In many cases, the dependent variable of interest is a proportion, i.e. its values range between 0 and 1. Wooldridge (1996, 2011) gives the example of the proportion of employees that participate in a company’s pension plan. Baum (2008) gives as examples the share of consumers’ spending on food, the fraction of the vote for a candidate, or the fraction of days when air pollution is above acceptable levels in a city. This handout will discuss a few different ways for analyzing such dependent variables: fractional response models (both heteroskedastic and non-heteroskedastic) and zero one-inflated beta models.

**Fractional Response Models.** As Wooldridge notes, many Stata commands (logit, probit, hetprob) could analyze DVs that are proportions, but they impose the data constraint that the dependent variable must be coded as either 0 or 1, i.e. you can’t have a proportion as the dependent variable even though the same formulas and estimation techniques would be appropriate with a proportion. Wooldridge offers his own short programs that relax this limitation; but a more flexible solution is offered by Richard Williams’ user-written routine, fracglm, currently in (perpetual) beta testing. fracglm is adapted from oglm, and is probably easier to use than oglm when the dependent variable is a dichotomy (rather than an ordinal variable with 3 or more categories.)

To get fracglm, from within Stata type

```
net install fracglm, from(https://www3.nd.edu/~rwilliam stata)
```

This is usually the best way to install. Files are placed in the right locations, and `adoupdate` will capture any updates.
That doesn’t always work though. If it doesn’t work for you, try pointing your browser to
https://www3.nd.edu/~rwilliam/stata/fracglmbeta.zip

Download the file (it may download automatically), unzip it, and follow the directions for
installing that are in the Readme.txt file.

The following is adapted from the help for fracglm:

fracglm estimates Fractional Response Generalized Linear Models (e.g. Fractional Probit,
Fractional Logit) with or without heteroskedasticity. Fractional response variables range in value
between 0 and 1. An example of a fractional response variable would be the percentage of
employees covered by an employer’s pension plan.

fracglm also works with binary 0/1 dependent variables. fracglm supports multiple link functions,
including logit (the default), probit, complementary log-log, log-log, log and cauchit. When these
models include equations for heteroskedasticity they are also known as heterogeneous choice/
location-scale / heteroskedastic regression models.

fracglm fills gaps left by other Stata commands. Commands like logit, probit and hetprob do not
allow for fractional response variables. glm can estimate some fractional response models but does
not allow an equation for heteroskedasticity.

Several special cases of generalized linear models can also be estimated by fracglm, including the
binomial generalized linear models of logit, probit and cloglog (which also assume
homoskedasticity), hetprob, as well as similar models that are not otherwise estimated by Stata.
This makes fracglm particularly useful for testing whether constraints on a model (e.g.
homoskedastic errors) are justified, or for determining whether one link function is more
appropriate for the data than are others.

In addition, Stata 14 introduced the fracreg command. It isn’t quite as flexible as fracglm
(e.g. it doesn’t support as many link functions) but if you have Stata 14 it may be fine for your
needs.

Example. Papke and Wooldridge (1996) give an example of participation rates in employer
401(k) pension plans. “Pension plan administrators are required to file Form 5500 annually with
the Internal Revenue Service, describing participation and contribution behavior for each plan
offered. Papke (1995) uses the plan level data to study, among other things, the relationship
between the participation rate and various plan characteristics, including the rate at which a firm
matches employee contributions.”

