Ordinal Logit and Fractional Regression: Analysis Options for Ordered and Continuous Bounded Variables

Mark Fossett
Sociology, Texas A&M University
m-fossett@tamu.edu

Sociology Quantitative Methods Series, Summer 2021
Overview of the Presentation

Review options for analyzing dependent variables that are continuous over a bounded range or involve three or more ranked categories.

- Consider pros and cons of OLS regression
- Consider pros and cons of Binary Logit regression
- Consider pros and cons of Multinomial Logit Regression
- Consider pros and cons of Ordered Logit Regression
- Consider pros and cons of Fractional Logit Regression
- Note related procedures such as Beta Regression

Review selected results from analyses performed using demonstration programs.
Assumptions

This presentation presumes familiarity with:
- Multiple regression analysis
- Logit regression analysis
- Multinomial Logit regression analysis

Also
- Odds ratio transformations of proportions \((OR = P/(1-P))\)
- Logit transformations of odds ratios \((L = \ln(OR))\)
- Inverse logit to odds ratio transforms \((OR = e^L)\)
- Inverse odds ratio to prop transform \((P = OR/(1+OR))\)
Dependent Variables – Level of Measurement

Nominal Dependent Variables
- Distinctive outcomes can be identified
- Theory and research question make distinctions meaningful
- Notions of distance or even rank order are not valid

Ordinal Dependent Variables
- Adds the ability to make valid rank order distinctions
- Weak notions of “distance” thresholds justify rank distinctions
- Strong notions of a “distance” scale are not valid

Interval Dependent Variables (Integer & Continuous)
- Adds the ability to make valid distinctions on a “distance” scale
- True zero and relative distance distinctions are not valid

Ratio Dependent Variables (Continuous)
- Adds true zero
- Adds the ability to make relative scale “distance” distinctions
Dependent Variables – Bounded vs. Unbounded

Unbounded - In principle, lower and upper values are unbounded
Income, log-odds ratios, zero-centered scaled variables

Single Bounded – Lower or upper values are bounded (capped)
Odds ratios, open-ended count variables such as years of education, number of children, soccer goals, years since immigration, etc.

Double Bounded – Lower and upper values are bounded (capped)
All ordinal variables (e.g., income quintile, low-medium-high, etc.), Likert style scales, SES indices, inequality measures (e.g., Gini), segregation indices (e.g., Dissimilarity), proportions and rates (e.g., poverty rate)

Bounds can be intrinsic (lower boundary for number of children) or bounds can be imposed by measurement procedures (e.g., placing top and bottom codes on income) for practical reasons, to protect confidentiality, etc.
# Selected Analysis Methods by Type of DV

<table>
<thead>
<tr>
<th></th>
<th>Nominal</th>
<th>Ordinal</th>
<th>Interval&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bounded Dependent Variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crosstab and X²</td>
<td>√√√</td>
<td>√</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Binary Logit Reg.</td>
<td>√√√</td>
<td>√</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>OLS Lin Prob Reg.</td>
<td>√</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Multinomial Logit</td>
<td>√√√</td>
<td>√</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Ordinal Logit Reg.</td>
<td>---</td>
<td>√√√</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Count Reg.</td>
<td>---</td>
<td>---</td>
<td>√√√&lt;sup&gt;b&lt;/sup&gt;</td>
<td>---</td>
</tr>
<tr>
<td>OLS Regression</td>
<td>---</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Fractional Reg.</td>
<td>√√√</td>
<td>√</td>
<td>√√√</td>
<td>√√√</td>
</tr>
<tr>
<td><strong>Unbounded Dependent Variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS Regression</td>
<td>---</td>
<td>---</td>
<td>√√√</td>
<td>√√√</td>
</tr>
</tbody>
</table>

“a” includes both integer (count) and continuous variables; “b” integer (count); “---” does not apply; “√” feasible, but suboptimal; “√√√” feasible and optimal.
Summary Remarks on Selected Analysis Methods

This section provides brief remarks on pros and cons of analysis methods for different types of dependent variables.
Analysis Methods – Dichotomous DV’s

Cross-tabulation and Chi Square
   Pros: Easy to implement
   Con1: No estimates of effects
   Con2: Multivariate analysis is difficult

Linear probability regression (OLS using 0,1 DV)
   Pros: Easy to interpret effects; easy to implement
   Con1: OLS assumptions for error term are not met
   Con2: Incorrectly assumes effects are linear & additive

