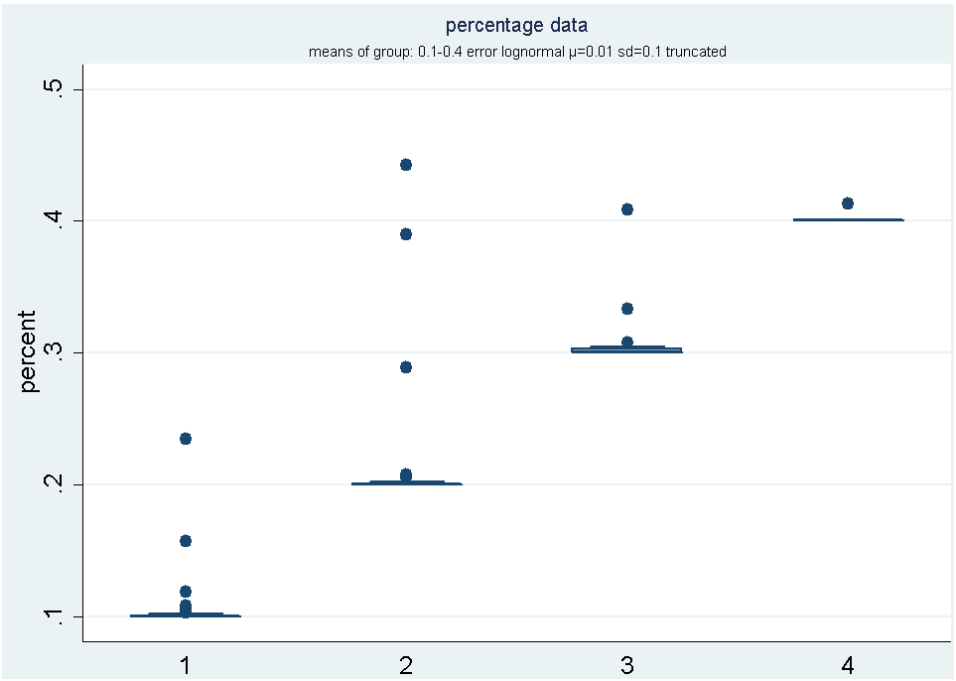
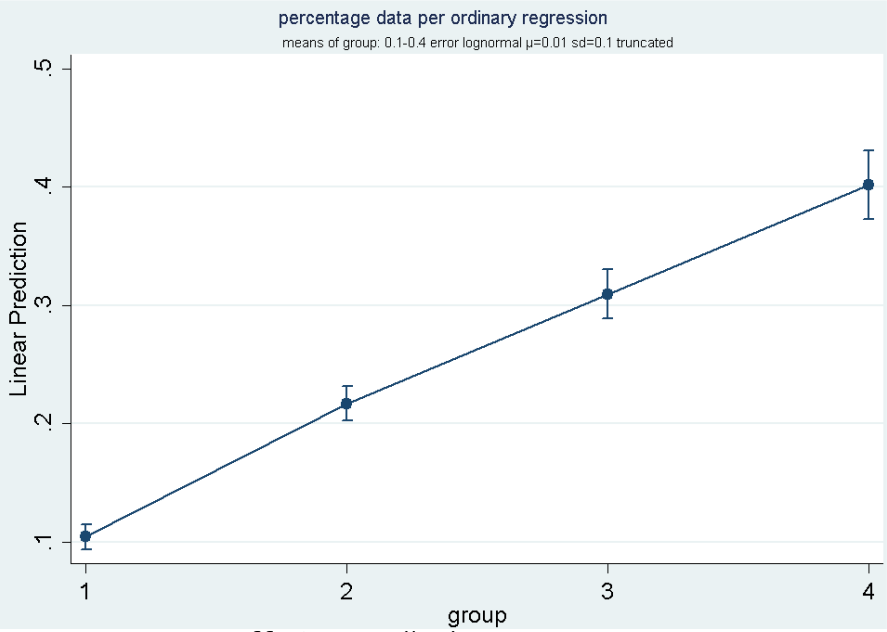
```
/* Literatur see below
   In order to show differences in respect to estimation (bias) and precision
   Data are simulated for 4 groups with mean 0.1 to 0.4
   The error term is lognormal distributed with mean=0.01 and std=0.1
   Heteroscedasticity appears data are truncated to [0;1]

   All regressions should be done with "robust" correction

*/

* Simulation
clear
input percent group freq
.10 1 64
.20 2 32
.30 3 16
.40 4  8
end
label variable group "group"
expand freq
sort group

scalar e=0.01
scalar s=1

scalar b=sqrt(ln(s^2/e^2+1))
scalar a=ln(e)-b^2/2

set seed 3

replace percent=min(max(percent+exp(rnormal(a,b)),0),1)
tabstat percent, by(group) stat(N mean median sd min max skew) format(%8.3f)

graph box percent, over(group) ///
      title(percentage data, size(small)) ///
      subtitle(means of group: 0.1-0.4 error lognormal μ=0.01 sd=0.1 truncated,
size(vsmall))
graph rename data, replace
```

```
********************************************** per ordinary reg
reg percent i.group
hettest
estimates store OrdReg
margins group, mcomp(bon)
marginsplot, title(percentage data per ordinary regression, size(small)) ///
             subtitle(means of group: 0.1-0.4 error lognormal μ=0.01 sd=0.1
truncated, size(vsmall))
graph rename per_ordinaryreg, replace
pwcompare group, effects mcomp(bon)
********************************************** per robust reg
reg percent i.group, robust
estimates store RobReg

margins group, mcomp(bon)
marginsplot, title(percentage data per robust regression, size(small)) ///
             subtitle(means of group: 0.1-0.4 error lognormal μ=0.01 sd=0.1
truncated, size(vsmall))
graph rename per_robustreg, replace
pwcompare group, effects mcomp(bon)

********************************************** per robust logit
gen logitperc=logit(percent)
reg logitperc i.group, robust
estimates store RobLogit
margins group, expression(invlogit(predict(xb))) mcomp(bon)
marginsplot, title(percentage data per robust logit, size(small)) ///
             subtitle(means of group: 0.1-0.4 error lognormal μ=0.01 sd=0.1
truncated, size(vsmall))
graph rename per_robustlogit, replace
margins group, expression(invlogit(predict(xb))) mcomp(bon) pwcompare(effects)
********************************************** per robust glm
glm percent i.group, family(binomial) link(logit) robust
estimates store RobGLM
margins group, mcomp(bon)
marginsplot, title(percentage data per robust glm, size(small)) ///
             subtitle(means of group: 0.1-0.4 error lognormal μ=0.01 sd=0.1
truncated, size(vsmall))
margins group, expression(invlogit(predict(xb))) mcomp(bon) pwcompare(effects)
