```
*First, download the following commands into your STATA version: tidwell, atkinson,
cwhetero,along with the data sets that go with them. The data consists of two
variables, area and perimeter and you may treat them as parameters taken from some
hospitals and you wish to construct a simple linear relationship between the two
variables after possibly transforming them to get a better fit.
.
. des
Contains data from ./weisberg.dta
  obs:            25                          Weisberg, 1985, p. 149
 vars:             3
 size:           400 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display    value
variable name   type   format     label      variable label
------------------------------------------------------------------------------
perimeter       float  %9.0g
area            float  %9.0g                 dependent variable
sqrta           float  %9.0g
-------------------------------------------------------------------------------
. sum
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     perimetr |         25      1.9736    1.270219        .41       4.78
        area |         25     17.1108    15.64692       1.13      51.19
       sqrta |         25    3.646396    1.993374   1.063015   7.154718
. list
     +-------------------------------+
     | perimetr     area        sqrta |
     |-------------------------------|
  1. |     3.48    38.83     6.231372 |
  2. |     3.69    43.92     6.627216 |
  3. |     1.43     9.14     3.023243 |
  4. |     2.05    16.66     4.081666 |
  5. |     3.05    36.16     6.013319 |
     |-------------------------------|
     |-------------------------------|
 25. |     2.99    29.75     5.454356 |
     +-------------------------------+

. graph twoway (scatter area perim) (lfit area perim,title("Plot of area versus perim"))

. fit area peri
      Source |       SS       df       MS              Number of obs =      25
-------------+------------------------------           F(  1,    23) =  597.04
       Model |  5657.86954     1  5657.86954            Prob > F      =  0.0000
    Residual |  217.960733    23   9.4765536           R-squared     =  0.9629
-------------+------------------------------           Adj R-squared =  0.9613
       Total |  5875.83027    24  244.826261           Root MSE      =  3.0784
------------------------------------------------------------------------------
        area |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   perimeter |    12.08766   .4946988     24.43   0.000     11.06429    13.11102
       _cons |   -6.745398   1.154252     -5.84   0.000    -9.13315   -4.357646
------------------------------------------------------------------------------

. predict resu,res
. predict yhatu,xb
. graph twoway (scatter resu yhatu ), yline(0) title("Residuals versus fitted from
untransformed response")
. graph save residplotuntransformed,replace
.
. *Let's transform the response using Box Cox transformation

. boxcox  area perime,nolog
Fitting comparison model
Fitting full model                              Number of obs    =         25
                                                LR chi2(1)       =      86.22
Log likelihood = -53.292366                     Prob > chi2      =      0.000
------------------------------------------------------------------------------
        area |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      /theta |   .6360019    .077472     8.21   0.000     .4841597    .7878442
```

```
------------------------------------------------------------------------------
Estimates of scale-variant parameters
------------------------------
             |      Coef.
-------------+----------------
Notrans      |
    perimetr |   4.406093
       _cons |  -1.701881
-------------+----------------
      /sigma |   .9022745
------------------------------
------------------------------------------------------------
   Test          Restricted      LR statistic       P-value
    H0:        log likelihood         chi2         Prob > chi2
------------------------------------------------------------
theta = -1       -100.95067          95.32             0.000
theta =  0       -69.400669          32.22             0.000
theta =  1       -62.541452          18.50             0.000
```

. *Result suggests that response should by transformed by 0.63

.

*Let's compare with Atkinson's technique for transforming the response:
atkinson area perime

score test for whether should transform area: t=4.868 p-value:0.0001
if above significant, transform area using 0.684 (round to .5)

*Result says response should by transformed by 0.68, which is similar to the one found by Box Cox's transformation. In practice, use the square root transformation.

.

. reg sqrta peri

```
      Source |       SS       df       MS              Number of obs =      25
-------------+------------------------------           F(  1,    23) =  686.31
       Model |  92.2726697     1  92.2726697            Prob > F      =  0.0000
    Residual |  3.09229873    23  .134447771            R-squared     =  0.9676
-------------+------------------------------           Adj R-squared =  0.9662
       Total |  95.3649684    24  3.97354035            Root MSE      =  .36667
------------------------------------------------------------------------------
       sqrta |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   perimeter |   1.543662   .058924    26.20   0.000     1.421768    1.665555
       _cons |   .5998246   .137484     4.36   0.000     .3154173    .8842319
------------------------------------------------------------------------------
```

. predict restos,res
. predict yhattos,xb

.