In Wooldridge’s (2011) version of this example, data are from 4,075 companies in 1987. The key
variables used in this analysis are:
. use https://www3.nd.edu/~rwilliam/statafiles/401kpart, clear
. codebook prate mrate ltotemp age sole, compact

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Unique</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>prate</td>
<td>4075</td>
<td>2597</td>
<td>.840607</td>
<td>.0036364</td>
<td>1</td>
<td>partic/employ</td>
</tr>
<tr>
<td>mrate</td>
<td>4075</td>
<td>3521</td>
<td>.463519</td>
<td>0</td>
<td>2</td>
<td>plan match rate, per $</td>
</tr>
<tr>
<td>ltotemp</td>
<td>4075</td>
<td>2147</td>
<td>6.97439</td>
<td>4.65396</td>
<td>13.00142</td>
<td>log(totemp)</td>
</tr>
<tr>
<td>age</td>
<td>4075</td>
<td>50</td>
<td>8.186503</td>
<td>1</td>
<td>71</td>
<td>age of the plan</td>
</tr>
<tr>
<td>sole</td>
<td>4075</td>
<td>2</td>
<td>.3693252</td>
<td>0</td>
<td>1</td>
<td>=1 if only pension plan</td>
</tr>
</tbody>
</table>

Wooldridge (2011) gives an example of a fractional probit model with heteroskedasticity. He recommends using robust standard errors (otherwise the standard errors are too large; you can confirm this by rerunning the following example with vce(oim); you will see dramatic differences in the test statistics and standard errors.) He wrote his own program for this but fracglm can easily reproduce his results.

. fracglm prate mrate ltotemp age i.sole, het(mrate ltotemp age i.sole) link(p)

Heteroskedastic Fractional Probit Regression
Number of obs =       4075
Wald chi2(4)    =     152.29
Prob > chi2     =     0.0000
Log pseudolikelihood = -1674.5212                 Pseudo R2       =     0.0632
------------------------------------------------------------------------------
|               Robust
| prate | Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>prate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mrate</td>
<td>1.3847</td>
<td>.2239</td>
<td>6.19</td>
<td>0.000</td>
<td>.9459349    1.823454</td>
</tr>
<tr>
<td>ltotemp</td>
<td>-.1495</td>
<td>.0139</td>
<td>-10.71</td>
<td>0.000</td>
<td>-.1768825   -.1221367</td>
</tr>
<tr>
<td>age</td>
<td>.0671</td>
<td>.0101</td>
<td>6.66</td>
<td>0.000</td>
<td>.0473474    .086797</td>
</tr>
<tr>
<td>1.sole</td>
<td>-.1183</td>
<td>.0932</td>
<td>-1.27</td>
<td>0.205</td>
<td>-.3010046   -.0644645</td>
</tr>
<tr>
<td>_cons</td>
<td>1.68</td>
<td>.1059</td>
<td>15.86</td>
<td>0.000</td>
<td>1.471818    1.886936</td>
</tr>
</tbody>
</table>

------------------------------------------------------------------------------
| lnsigma | Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>mrate</td>
<td>.2404</td>
<td>.0538</td>
<td>4.47</td>
<td>0.000</td>
<td>.1349479    .3547662</td>
</tr>
<tr>
<td>ltotemp</td>
<td>.0375</td>
<td>.0144</td>
<td>2.60</td>
<td>0.009</td>
<td>.0092525    .0657845</td>
</tr>
<tr>
<td>age</td>
<td>.0171</td>
<td>.0027</td>
<td>6.29</td>
<td>0.000</td>
<td>.0118229    .0225199</td>
</tr>
<tr>
<td>1.sole</td>
<td>-.1628</td>
<td>.0631</td>
<td>-2.58</td>
<td>0.010</td>
<td>-.2864417   -.0390674</td>
</tr>
</tbody>
</table>

[NOTE: vce(robust) is the default for both fracglm and fracreg. If you are using fracglm with a binary dependent variable, you may wish to specify vce(oim) instead.]

Wooldridge (2011) notes that a simple Wald test can be used to determine whether the coefficients in the heteroskedasticity equation are significantly different from zero. (I believe this is better than a likelihood ratio test because LR tests are problematic when using robust standard errors).

. test [lnsigma]

( 1) [lnsigma]mrate = 0
( 2) [lnsigma]ltotemp = 0
( 3) [lnsigma]age = 0
( 4) [lnsigma]sole = 0
( 5) [lnsigma]1.sole = 0
Constraint 4 dropped
    chi2(  4) = 109.26
    Prob > chi2 = 0.0000
You interpret these results pretty much the same way you would interpret the results from a hetprobit model. A higher match rate, an older fund, and having fewer employees all increase participation rates.