graph rename per_robustglm, replace

********************************************** Graphs
graph combine data per_ordinaryreg per_robustreg per_robustlogit per_robustglm,
ycommon
graph export "percent ordinary vs robust.wmf", replace

graph combine  per_robustreg per_robustlogit per_robustglm, ycommon
graph export "percent various robust methods.wmf", replace

********************************************** Results (only reg because logit
and GLM not transformed)
esttab OrdReg RobReg, b(%10.4f) se mtitles title(Compare various models)
```

```
. tabstat percent, by(group) stat(N mean median sd min max skew) format(%8.3f)

Summary for variables: percent
     by categories of: group (group)

   group |         N      mean       p50        sd       min       max  skewness
---------+--------------------------------------------------------------------------
       1 |    64.000     0.104     0.100     0.018     0.100     0.235     6.297
       2 |    32.000     0.217     0.200     0.055     0.200     0.443     3.334
       3 |    16.000     0.310     0.300     0.028     0.300     0.409     3.161
       4 |     8.000     0.402     0.400     0.004     0.400     0.413     2.233
---------+--------------------------------------------------------------------------
   Total |   120.000     0.182     0.108     0.100     0.100     0.443     1.007
--------------------------------------------------------------------------------
```

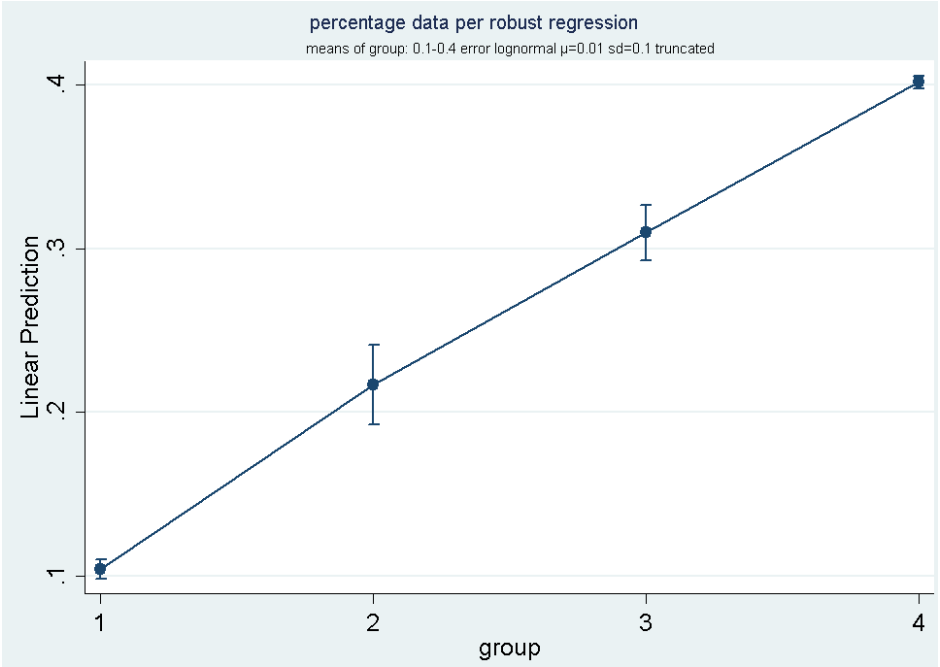


```
. ****************************************** per ordinary reg
. reg percent i.group
      Source |       SS       df       MS              Number of obs =     120
-------------+------------------------------           F(  3,   116) =  329.15
       Model |  1.07428071     3  .358093571           Prob > F      =  0.0000
    Residual |  .126202043   116  .001087949           R-squared     =  0.8949
-------------+------------------------------           Adj R-squared =  0.8922
       Total |  1.20048275   119   .01008809           Root MSE      =  .03298
------------------------------------------------------------------------------
     percent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |
          2  |   .1128208   .0071413    15.80   0.000     .0986766    .1269649
          3  |   .2055694   .0092193    22.30   0.000     .1873094    .2238295
          4  |   .2975948    .012369    24.06   0.000     .2730964    .3220932
             |
       _cons |   .1041657    .004123    25.26   0.000     .0959996    .1123319
------------------------------------------------------------------------------
. hettest
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
