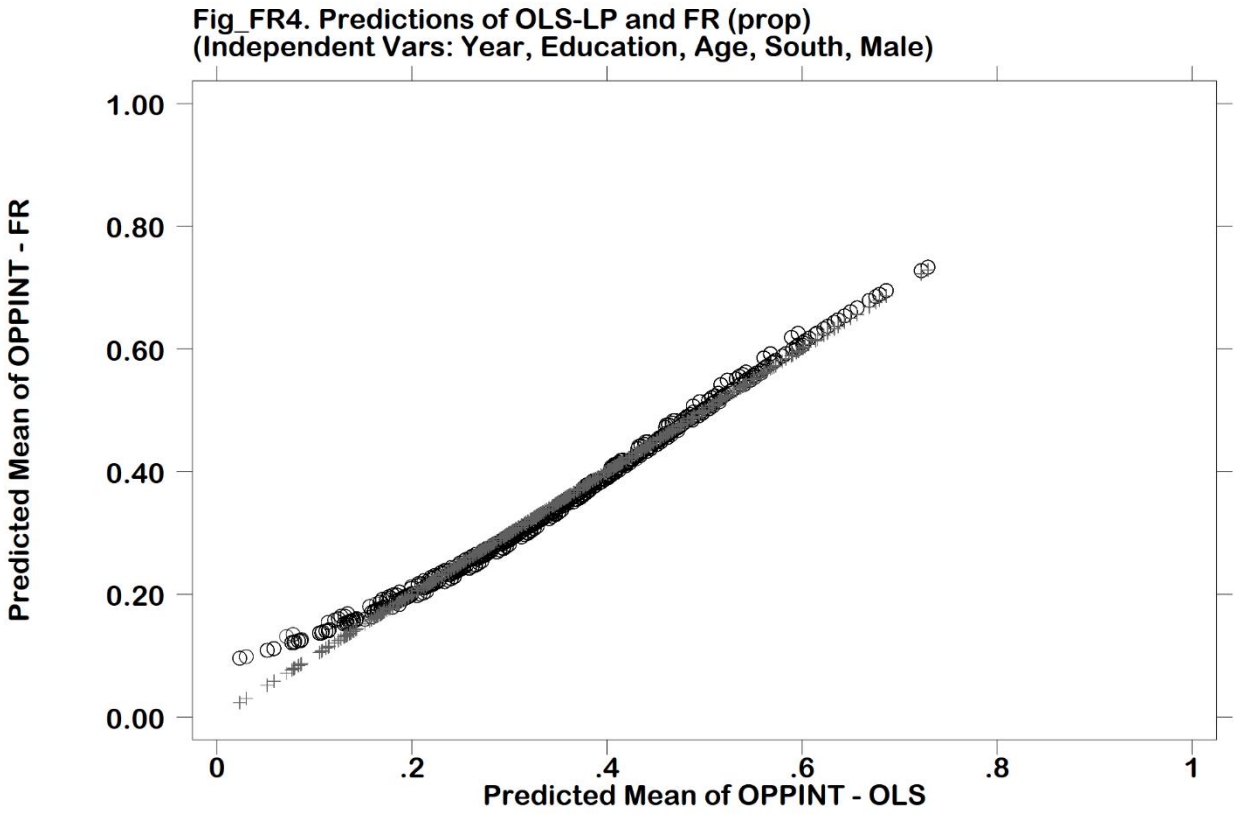


The following results from the table command lists differences in mean predictions that reach 0.02 in absolute value. (Note: The variable of male is not represented in the tabulation. More discrepancies in predictions might be observed if the predictions were grouped by male.)

```
. table educ6 age3 south if (abs(diff2_p) > 0.02) , by(year) c(mean diff2_p) format(%8.3f)

            | Resides in South and Age 0=18-29, 1=30-59, 2=60+
Year and    | -------- 0=No --------    -------- 1=Yes -------
Education   |  18-29   30-59   60-99     18-29   30-59   60-99
------------+-----------------------------------------------
1972        |
  0-8 Years |         -0.026  -0.030
 9-11 Years |                 -0.020
1976        |
  0-8 Years |                 -0.025
1980        |
  18+ Years | -0.021
1984        |
  18+ Years | -0.029
1988        |
16-17 Years | -0.021
  18+ Years | -0.042                   -0.025
1992        |
16-17 Years | -0.031                   -0.021
  18+ Years | -0.057  -0.028           -0.041  -0.022
1996        |
13-15 Years | -0.022
16-17 Years | -0.044                   -0.036
  18+ Years | -0.073  -0.042           -0.061  -0.037
```