Logit regression
   Pro1: Assumptions for error term are met
   Pro2: Correctly assumes effects are nonlinear & nonadditive
   Con1: Effects are harder to interpret
   Con2: Harder to implement and requires larger samples
Analysis Methods – Polytomous DV’s

Cross-tabulation and Chi Square
Pros: Easy to implement
Cons: Multivariate analysis is difficult

OLS regression – No options

Multinomial Logistic Regression
Pro1: Assumptions for error term are met
Pro2: Correctly assumes effects are nonlinear & nonadditive
Con1: Effects are very hard to interpret
Con2: Harder to implement and requires larger samples
Analysis Methods – Ordinal DV’s

OLS Regression Using “Rank” (1, 2, 3, etc.) as DV
Pros: Easy to interpret and easy to implement
Con1: OLS assumptions for error term are not met
Con2: Incorrectly assumes effects are linear & additive
Con3: Incorrectly assumes equal “distance” between ranks

Multinomial Logit regression
Pro1: Assumptions for error term are met
Pro2: Correctly assumes effects are nonlinear & nonadditive
Con1: Effects are unnecessarily complicated and hard to interpret
Con2: Harder to implement and requires larger samples
Fractional Regression (scale ranks to fall in 0-1 range)
Pro1: Assumptions for error term are met
Pro2: Correctly assumes effects are nonlinear & nonadditive
Con1: Parsimonious and easier to interpret
Con2: Harder to implement and requires larger samples
Con3: Incorrectly assumes equal “distance” between ranks

Ordinal Logit Regression
Pro1: Assumptions for error term are met
Pro2: Correctly assumes effects are nonlinear & nonadditive
Pro3: Effects are parsimonious and easier to interpret (in comparison to multinomial logit regression)
Pro4: Ordinality assumption can be tested (ordered logit is “nested” under multinomial logit regression)
Cons: Harder to implement and requires larger samples
Analysis Methods – Unbounded Interval DV’s

OLS Regression Using “Raw Scores”
  Pro1: Easy to interpret and easy to implement
  Pro2: OLS assumptions for error term are appropriate
  Pro3: Assumptions of linear, additive effects are appropriate
  Cons: Limited to usual OLS concerns

Other Methods
  Not indicated unless OLS assumptions are violated
  Next logical alternatives are to consider extensions of OLS regression such as Weight Least Squares regression, Robust regression, and Bootstrapped OLS regression
Concerns – Double Bounded Continuous DV’s

OLS regression assumes the dependent variable is “unbounded” on both the lower and upper ends of the relevant scale.

This follows directly from the OLS assumptions that the errors of prediction ($e_i$) at any combination of values on the independent variables ($X$’s) are: (a) normally distributed and (b) have uniform (equal) variance.

OLS errors for bounded variables do not meet this assumption; errors are non-normal and have unequal variance.

Significance tests are compromised.

OLS regression assumes effects are linear and additive.

This is inappropriate.

It can lead to larger errors of prediction, especially near the lower and upper boundaries of the DV.

In the extreme, can lead to predictions that are impossible (out-of-bounds).
Analysis Methods – Double Bounded Continuous DV’s

OLS Regression Using “Raw Scores” or, alternatively, Using a 0-1 Transform (to highlight bounds)

Pro1: Easy to interpret and easy to implement

Pro2: In some cases, results can be fairly robust to mild violations of assumptions. For example, OLS analysis of SES scores coded 1-99 with predicted values in range 20-80.

Con1: OLS assumptions for error term are not met

Con2: Incorrectly assumes effects are linear & additive

OLS Regression Using Logit (or Similar) Transform for “Raw Scores” Converted to 0-1 Range

Pro1: Easy to interpret and easy to implement

Pro2: Appropriately assumes effects are nonlinear & nonadditive

Con1: The logit transform is undefined at the boundary values (0 & 1) and results can be sensitive to arbitrary choices for how boundary values are scored

Con2: Models the mean of logits, not the mean of “raw scores”
Analysis Methods – Double Bounded Cont. DV’s (II)

Fractional Regression (DV scaled to fall in 0-1 range)

Pro1: Assumptions for error term are met
Pro2: Appropriately assumes effects are nonlinear & nonadditive
Pro3: Especially flexible for modeling proportions
Pro4: Directly models the mean of the DV (not the mean of a logit transform of the DV)