*Next consider transforming the independent variable

. tidwell sqrta perime

score test for whether should transform perimeter:t=-4.454 p-value: 0.0002
if above significant, transform perimeter using 0.564 (round to .5)

.

*Result says the independent variable should be transformed by 0.56 or by 0.5.

.

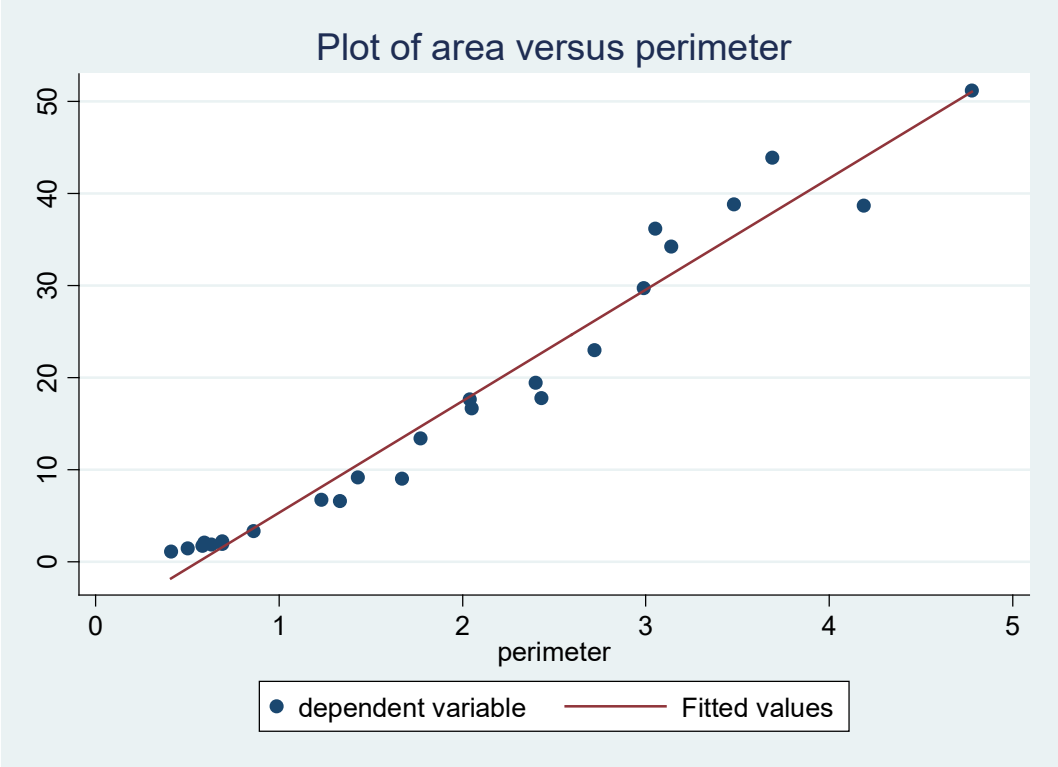
. reg sqrta sqrtp

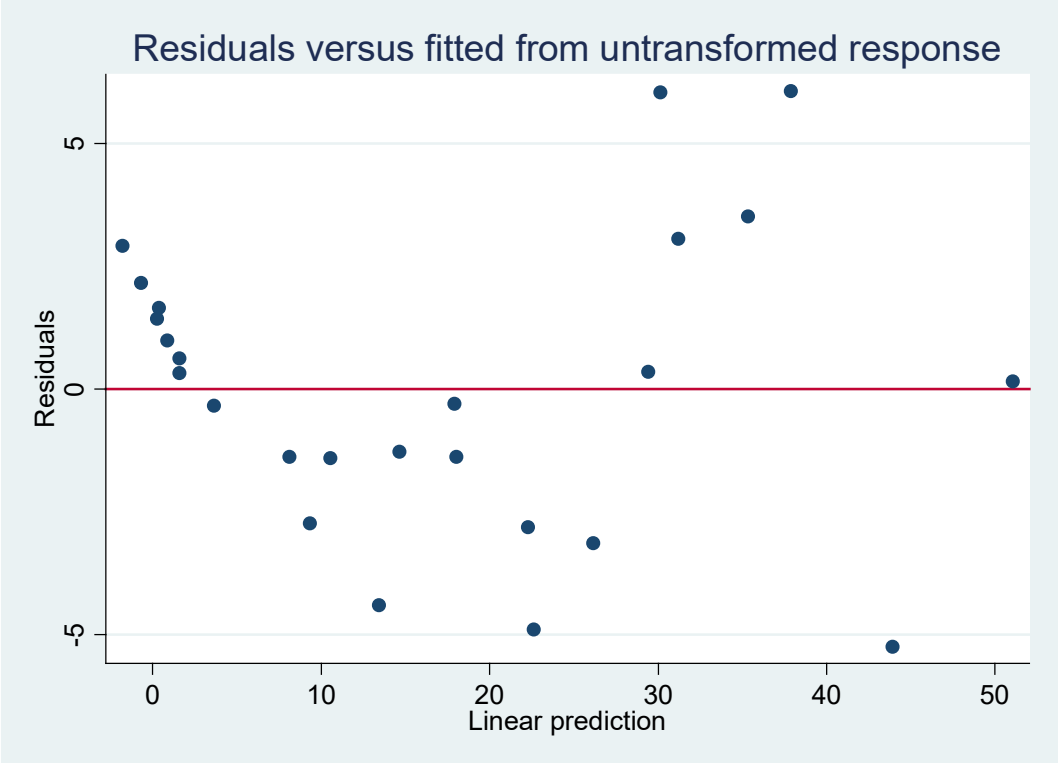
```
      Source |       SS       df       MS              Number of obs =      25
-------------+------------------------------           F(  1,    23) = 1147.50
       Model |  93.4910736     1  93.4910736            Prob > F      =  0.0000