…
         Prob > chi2  =   0.0378
. margins group, mcomp(bon)
…
------------------------------------------------------------------------------
             |            Delta-method    Bonferroni           Bonferroni
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |
          1  |   .1041657    .004123    25.26   0.000     .0938677    .1144638
          2  |   .2169865   .0058308    37.21   0.000     .2024228    .2315501
          3  |   .3097352    .008246    37.56   0.000      .289139    .3303313
          4  |   .4017605   .0116616    34.45   0.000     .3726332    .4308878
------------------------------------------------------------------------------
```



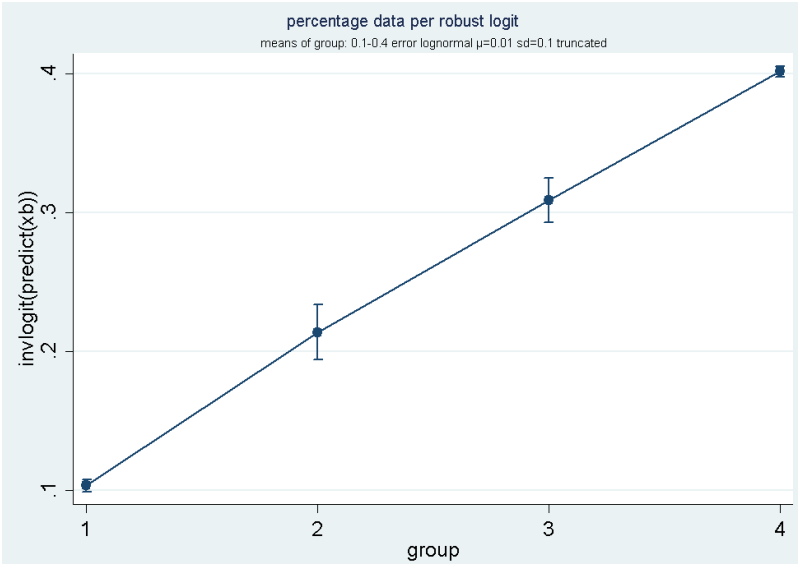

```
. pwcompare group, effects mcomp(bon)
------------------------------------------------------------------------------
             |                              Bonferroni           Bonferroni
             |   Contrast   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |
     2 vs 1  |   .1128208   .0071413    15.80   0.000     .0936518    .1319897
     3 vs 1  |   .2055694   .0092193    22.30   0.000     .1808224    .2303165
     4 vs 1  |   .2975948    .012369    24.06   0.000     .2643932    .3307964
     3 vs 2  |   .0927487   .0100993     9.18   0.000     .0656396    .1198577
     4 vs 2  |    .184774   .0130381    14.17   0.000     .1497764    .2197716
     4 vs 3  |   .0920254   .0142825     6.44   0.000     .0536874    .1303633
------------------------------------------------------------------------------
```

```
. ********************************************** per robust reg
. reg percent i.group, robust
Linear regression                                      Number of obs =     120
                                                       F(  3,   116) = 3962.51
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.8949
                                                       Root MSE      =  .03298
------------------------------------------------------------------------------
             |               Robust
     percent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |
           2 |   .1128208   .0099969    11.29   0.000     .0930205     .132621
           3 |   .2055694   .0071764    28.65   0.000     .1913557    .2197832
           4 |   .2975948   .0027407   108.58   0.000     .2921665    .3030231
             |
       _cons |   .1041657   .0022995    45.30   0.000     .0996113    .1087201
------------------------------------------------------------------------------
. margins group, mcomp(bon)
------------------------------------------------------------------------------