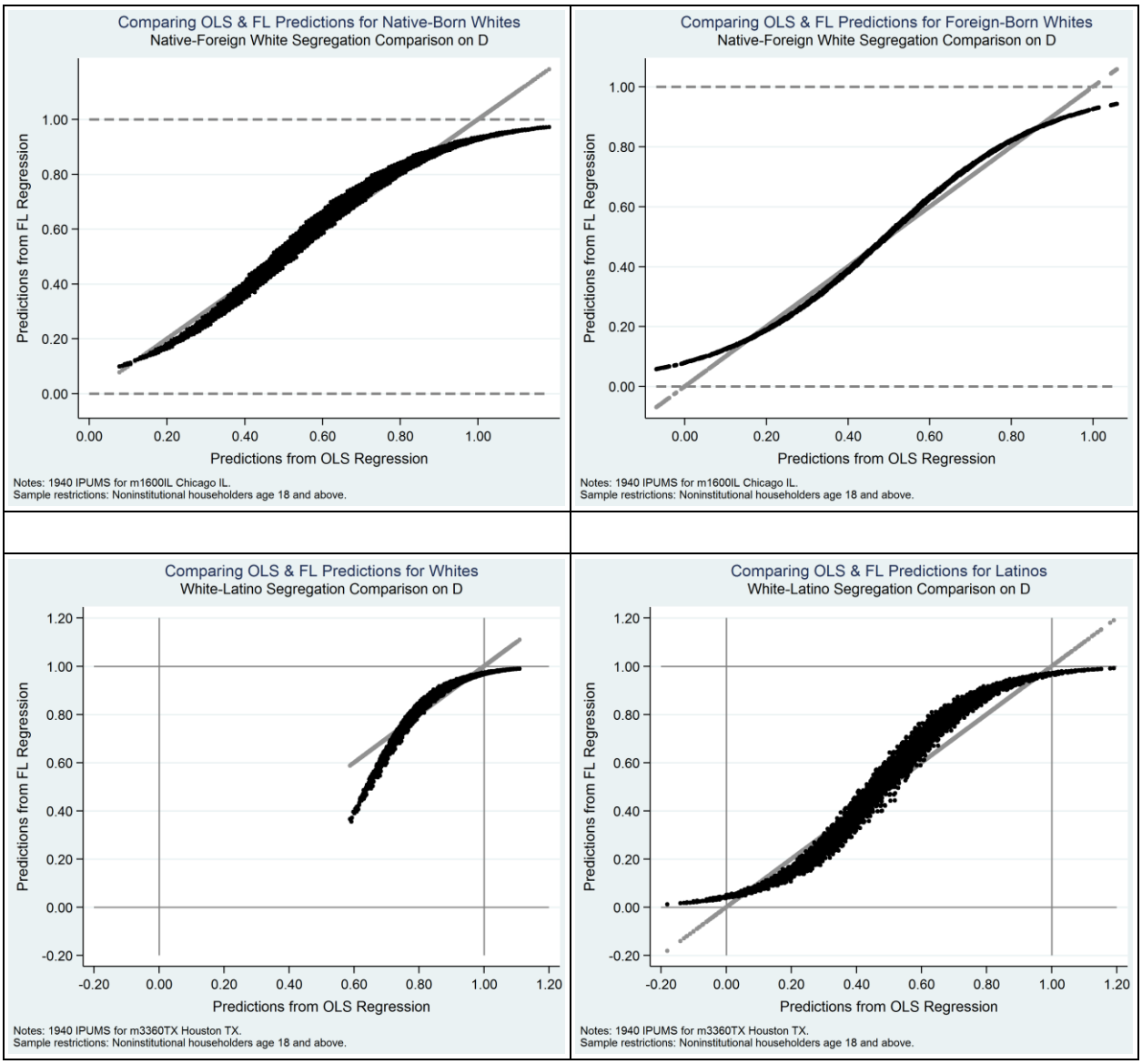
Simple summary: The largest misses are seen for cases where the X variables combine to put OLS predictions near the boundary. Fractional regression handles such cases more appropriately and accurately. In general, OLS is too quick to put cases near the boundary and, in the extreme, OLS will put cases beyond the boundary.