    Residual |  1.87389479    23  .081473686            R-squared     =  0.9804
-------------+------------------------------           Adj R-squared =  0.9795
       Total |  95.3649684    24  3.97354035            Root MSE      =  .28544
------------------------------------------------------------------------------
       sqrta |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       sqrtp |   4.246699   .1253646    33.87   0.000     3.987362    4.506035
       _cons |  -1.997464   .1761184   -11.34   0.000    -2.361792   -1.633135
------------------------------------------------------------------------------
```

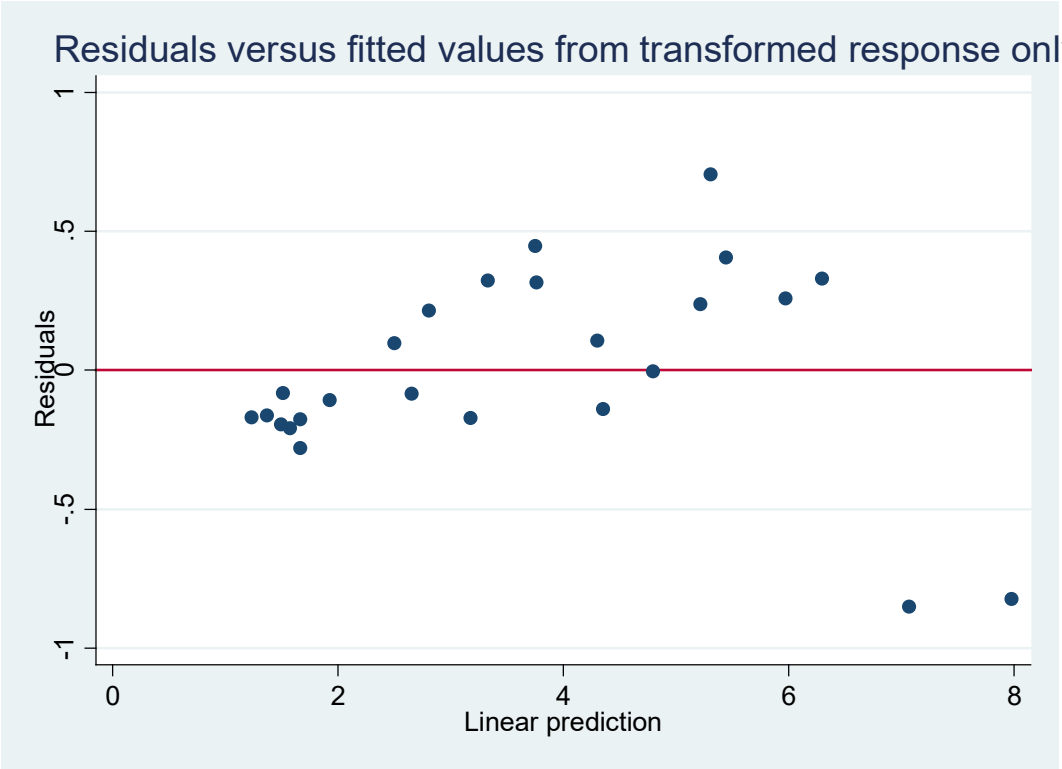
. predict restbs,res
. predict yhattbs,xb