             |            Delta-method    Bonferroni           Bonferroni
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |
           1 |   .1041657   .0022995    45.30   0.000     .0984223    .1099091
           2 |   .2169865   .0097289    22.30   0.000     .1926866    .2412864
           3 |   .3097352    .006798    45.56   0.000     .2927557    .3267146
           4 |   .4017605   .0014913   269.41   0.000     .3980358    .4054853
------------------------------------------------------------------------------
```



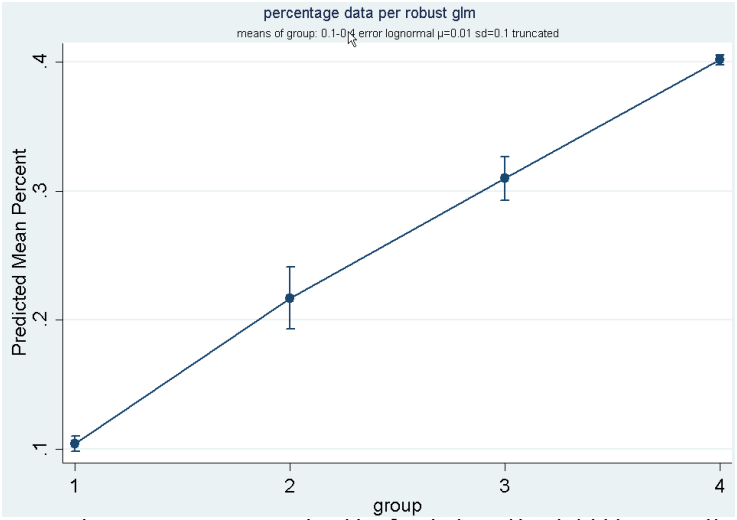

```
. pwcompare group, effects mcomp(bon)
…
------------------------------------------------------------------------------
             |                             Bonferroni           Bonferroni
             |   Contrast   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |
     2 vs 1  |   .1128208   .0099969    11.29   0.000     .0859864    .1396551
     3 vs 1  |   .2055694   .0071764    28.65   0.000     .1863061    .2248327
     4 vs 1  |   .2975948   .0027407   108.58   0.000      .290238    .3049516
     3 vs 2  |   .0927487   .0118686     7.81   0.000     .0608902    .1246071
     4 vs 2  |    .184774   .0098425    18.77   0.000     .1583542    .2111939
     4 vs 3  |   .0920254   .0069597    13.22   0.000     .0733438    .1107069
------------------------------------------------------------------------------
```

```
. ********************************************** per robust logit
. gen logitperc=logit(percent)
. reg logitperc i.group, robust
Linear regression                                       Number of obs =     120
                                                        F(  3,   116) = 2956.29
                                                        Prob > F      =  0.0000
                                                        R-squared     =  0.9227
                                                        Root MSE      =  .17931
------------------------------------------------------------------------------
             |               Robust
   logitperc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |
          2  |   .8593403   .0507076    16.95   0.000     .7589075    .9597731
          3  |   1.357005   .0349643    38.81   0.000     1.287754    1.426256
          4  |   1.762558   .0190093    92.72   0.000     1.724908    1.800208
             |
       _cons |  -2.160721   .0179759  -120.20   0.000    -2.196325   -2.125118
------------------------------------------------------------------------------
. margins group, expression(invlogit(predict(xb))) mcomp(bon)
…
------------------------------------------------------------------------------