If you want to make results more tangible, you can use methods like we have used before. For example,

```
. margins, dydx(*)
```

<table>
<thead>
<tr>
<th></th>
<th>Delta-method</th>
<th></th>
<th></th>
<th></th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dy/dx</td>
<td>Std. Err.</td>
<td>z</td>
<td>P&gt;</td>
<td>z</td>
</tr>
<tr>
<td>mrate</td>
<td>.159632</td>
<td>.0114214</td>
<td>13.98</td>
<td>0.000</td>
<td>.1372466 .1820175</td>
</tr>
<tr>
<td>ltotemp</td>
<td>-.0306262</td>
<td>.0020032</td>
<td>-15.29</td>
<td>0.000</td>
<td>-.0345524 -.0267</td>
</tr>
<tr>
<td>age</td>
<td>.0065659</td>
<td>.0006323</td>
<td>10.38</td>
<td>0.000</td>
<td>.0053266 .0078052</td>
</tr>
<tr>
<td>1.sole</td>
<td>.016089</td>
<td>.0061307</td>
<td>2.62</td>
<td>0.009</td>
<td>.0040729 .028105</td>
</tr>
</tbody>
</table>

Note: dy/dx for factor levels is the discrete change from the base level.

```
. mcp mrate, at1(0 (.05) 2)
```

Wooldridge (2011) also says you “should do a comparison of average partial effects [aka average marginal effects] between ordinary fractional probit and heteroskedastic fractional probit.” Non-heteroskedastic models can also be estimated with fracglm:
. fracglm prate mrate ltotemp age i.sole, link(p)

Fractional Probit Regression

Number of obs = 4075
Wald ch2(4) = 695.89
Prob > ch2 = 0.0000

Log pseudolikelihood = -1681.9607
Pseudo R2 = 0.0591

------------------------------------------------------------------------------
|               Robust
|     Coef.   Std. Err.     z     P>|z|    [95% Conf. Interval]
-------------+----------------------------------------------------------------
prate |    .595573    .038756   15.37   0.000     .519612    .671533
mrate | -.117285   .0080003  -14.66   0.000    -.132965    -.101605
ltotemp |  -.117285   .0080003  -14.66   0.000    -.132965    -.101605
age |   .018026    .001422   12.68   0.000     .015239    .020813
1.sole |   .094416    .027169    3.48   0.001     .041164    .147672
_cons |   1.42885   .0593694   24.07   0.000     1.31249    1.545216
------------------------------------------------------------------------------

. margins, dydx(*)

Average marginal effects

Number of obs = 4075
Model VCE : Robust

Expression : Pr(prate), predict()
dy/dx w.r.t. : mrate ltotemp age 1.sole

------------------------------------------------------------------------------
| Delta-method
|     dy/dx   Std. Err.     z     P>|z|    [95% Conf. Interval]
-------------+----------------------------------------------------------------
prate |    .1363    .00881   15.47   0.000     .119    .1535
mrate |   -.0268    .00184   -14.54   0.000    -.03   -.0232
ltotemp |  -.0268    .00184   -14.54   0.000    -.03   -.0232
age |   .0041    .00033    12.59   0.000     .0035    .0047
1.sole |   .0213    .00604    3.53   0.000     .01  .0332
_cons |   1.68    .10589    15.86   0.000     1.48    1.89
------------------------------------------------------------------------------

Note: dy/dx for factor levels is the discrete change from the base level.

Based on the earlier Wald test we would prefer the heteroskedastic model. You can also see that there are some modest differences in the Average Marginal Effects estimated by the two models.