Cons: Harder to implement, requires larger samples, uses quasi-likelihood theory (not maximum likelihood)

Beta Regression (DV scaled to fall in 0-1 range)

Pros: Same as fractional regression
Con1: Harder to implement, requires larger samples, uses quasi-likelihood theory (not maximum likelihood)
Con2: The model is complicated – it involves simultaneously estimating separate parameters for effects on both the mean and the dispersion in the distribution
Hands on with Ordinal Logit Regression – I

When is Ordinal Logit Regression relevant?
The DV has 3 or more categories
   If only 2 categories, use regular logit regression
The DV has rank order quality
   If rank order is questionable, use multinomial logit regression
The Distances between ranks are not equal
   If distances are approximately equal, consider fractional regression

What is the contrast with Multinomial Logit Regression?
   Ordinal regression is simpler and more parsimonious, only one effect parameter for each X
   Ordinal logit regression “nests” under multinomial logit, so the assumption of simpler effects can be tested
Hands on with Ordinal Logit Regression – II

What is the contrast with OLS Regression with Rank Scores for Y?
OLS assumptions for the error term are not met. So, significance tests are compromised.

OLS incorrectly assumes effects are linear and additive. So, predictions can be flawed and even fall “out-of-bounds”.

OLS imposes an arbitrary assumption of equal distance between ranks. This can lead to incorrect predictions of the distribution of cases over categories of Y (the DV).

What is the contrast with Binary Logit Regression?
One can “collapse” the ordinal variable into a simple dichotomy to make binary logit regression feasible.

The advantage is the model is easier to interpret.

The cost is potentially important loss of information regarding the dependent variable. Generally speaking, this is undesirable.
What is the contrast with Fractional Regression with Continuous Scores for Y over the 0-1 Range?

FR incorrectly assumes distances between categories are known.
FR assumes scores on Y can be continuous. This is inappropriate when the ordinal variable has fewer than 10 categories.
FR incorrectly assumes equal precision in measuring distinctions on Y across the full range of Y.
Hands on with Ordinal Logit Regression – IV

The Logic of Ordered Logit Regression

As with multinomial logit regression, the dependent variable is the relative distribution of cases across categories of Y.

In graphical terms, the DV is the shape of the histogram for Y across combinations of values on X’s.

A baseline histogram is captured with “cut-point” coefficients.

The cut-point coefficients predict the shape of the histogram for Y when X’s are at 0.

With no X’s in the model, the cut-point coefficients are the logits for the ratio of cases above and below the cut point.

Effects of X’s shift the shape of the predicted histogram.

Coefficients for X’s increment or decrement the cut points coefficients; this shifts all bars in the predicted histogram.

Negative effects shift the relative frequency distribution to the left (thus increasing left-most bars).

Positive effects shift the relative frequency distribution to the right (thus increasing right-most bars)
White Opposition to Residential Integration

No Effects Over Time (4-Year Intervals), Histogram Bars are Constant

Fig 2a. Opposed to Residential Integration - All Regions (OLOGIT Predictions for All Whites)

White Opposition to Residential Integration

Negative Effect Over Time (4-Year Intervals), Histogram Bars Shift Left

Fig2b1. Opposed to Residential Integration - Non-South
(OLOGIT Predictions for All Whites)

From Tabulation to MLOGIT Relative Risk Coefficients

. tab oppint4 // simple tabulation of dependent variable

Okay to Oppose Integration
<table>
<thead>
<tr>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Firm-No</td>
<td>3,801</td>
<td>46.16</td>
</tr>
<tr>
<td>Mild-No</td>
<td>2,099</td>
<td>25.49</td>
</tr>
<tr>
<td>Mild-Yes</td>
<td>1,212</td>
<td>14.72</td>
</tr>
<tr>
<td>Firm-Yes</td>
<td>1,123</td>
<td>13.64</td>
</tr>
<tr>
<td>Total</td>
<td>8,235</td>
<td>100.00</td>
</tr>
</tbody>
</table>

. mlogit oppint4, base(1) rr

Multinomial logistic regression
Number of obs = 8,235
LR chi2(0) = 0.00
Prob > chi2 = .
Log likelihood = -10367.648
Pseudo R2 = 0.00

| oppint4 | Rel. risk  | Std. err. | z    | P>|z| | [95% conf. interval] |
|---------|------------|-----------|------|-----|---------------------|
| Firm_No | (base outcome) |          |      |     |                     |
| Mild_No |            | .5522231  | .0150171 | -21.84 | 0.000      | .5235608 | .5824545 |
| Mild_Yes| _cons     | .3188635  | .0105185 | -34.65 | 0.000      | .2988999 | .3401604 |
| Firm_Yes| _cons     | .2954486  | .0100346 | -35.90 | 0.000      | .2764214 | .3157855 |