             |            Delta-method    Bonferroni           Bonferroni
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |
          1  |   .1033336   .0016656    62.04   0.000     .0991735    .1074937
          2  |   .2139327   .0079735    26.83   0.000     .1940173    .2338481
          3  |   .3092311    .006406    48.27   0.000     .2932309    .3252314
          4  |   .4017537   .0014859   270.38   0.000     .3980425     .405465
------------------------------------------------------------------------------
```



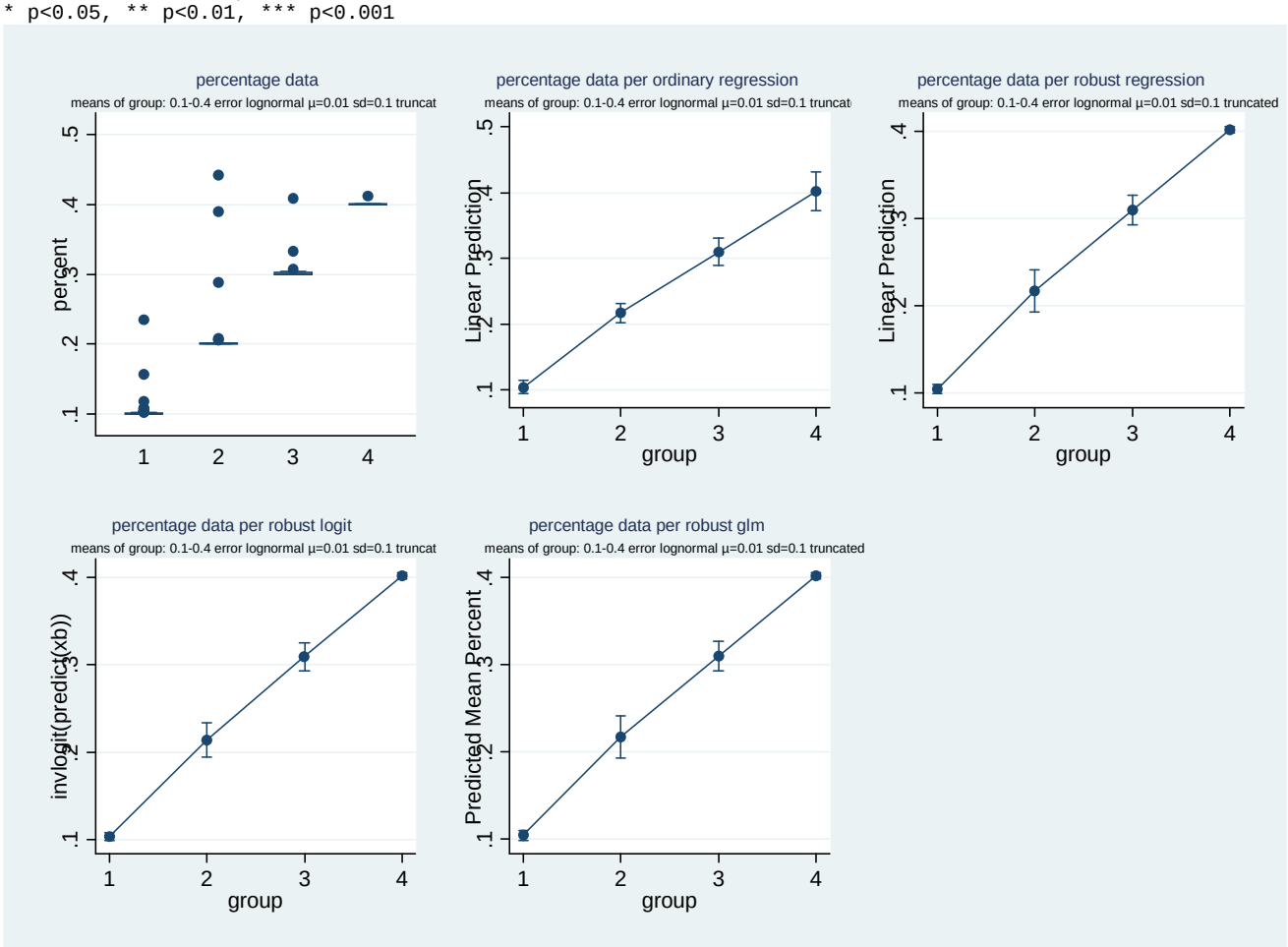

```
. margins group, expression(invlogit(predict(xb))) mcomp(bon) pwcompare(effects)
…
------------------------------------------------------------------------------
             |            Delta-method    Bonferroni           Bonferroni
             |   Contrast   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |
     2 vs 1  |   .1105991   .0081456    13.58   0.000      .089109    .1320892
     3 vs 1  |   .2058975    .006619    31.11   0.000      .188435      .22336
     4 vs 1  |   .2984201    .002232   133.70   0.000     .2925315    .3043088
     3 vs 2  |   .0952984    .010228     9.32   0.000     .0683142    .1222826
     4 vs 2  |    .187821   .0081107    23.16   0.000     .1664228    .2092192
     4 vs 3  |   .0925226    .006576    14.07   0.000     .0751733    .1098719
------------------------------------------------------------------------------
. /* Cave!
> Neither in this case nor in the case GLM below:
> the "margins, exp( )... pwcompare..." command can *** not *** be abbreviated to
> margins group, mcomp(bon) pwcompare(effects)
> as it is the case if a logistic regression is used
> */
```

```
. ****************************************** per robust glm
. glm percent i.group, family(binomial) link(logit) robust
Generalized linear models                         No. of obs      =       120
Optimization     : ML                             Residual df     =       116
                                                  Scale parameter =         1
Deviance         =  .7094192898                   (1/df) Deviance =  .0061157
Pearson          =  .8298653415                   (1/df) Pearson  =   .007154
Variance function: V(u) = u*(1-u/1)               [Binomial]
Link function    : g(u) = ln(u/(1-u))             [Logit]
                                                  AIC             =  .6859103
Log pseudolikelihood = -37.15461656               BIC             = -554.6396
------------------------------------------------------------------------------
             |               Robust
     percent |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |