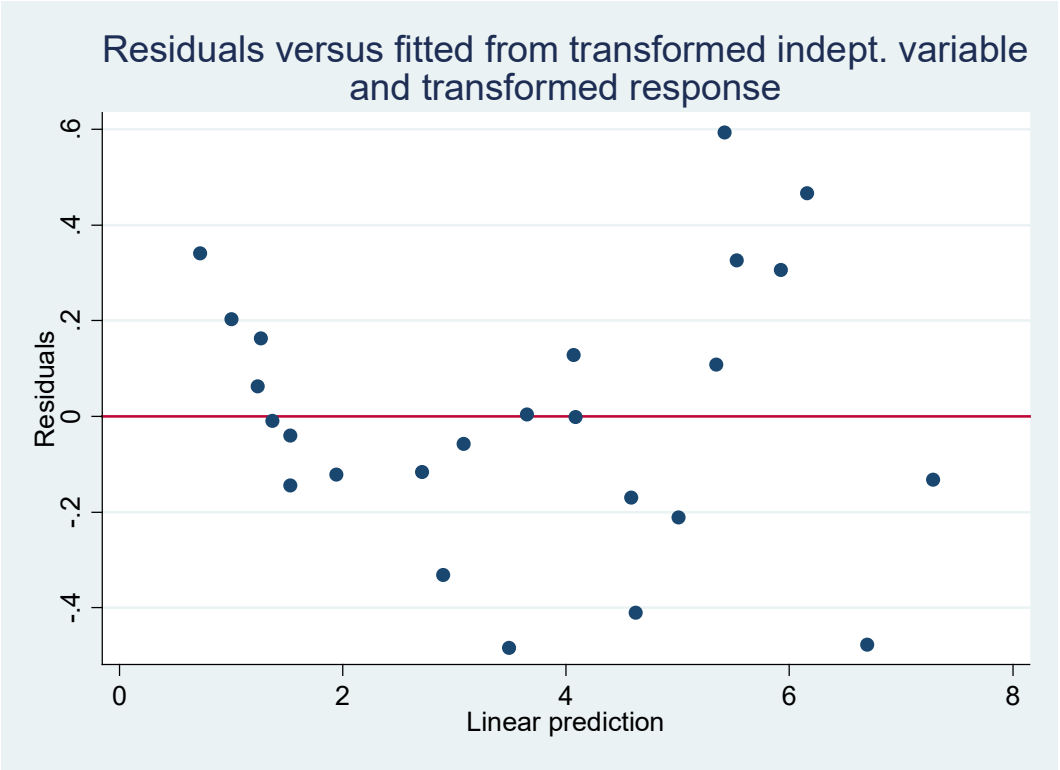
. graph twoway (scatter restbs yhattbs), yline(0) title("Residuals versus fitted from transformed response")
. graph save residplottbs,replace

```
.
*Let's see whether residuals seem more normally distributed after the transformations

. swilk res* yhat*
                    Shapiro-Wilk W test for normal data
    Variable |     Obs        W              V              z          Prob>z
-------------+-------------------------------------------------------------
        resu |      25     0.97443         0.711         -0.699      0.75761
      restos |      25     0.93272         1.869          1.279      0.10047
      restbs |      25     0.97944         0.571         -1.144      0.87377
       yhatu |      25     0.92995         1.947          1.362      0.08667
      yhatos |      25     0.92995         1.947          1.362      0.08667
     yhattbs |      25     0.94665         1.482          0.805      0.21052

. sktest res* yhat*
                    Skewness/Kurtosis tests for Normality
                                                  ------- joint ------
    Variable |     Obs    Pr(Skewness)    Pr(Kurtosis)   adj chi2(2)     Prob>chi2
-------------+-----------------------------------------------------------------
        resu |      25           0.626           0.983          0.24          0.8877
      restos |      25           0.205           0.250          3.26          0.1958
      restbs |      25           0.702           0.924          0.16          0.9253
       yhatu |      25           0.225           0.444          2.27          0.3215
      yhatos |      25           0.225           0.444          2.27          0.3215
     yhattbs |      25           0.810           0.044          4.28          0.1175

. *Let's look at heteroscedasticity
. qui reg area peri
. estat hettest
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
       Ho: Constant variance
       Variables: fitted values of area
       chi2(1) = 5.04
       Prob > chi2 = 0.0248

. qui reg sqrta sqrtp
. estat hettest
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
       Ho: Constant variance
       Variables: fitted values of sqrta
       chi2(1) = 3.43
       Prob > chi2  = 0.0640

*Results show untransformed model shows some evidence of heteroscedasticity but
not so for the transformed model
.
*One can also use the cwhetero command, which may produce slightly different result
as in this case.  The cwhetero command also gives the test result when the response
variance is assumed to depend on the predicted values. Note that this is an old
command that must follow the old commend fit (instead of reg) in STATA and has the
disadvantage that it eliminates any previous output results automatically.  This
explains why I have to generate the variable sqrtp below again.

. qui fit area peri
. cwhetero area peri

Score test (area perimeter)=5.542; chi-square p-value (df)=0.063 (2)
Score test (pred fit)=5.039; chi-square p-value (df)=0.025 (1)

. gen sqrtp=peri^.5
. qui fit sqrta sqrtp
. cwhetero sqrta sqrtp

Score test (sqrta sqrtp)=3.740; chi-square p-value (df)=0.154 (2)
Score test (pred fit)=3.431; chi-square p-value (df)=0.064 (1)
```