Note: Mild No _cons = 0.5522 = 2099/3801
From Tabulation to MLOGIT Logit Coefficients

```
. mlogit oppint4, base(1)

Multinomial logistic regression
Number of obs = 8,235
LR chi2(0) = 0.00
Prob > chi2 = .
Log likelihood = -10367.648
Pseudo R2 = 0.0000

------------------------------------------------------------------------------
| oppint4 | Coefficient  Std. err.  z     P>|z|  [95% conf. interval] |
|----------|---------------|--------|--------|---------------------------|
| Firm_No  | (base outcome)|        |        |                           |
| Mild_No  |               |        |        |                           |
| _cons    | -.5938031     | .0271939 | -21.84| 0.000         | -.6471021   | -.5405042 |
| Mild_Yes |               |        |        |                           |
| _cons    | -1.142992     | .0329874 | -34.65| 0.000         | -1.207646   | -1.078338 |
| Firm_Yes |               |        |        |                           |
| _cons    | -1.219261     | .0339641 | -35.90| 0.000         | -1.285829   | -1.152692 |
------------------------------------------------------------------------------
Note: Mild No _cons = -0.5938 = ln(0.5522) = ln(2099/3801)
```
From Tabulation to OLOGIT Logit Coefficients

. tab oppint4 // simple tabulation of dependent variable

Okay to Oppose Integration

<table>
<thead>
<tr>
<th></th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Firm-No Y1</td>
<td>3,801</td>
<td>46.16</td>
<td>46.16</td>
</tr>
<tr>
<td>Mild-No Y2</td>
<td>2,099</td>
<td>25.49</td>
<td>71.65</td>
</tr>
<tr>
<td>Mild-Yes Y3</td>
<td>1,212</td>
<td>14.72</td>
<td>86.36</td>
</tr>
<tr>
<td>Firm-Yes Y4</td>
<td>1,123</td>
<td>13.64</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>8,235</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

Baseline cut point odds ratios calculated from frequency distribution

- **cut1** odds ratio = \((Y1N) / (Y2N+Y3N+Y4N)\)  
  \[ \text{logit cut1} = -0.15404 \]

- **cut2** odds ratio = \((Y1N+Y2N) / (Y3N+Y4N)\)  
  \[ \text{logit cut2} = 0.92694 \]

- **cut3** odds ratio = \((Y1N+Y2N+Y3N) / (Y4N)\)  
  \[ \text{logit cut3} = 1.84578 \]

. ologit oppint4

Ordered logistic regression  
Number of obs = 8,235  
Log likelihood = -10367.648  
Pseudo R2 = 0.0000

| oppint4 | Coefficient  | Std. err.  | z     | P>|z|   | [95% conf. interval] |
|---------|--------------|------------|-------|-------|----------------------|
| /cut1   | -.1540379    | .0221047   | -.697 | .488  | -.3170595 to .0096911|
| /cut2   | .9269405     | .0244491   | 3.796 | .000  | .8790212 to .9748598 |
| /cut3   | 1.84578      | .0321104   | 5.829 | .000  | 1.782845 to 1.908715 |
MLOGIT and OLOGIT predictions are the same for baseline models that have no predictors (no X variables).

Predictions from mlogit (model with no X’s)

. table, stat(mean pv1z_y1 pv1z_y2 pv1z_y3 pv1z_y4) nformat(%9.4f)

<table>
<thead>
<tr>
<th>Pr(oppint4==Firm_No)</th>
<th>0.4616</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr(oppint4==Mild_No)</td>
<td>0.2549</td>
</tr>
<tr>
<td>Pr(oppint4==Mild_Yes)</td>
<td>0.1472</td>
</tr>
<tr>
<td>Pr(oppint4==Firm_Yes)</td>
<td>0.1364</td>
</tr>
</tbody>
</table>

Predictions from ologit (model with no X’s)

. table, stat(mean pv2a_y1 pv2a_y2 pv2a_y3 pv2a_y4) nformat(%8.4f)