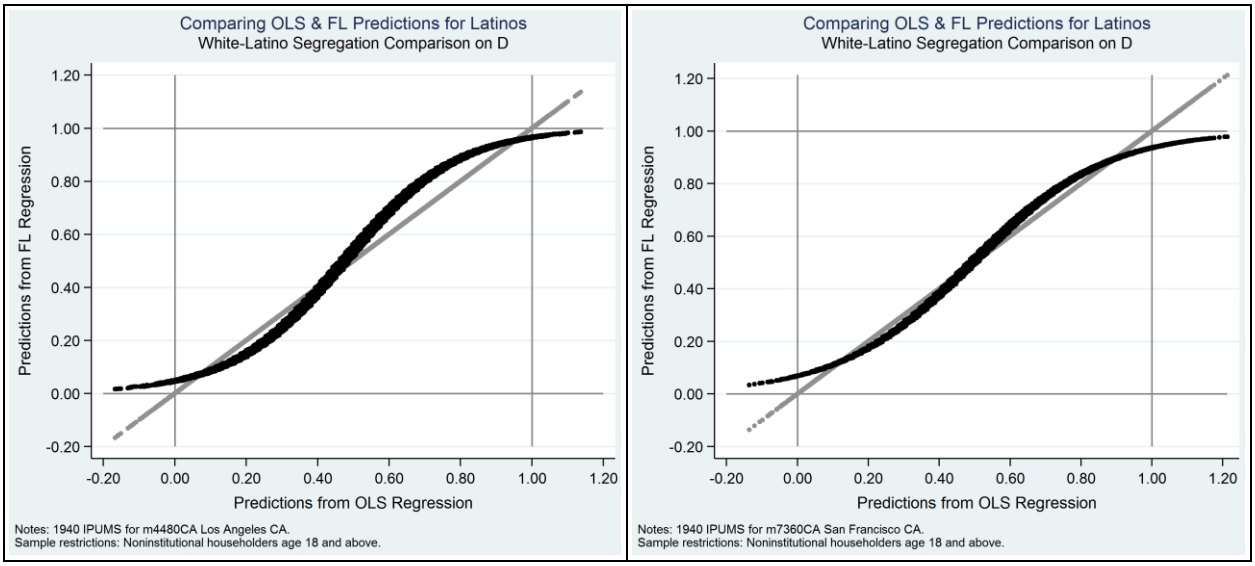
The following graph highlights a general pattern one will usually see in analyses of this type. The pattern is that predictions of OLS regression and fractional regression are usually very similar when the predictions fall in the range 0.20 to 0.80. But the predictions diverge when OLS predictions get outside of the 0.20 to 0.80 range as seen here for predictions in the range near the lower boundary (i.e., 0.00 to 0.20) and as might occur for predictions in the range near the upper boundary (i.e., 0.80 to 1.00).



The following are graphs from research in progress by Fossett and Crowell. The research is investigating how residential attainment processes shape residential segregation in metropolitan areas in 1940. The models involved are micro-level regressions predicting contact with native-born whites at the level of neighborhoods. Following conventions in segregation research, "contact" is conceived as "co-residence" as measured by proportion native-born white in the neighborhood. The models are estimated separately by racial-ethnic groups. The predictors include: nativity (US-born), age, education, income, gender, presence of one or more foreign-born household members.

The graphs plot fractional logit predictions of the mean contact against the predictions from OLS regression. The takeaway point from these figures is that the OLS regressions routinely make poor predictions when contact approaches boundaries and in many cases the predictions are out of bounds by large amounts.





References

Crowell, Amber R. and Mark Fossett. 2018. "White and Latino Locational Attainments: Assessing the Role of Race and Resources in U.S. Metropolitan Residential Segregation." *Sociology of Race and Ethnicity* (4):491-507.

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: 619–632.

Papke, L. E., and J. M. Wooldridge. 2008. "Panel data methods for fractional response variables with an application to test pass rates." *Journal of Econometrics* 145: 121-133.

Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. (2nd ed.) Cambridge, MA: MIT Press.

END OF NOTES