          2  |   .8684574   .0615477    14.11   0.000     .7478261    .9890887
          3  |   1.350415   .0397169    34.00   0.000     1.272571    1.428258
          4  |   1.753637   .0250888    69.90   0.000     1.704464    1.802811
             |
       _cons |  -2.151772   .0243294   -88.44   0.000    -2.199457   -2.104087
------------------------------------------------------------------------------
. margins group, mcomp(bon) /* the margin command can be abbreviated *** here in GLM ***  */
…
------------------------------------------------------------------------------
             |            Delta-method    Bonferroni           Bonferroni
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |
          1  |   .1041657   .0022703    45.88   0.000     .0984952    .1098363
          2  |   .2169865   .0096055    22.59   0.000     .1929948    .2409781
          3  |   .3097352   .0067118    46.15   0.000     .2929711    .3264992
          4  |   .4017605   .0014724   272.87   0.000      .398083     .405438
------------------------------------------------------------------------------
```




```
. margins group, expression(invlogit(predict(xb))) mcomp(bon) pwcompare(effects)
------------------------------------------------------------------------------
             |            Delta-method    Bonferroni           Bonferroni
             |   Contrast   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |
     2 vs 1  |   .1128208   .0098701    11.43   0.000     .0867808    .1388607
     3 vs 1  |   .2055694   .0070854    29.01   0.000     .1868764    .2242624
     4 vs 1  |   .2975948   .0027059   109.98   0.000     .2904558    .3047338
     3 vs 2  |   .0927487   .0117181     7.92   0.000     .0618334     .123664
     4 vs 2  |    .184774   .0097177    19.01   0.000     .1591363    .2104117
     4 vs 3  |   .0920254   .0068714    13.39   0.000     .0738969    .1101538
------------------------------------------------------------------------------
. /* Cave!
> Neither in this case nor in the case RegLogit above:
> the "margins, exp( )... pwcompare..." command can *** not *** be abbreviated to
> margins group, mcomp(bon) pwcompare(effects)
> as it is the case if a logistic regression is used
> */
.
. graph rename per_robustglm, replace
(note: graph per_robustglm not found)
```

```
.
. ********************************************** Graphs
. graph combine data per_ordinaryreg per_robustreg per_robustlogit per_robustglm, ycommon

. graph export "percent ordinary vs robust.wmf", replace
(note: file percent ordinary vs robust.wmf not found)
(file D:\@Guidos Eigene Dateien auf D\Desktop\percent ordinary vs robust.wmf written in Windows
Metafile format)