<table>
<thead>
<tr>
<th>Pr(oppint4==1)</th>
<th>0.4616</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr(oppint4==2)</td>
<td>0.2549</td>
</tr>
<tr>
<td>Pr(oppint4==3)</td>
<td>0.1472</td>
</tr>
<tr>
<td>Pr(oppint4==4)</td>
<td>0.1364</td>
</tr>
</tbody>
</table>
## Parsimony of OLOGIT vs MLOGIT – I

```
.mlogit oppint4 south , base(1)
```

### Multinomial logistic regression

- Number of obs = 8,235
- LR chi2(3) = 193.68
- Prob > chi2 = 0.0000
- Log likelihood = -10270.81
- Pseudo R2 = 0.0093

The table below presents the coefficients, standard errors, z-values, p-values, and confidence intervals for the different outcomes:

| oppint4 | Coefficient | Std. err. | z     | P>|z|   | [95% conf. interval] |
|---------|-------------|-----------|-------|-------|----------------------|
| Firm_No | (base outcome) |           |       |       |                      |
| Mild_No |              |           |       |       |                      |
| south   | 0.4419306   | 0.0598196 | 7.39  | 0.000 | 0.3246863 - 0.5591748|
| _cons   | -0.7202567  | 0.0325748 | -22.11| 0.000 | -0.7841022 - 0.6564113|
| Mild_Yes|              |           |       |       |                      |
| south   | 0.5116254   | 0.0711859 | 7.19  | 0.000 | 0.3721036 - 0.6511472|
| _cons   | -1.293165   | 0.0401642 | -32.20| 0.000 | -1.371886 - -1.214445 |
| Firm_Yes|              |           |       |       |                      |
| south   | 0.9367234   | 0.0709243 | 13.21 | 0.000 | 0.7977142 - 1.075733  |
| _cons   | -1.53871    | 0.0443336 | -34.71| 0.000 | -1.625602 - -1.451818 |
### Parsimony of OLOGIT vs MLOGIT – II

```
. ologit oppint4 south ,

Ordered logistic regression

Number of obs =  8,235
LR chi2(1)    = 184.57
Prob > chi2   = 0.0000
Log likelihood = -10275.361
Pseudo R2     = 0.0089

----------------------------------------------------------------------------
        oppint4 | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+------------------------------------------------------------
          south |   .5963232   .0438793    13.59   0.000     .5103214     .682325
-------------+------------------------------------------------------------
/cut1        |   .0258228   .0259769   -.0250909     .0767365
/cut2        |   1.125163   .0289257    1.06847    1.181857
/cut3        |   2.056144   .036175    1.985242    2.127045
-------------+------------------------------------------------------------
```
MLOGIT & OLOGIT Predictions Now May Differ

Predictions from mlogit (model with one predictor - South)

. table, stat(mean pv1x_y1m pv1x_y2m pv1x_y3m pv1x_y4m) nformat(%9.4f)

<table>
<thead>
<tr>
<th>Pr(oppint4==Firm_No)</th>
<th>0.4616</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr(oppint4==Mild_No)</td>
<td>0.2549</td>
</tr>
<tr>
<td>Pr(oppint4==Mild_Yes)</td>
<td>0.1472</td>
</tr>
<tr>
<td>Pr(oppint4==Firm_Yes)</td>
<td>0.1364</td>
</tr>
</tbody>
</table>

Predictions from ologit (model with one predictor - South)

. table, stat(mean pv1x_y1o pv1x_y2o pv1x_y3o pv1x_y4o) nformat(%9.4f)

<table>
<thead>
<tr>
<th>Pr(oppint4==1)</th>
<th>0.4615</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr(oppint4==2)</td>
<td>0.2546</td>
</tr>
<tr>
<td>Pr(oppint4==3)</td>
<td>0.1473</td>
</tr>
<tr>
<td>Pr(oppint4==4)</td>
<td>0.1366</td>
</tr>
</tbody>
</table>
A Technical Aside RE: Stata’s Ordinal Logit Routine

The “Notes” document for ordered logit discusses a technical issue concerning how the Stata implementation of the ologit procedure “parameterizes” the coefficients reported by the model.

The issue is relevant when one uses ologit results to generate predicted values by manual calculation.

In brief, the calculations are slightly different from what one might expect. Nothing is incorrect. However, the model parameterization is different from what researcher might expect (but mathematically equivalent). As a result, the calculations need to be implemented in a particular way to get correct results.