Using Stata 14’s fracreg instead, the heteroskedastic and non-heteroskedastic models are estimated by

. use https://www3.nd.edu/~rwilliam statafiles/401kpart, clear
. fracreg probit prate mrate ltotemp age i.sole, het(mrate ltotemp age i.sole) nolog

Fractional heteroskedastic probit regression

Number of obs = 4,075
Wald ch2(5) = 152.30
Prob > ch2 = 0.0000

Log pseudolikelihood = -1674.5212
Pseudo R2 = 0.0088

------------------------------------------------------------------------------
|               Robust
|     Coef.   Std. Err.     z     P>|z|    [95% Conf. Interval]
-------------+----------------------------------------------------------------
prate |     1.38    .22384    6.19   0.000     .94    1.82
mrate |   -.1495    .0140   -10.71   0.000    -.17   -.122
ltotemp |  -.0269    .0014    -14.54   0.000    -.03   -.0232
age |   .067   .0101    6.67   0.000     .05    .087
1.sole |  -.12    .093    -1.27   0.205    -.3    .0645
:_cons |   1.68    .1059    15.86   0.000     1.48    1.89
-------------+----------------------------------------------------------------
lnsigma |    .24    .0538    4.47   0.000     .14      3.45
mrate |    .03    .0144    2.60   0.009     .00    .0657
ltotemp |   .0171    .0027    6.29   0.000     .01    .0225
age |    .01    .0027    2.60   0.009     .00    .0066
1.sole |   -.16    .063    -2.58   0.010    -.28    -.039
(_cons |    1.67    .1059    15.86   0.000     1.48    1.89
------------------------------------------------------------------------------
. fracreg probit prate mrate ltotemp age i.sole, nolog

Fractional probit regression

Number of obs =  4,075
Wald chi2(5)   =  695.89
Prob > chi2    =  0.0000
Log pseudolikelihood = -1681.9607
Pseudo R2      =  0.0591

------------------------------------------------------------------------------
|               Robust          
|               Coef.  Std. Err.  z     P>|z|    [95% Conf. Interval]
-------------+------------------------------------------------------------------
prate        |  .5955726   .038756  15.37    0.000   .5196123   .6715329
mrate        | -.1172851  .0080003 -14.66    0.000  -.1329655  -.1016048
ltotemp      |  .0180259   .0014218  12.68    0.000   .0152392   .0208126
age          |  .0944158   .0271696   3.48    0.001   .0411645   .1476672
1.sole       |  1.428854   .0593694  24.07    0.000   1.312493   1.545216
_cons        |  1.428854   .0593694  24.07    0.000   1.312493   1.545216
------------------------------------------------------------------------------

Other Comments on Fractional Response Models:

1. Other than the fact that the heteroskedastic model fits better in this case, what is the rationale for it? I asked Jeffrey Wooldridge about this and he emailed me the following:

   I think of it [the heteroskedastic model] mainly as a simple way to get a more flexible functional form. But this can also be derived from a model where, say, y(i) is the fraction of successes out of n(i) Bernoulli trials, where each binary outcome, say w(i,j), follows a heteroskedastic probit. Then E(y(i)|x(i)) would have the form estimated by your Stata command.

   Or, we could start with an omitted variable formulation: E[y(i)|x(i),c(i)] = PHI[x(i)*b + c(i)], where the omitted variable c(i) is distributed as Normal with mean zero and variance h(x(i)). As an approximation, we might use an exponential function for 1 + h(x(i)), and then that gives the model, too.

2. As noted in Wooldridge (2011) and in the Stata FAQ cited above, the glm command can also be used to estimate non-heteroskedastic models. Specify family(binomial) and either link(p) or link(l). These are the same results that fracglm gave for the non-heteroskedastic model.