. graph combine  per_robustreg per_robustlogit per_robustglm, ycommon
. graph export "percent various robust methods.wmf", replace
(note: file percent various robust methods.wmf not found)
(file D:\@Guidos Eigene Dateien auf D\Desktop\percent various robust methods.wmf written in Windows
Metafile format)

. ********************************************** Results (only reg because logit and GLM not
transformed)
. esttab OrdReg RobReg, b(%10.4f) se mtitles title(Compare various models)
Compare various models
--------------------------------------------
                      (1)             (2)
                   OrdReg          RobReg
--------------------------------------------
1bn.group               .               .
2.group            0.1128***       0.1128***
                  (0.0071)        (0.0100)
3.group            0.2056***       0.2056***
                  (0.0092)        (0.0072)
4.group            0.2976***       0.2976***
                  (0.0124)        (0.0027)
_cons              0.1042***       0.1042***
                  (0.0041)        (0.0023)
--------------------------------------------
N                      120             120
--------------------------------------------
Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```



```
/* if percentages are in the form
   #success #pop x1 x2 ...
   then use
Logistic regression for grouped data: blogit
Probit regression for grouped data: bprobit
Weighted least-squares logistic regression for grouped data: glogit
Weighted least-squares probit regression for grouped data: gprobit


* For just percentages use the following

http://www.stata.com/support/faqs/stat/logit.html

How do you fit a model when the dependent variable is a proportion?
Title          Logit transformation
Author  Allen McDowell, StataCorp
Nicholas J. Cox, Durham University, UK
Date    August 2001; updated August 2004

A traditional solution to this problem 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
                    1
        y =   ----------------
                1 + exp(-XB)
If one then performs the logit transformation, the result is
        ln( y / (1 - y) ) = 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).
At the time this article was published, Stata's glm command could not fit such models,
and this fact is noted in the article.
glm has since been enhanced specifically to deal with fractional response data.
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.
Reference

Papke, L. E. and J. Wooldridge. 1996.
    Econometric methods for fractional response variables with an application to 401(k) plan
participation rates. Journal of Applied Econometrics 11: 619–632.
*/
```