Yes. This is a somewhat esoteric issue. But it helps avoid confusion regarding how to generate predictions using results from ologit regressions.
The “Notes” document discusses the following points.

The OLOGIT model “nests” under the MLOGIT model.

So, one can test the assumption of “fixed” ordinality by comparing OLOGIT and MLOGIT models with the same X’s.

In large samples, the MLOGIT results may test out as a statistically significant improvement over OLOGIT results.

Should MLOGIT results be automatically preferred in this situation? Not necessarily.

OLOGIT results are much more parsimonious and easier to interpret.

MLOGIT can capitalize on idiosyncratic patterns and “overfit” the effects. Ask three questions.

Are you confident the differences to hold up in new analyses?
Do the differences have important substantive implications?
Do you have a sound theory to make sense of the differences?
OLOGIT Prediction Success

How does one assess OLOGIT model fit?

Pseudo R2 statistics are well named. They really are “VERY pseudo”; they are not comparable to R2 in OLS regression.

Remember, the DV is a relative frequency distribution. The notion of explained “variance” does not really apply.

Dissimilarity between observed and predicted distributions provides some insight into whether predictions are “good” (see the “Notes” document for discussion).

BIC and AIC are also options to consider.

Think carefully before placing weight on model fit statistics.

Tests of theories do not hinge on model fit. They hinge on direction and magnitude of effects of particular X’s.

Model fit is mainly relevant for accuracy of prediction.

Ill-conceived models can produce better predictions than well-conceived models. But we would not accept them.
Fractional regression models the mean of Y as following a logistic “S” curve such that predicted means stay in the range 0-1.

Fractional regression can be a good option to consider when:

Scores for the dependent variable are continuous and are bounded in a limited range.
The classic case is proportions.
Many other bounded variables can be converted to the range of 0-1 (by applying a simple transformation formula).

OLS Regression (e.g., Linear Probability Models) are Inferior

OLS assumptions for errors of prediction are not met. So, significance tests are unsound.
OLS assumes effects are linear and additive. This is unsound and can lead to inaccurate and even impossible (e.g., out-of-bounds) predictions.
An Attractive Practical 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 easily lead to misleading findings.

Implication

Compare fractional regression and OLS regression to validate OLS inferences.

FR is always technical superior and, importantly, makes it clear that effects should be understood as nonlinear and non-additive.

But OLS results can be easier to explain.
Fractional Regression models are estimated using the GLM (General Linear Model) framework. The framework can fit a wide range of models. The model involves specifying two items.

The relevant distributional model.

In the case of FR, it is the binomial model (variation in Y values is between 0 and 1 inclusive).

The “link” function for the path of the mean.

In the case of FR, the link function is the “logit” (or probit) function.

Under this specification,

Predicted means for Y will take values between 0 and 1.
Predictions will follow a logistic “S” curve.
Errors of prediction at any point on the prediction curve will follow a binomial distribution (not a normal distribution).
Fractional regression models are fit by the method of quasi-maximum likelihood estimation (QMLE).

This 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. However, QMLE makes fewer strong assumptions about the specification of the model (e.g., the form of the distribution of errors around the estimate of the predictions).

Pros and Cons of QMLE Estimates

QMLE estimates have the desirable properties of being consistent and asymptotically normal (in large samples, as with maximum likelihood).

However, QMLE estimates are less efficient than comparable MLE estimates (i.e., they exhibit greater sample-to-sample variability).
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.

That is, it may be okay to use standard errors calculated using analytic formulas that rest on assumptions that may not be met.

More generally one should take a conservative approach and assume standard tests are too optimistic (standard errors based on analytic formulas are too small and generate too many false positives for statistical significance).

Stata’s fracreg procedure follows this recommended approach. Accordingly, it estimates “robust standard errors” by default. The fracreg procedure also provides the option to use bootstrap methods to obtain standard errors.
The following two Stata commands will implement fractional regression and generate identical results.

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

The fracreg command can be described as a convenient command for automating the glm estimation procedure.


Modeling a Continuous Version of OPPINT

The example analysis here models OPPINT. The values of OPPINT are continuous over the range 0-1. The value of 1 indicates opposition to integration. (See the “Notes” documents for details.)

We will compare results obtained when modeling this DV using OLS regression and fractional regression (FR).