   . glm prate mrate ltotemp age i.sole, vce(robust) link(p) family(binomial) nolog
   note: prate has noninteger values

   Generalized linear models
   No. of obs =  4075
   Optimization : ML
   Residual df  =  4070
   Scale parameter =  1
   (1/df) Deviance =  .2176709
   (1/df) Pearson =  .2203313
   Variance function: V(u) = u*(1-u/1)  [Binomial]
   Link function  : g(u) = invnorm(u)  [Probit]
   AIC            =  .6352659
   BIC            =  -.32946.47
   Log pseudolikelihood = -1289.354251

Analyzing Proportions: Fractional Response and Zero One Inflated Beta Models
|               Robust    |       Coef. |     Std. Err. |      z   |     P>|z|     | [95% Conf. Interval] |
|----------------------|-------------|---------------|----------|-----------|---------------------|
| mrate | .5955726    | .038756       | 15.37    | 0.000     | .5196123            | .6715329            |
| ltotemp | -.1172851   | .0080003      | -14.66   | 0.000     | -.1329655           | -.1016048           |
| age | .0180259     | .0014218      | 12.68    | 0.000     | .0152392            | .0208126            |
| 1.sole | .0944158    | .0271696      | 3.48     | 0.001     | .0411645            | .1476672            |
| _cons | 1.428854    | .0593694      | 24.07    | 0.000     | 1.312493            | 1.545216            |

Zero Inflated Beta Models. The Stata FAQ (http://www.stata.com/support/faqs/statistics/logit-transformation/) warns that other types of models may be advisable depending on why the 0s or 1s exist. From the FAQ (it talks about a logit transformation but the same is true for probit):

A traditional solution to this problem [the dependent variable is a proportion] is to perform a logit transformation on the data. Suppose that your dependent variable is called y and your independent variables are called X. Then, one assumes that the model that describes y is

\[ y = \text{invlogit}(XB) \]

If one then performs the logit transformation, the result is

\[ \ln\left(\frac{y}{1-y}\right) = XB \]

We have now mapped the original variable, which was bounded by 0 and 1, to the real line. One can now fit this model using OLS or WLS, for example by using `regress`. Of course, one cannot perform the transformation on observations where the dependent variable is zero or one; the result will be a missing value, and that observation would subsequently be dropped from the estimation sample.

A better alternative is to estimate using `glm` with `family(binomial)`, `link(logit)`, and `robust`; this is the method proposed by Papke and Wooldridge (1996).

In either case, there may well be a substantive issue of interpretation. Let us focus on interpreting zeros: the same kind of issue may well arise for ones. Suppose the y variable is proportion of days workers spend off sick. There are two extreme possibilities. The first extreme is that all observed zeros are in effect sampling zeros: each worker has some nonzero probability of being off sick, and it is merely that some workers were not, in fact, off sick in our sample period. Here, we would often want to include the observed zeros in our analysis and the `glm` route is attractive. The second extreme is that some or possibly all observed zeros must be considered as structural zeros: these workers will not ever report sick, because of robust health and exemplary dedication. These are extremes, and intermediate cases are also common. In practice, it is often helpful to look at the frequency distribution: a marked spike at zero or one may well raise doubt about a single model fitted to all data.

A second example might be data on trading links between countries. Suppose the y variable is proportion of imports from a certain country. Here a zero might be structural if two countries never trade, say on political or cultural grounds. A model that fits over both the zeros and the nonzeros might not be advisable, so that a different kind of model should be considered.
Baum (2008) elaborates on the problem of structural zeros and 1s. He notes “the managers of a city that spends none of its resources on preschool enrichment programs have made a discrete choice. A hospital with zero heart transplants may be a facility whose managers have chosen not to offer certain advanced services. In this context, the glm approach, while properly handling both zeros and ones, does not allow for an alternative model of behavior generating the limit values.”

He suggests alternatives such as the “zero-inflated beta” model, which allows for zero values (but not unit values) in the proportion and for separate variables influencing the zero and nonzero values (i.e. something similar to the zero-inflated or hurdle models that you have for count data). A one-inflated beta model allows for separate variables influencing the one and non-one values.