Plot of area versus perimeter
0
10
20
30
40
50
0
1
2
3
4
5
perimeter
dependent variable
Fitted values


Residuals versus fitted from untransformed response
Residuals
-5
0
5
0
10
20
30
40
50
Linear prediction


Residuals versus fitted values from transformed response only
Residuals
1
.5
0
-.5
-1
0
2
4
6
8
Linear prediction


Residuals versus fitted from transformed indept. variable and transformed response
Residuals
.6
.4
.2
0
-.2
-.4
0
2
4
6
8
Linear prediction

```
*************************** Here is the do file (minus all commentaries) ********************
capture log close
log using weisbergp149,text replace
sysuse weisberg,clear

des
sum
list
graph twoway (scatter area perim) (lfit area perim,title("Plot of area versus perimeter"))
graph save scatterareaperi,replace
reg area peri
predict resu,res
predict yhatu,xb
graph twoway (scatter resu yhatu ), yline(0) title("Residuals versus fitted values from untransformed
response")
graph save residplotnot,replace

boxcox  area perime
atkinson area perime

reg sqrta peri
predict restos,res
predict yhatos,xb
graph twoway (scatter resutos yhattos ), yline(0) title("Residuals versus fitted values from transformed
response only")
graph save residplottos,replace

tidwell sqrta perime
*Result says the independent variable should be transformed by 0.56 or roughly by its square root
gen sqrtp=perimet^.5
reg sqrta sqrtp
predict restbs,res
predict yhattbs,xb
graph twoway (scatter restbs yhatbs), yline(0) title("Residuals versus fitted from transformed independent"
" variable and transformed response")
graph save residplotbs,replace

swilk res* yhat*
sktest res* yhat*

qui reg area peri
estat hettest
qui reg sqrta sqrtp
estat hettest

qui fit area peri
cwhetero area peri
gen sqrtp=peri^.5
qui fit sqrta sqrtp
cwhetero sqrta sqrtp
log close
```