First OLS with no X variables

```
. 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   0.00   .3357766    .3494855
------------------------------------------------------------------------------
```

The OLS regression constant of 0.3426 indicates the average score on OPPINT.
Modeling a Continuous Version of OPPINT

Next Fractional Regression (FR) with no X variables

```
.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

```
<table>
<thead>
<tr>
<th>Robust</th>
</tr>
</thead>
<tbody>
<tr>
<td>oppint</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>_cons</td>
</tr>
</tbody>
</table>
```

The FR regression constant of -0.6516 indicates the logit (log odds) for the predicted average score on OPPINT. The inverse logit transformation of this value yields the raw score prediction of 0.3426 (based on OR = $e^{-0.6516} = 0.5212$ and proportion (0-1) score = OR/(1+OR) = 0.5212/1.5212 = 0.3426.
Expanded OLS Regression Analysis

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

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 8,235</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>143.048245</td>
<td>8</td>
<td>17.8810306</td>
<td>F(8, 8226) = 214.50</td>
</tr>
<tr>
<td>Residual</td>
<td>685.738503</td>
<td>8,226</td>
<td>.083362327</td>
<td>Prob &gt; F = 0.000</td>
</tr>
<tr>
<td>Total</td>
<td>828.786748</td>
<td>8,234</td>
<td>.100654208</td>
<td>R-squared = 0.1726</td>
</tr>
</tbody>
</table>

| 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 | .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 |

See "Notes" document for details.
### Expanded Fractional Regression Analysis

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

Fractional logistic regression

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

See “Notes” document for details.
Comparing OLS Results with FR Results – I

In this case, the OLS and FR results appear to be fairly close. That is, they tend to agree on the sign (direction) of the effects of the independent variables and on reported statistical significance.

The OLS $t$-tests are less valid than the FR $Z$-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 nonlinear, nonadditive frame of reference (based on the settings of other $X$’s).

OLS errors will be largest when $X$’s occur in combinations that put predictions near the upper or lower boundaries for $Y$. This is evident in the following graph of OLS predictions.
Larger differences between OLS and FR predictions occur when predicted values of Y are near the boundaries of 0 and 1. This is because the OLS model incorrectly assumes linear, additive effects.
Comparing OLS Results with FR Results – III

Differences between OLS Regression and Fractional Regression will be more dramatic in other situations.

To illustrate, the graphs below compare OLS and FR predictions for 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 standard practice, “contact” is measured by proportion native-born white in the neighborhood. The predictors (X’s) include: nativity (US-born), age, education, income, gender, presence of one or more foreign-born household members.

The takeaway point is that the OLS predictions are clearly inferior when contact approaches boundaries. In many cases the predictions are “out of bounds” by large amounts.
Comparing OLS Results with FR Results – IV

White-Minority Segregation – Predicting Co-Residence with Whites

Native-Born Whites in Chicago IL 1940

Foreign-Born Whites in Chicago IL 1940

Notes: 1940 IPUMS for m1600L Chicago IL.
Sample restrictions: Noninstitutional households age 15 and above.
Comparing OLS Results with FR Results – V

White-Minority Segregation – Predicting Co-Residence with Whites

<table>
<thead>
<tr>
<th>Blacks in Houston TX 1940</th>
<th>Latinos in Houston TX 1940</th>
</tr>
</thead>
</table>

Comparing OLS & FL Predictions for Whites
White-Latino Segregation Comparison on D

Notes: 1940 PUMS for m13607X Houston TX.
Sample restrictions: Noninstitutional households age 18 and above.

Comparing OLS & FL Predictions for Latinos
White-Latino Segregation Comparison on D

Notes: 1940 PUMS for m13607X Houston TX.
Sample restrictions: Noninstitutional households age 18 and above.
Comparing OLS Results with FR Results – VI

White-Minority Segregation – Predicting Co-Residence with Whites

<table>
<thead>
<tr>
<th>Latinos in Los Angeles CA 1940</th>
<th>Latinos in San Francisco CA 1940</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparing OLS &amp; FL Predictions for Latinos White-Latino Segregation Comparison on D</td>
<td>Comparing OLS &amp; FL Predictions for Latinos White-Latino Segregation Comparison on D</td>
</tr>
</tbody>
</table>

Notes: 1940 PUMS for m4480CA Los Angeles CA. Sample restrictions: noninstitutional householders age 18 and above.

Notes: 1940 PUMS for m1760CA San Francisco CA. Sample restrictions: noninstitutional householders age 18 and above.
End of Slides

Thank you for your attention.