Both the zero and one inflated beta models can be estimated via Maarten Buis’s zoib program, available from SSC. The help file for zoib says

zoib fits by maximum likelihood a zero one inflated beta distribution to a distribution of a variable depvar. depvar ranges between 0 and 1: for example, it may be a proportion. It will estimate the probabilities of having the value 0 and/or 1 as separate processes. The logic is that we can often think of proportions of 0 or 1 as being qualitatively different and generated through a different process as the other proportions.

Here is how we can apply the one-inflated beta model to the current data. In these data, no company has a value of zero, but about a third of the cases have a value of 1, so we use the oneinflate option to model the 1s separately.

```
. zoib prate mrate ltotemp age i.sole, oneinflate( mrate ltotemp age i.sole)
Iteration 0:   log likelihood = -1350.3099
Iteration 1:   log likelihood = -881.01326
Iteration 2:   log likelihood = -860.4238
Iteration 3:   log likelihood = -860.34541
Iteration 4:   log likelihood = -860.34541
ML fit of oib                     Number of obs   =       4075
Wald chi2(4)    =     438.05
Log likelihood = -860.34541                     Prob > chi2     =     0.0000
------------------------------------------------------------------------------
prate |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
proportion   |
         mrate |   .7549614   .0570378    13.24   0.000     .6431693    .8667535
        ltotemp |  -.1138555   .0111967   -10.17   0.000    -.1358006   -.0919104
         age |   .0236683   .0022057    10.73   0.000     .0193452    .0279915
        _cons |   1.507286   .0870497    17.32   0.000     1.336672    1.677901
oneinflate   |
         mrate |   .9482935   .0859378    11.03   0.000     .7798591    1.116728
        ltotemp |  -.2918532    .027768   -10.51   0.000    -.3462776   -.2374288
         age |   .0190046   .0038664     4.92   0.000     .011426   .0265826
        _cons |   .6041419   .0762853     7.92   0.000     .4546255    .7536583
ln_phi       |
         _cons |   1.621576   .0262855    61.69   0.000     1.570057    1.673094
------------------------------------------------------------------------------
```
The one-inflate equation shows that companies with higher match rates, fewer total employees, older plans, and that have only one pension plan available are more likely to have 100% participation in their plans. When participation is not 100%, these same variables (except sole) also are associated with higher participation rates.

**Other Models & Programs.** I am not familiar with most of these, but the help for zoib suggests that these programs may also sometimes be helpful when modeling proportions:

- **betafit** fits by maximum likelihood a two-parameter beta distribution to a distribution of a variable depvar. depvar ranges between 0 and 1: for example, it may be a proportion.

- **dirifit** fits by maximum likelihood a Dirichlet distribution to a set of variables depvarlist. Each variable in depvarlist ranges between 0 and 1 and all variables in depvarlist must, for each observation, add up to 1: for example, they may be proportions.

- **fmlogit** fits by quasi maximum likelihood a fractional multinomial logit model. Each variable in depvarlist ranges between 0 and 1 and all variables in depvarlist must, for each observation, add up to 1: for example, they may be proportions. It is a multivariate generalization of the fractional logit model proposed by Papke and Wooldridge (1996).

For the latter two programs, the help files give as examples models where the dependent variables are the proportions of a municipality’s budget that are spent on governing, public safety, education, recreation, social work, and urban planning. Independent variables include whether or not there are any left-wing parties in city government. If I understand the models correctly, the coefficients tell you how the independent variables increase or decrease the proportion of spending in each area. For example, the results show that when there is no left wing party in city government, less of the city budget tends to get spent on education.

- **fracivp – fractional iv probit.** fracivp is a beta program adapted from Stata 12’s ivprobit program. It relaxes the assumption that the dependent variable be coded 0/1 and allows it to be a proportion instead. fracivp estimates Fractional Response Probit models with continuous endogenous regressors. This is a use at your own risk program; it seems to work ok but I haven’t been able to verify it against other programs or known results. Comments are welcome. To get fracivp, from within Stata type

```stata
net install fracivp, from(https://www3.nd.edu/~rwilliam/stata)
```