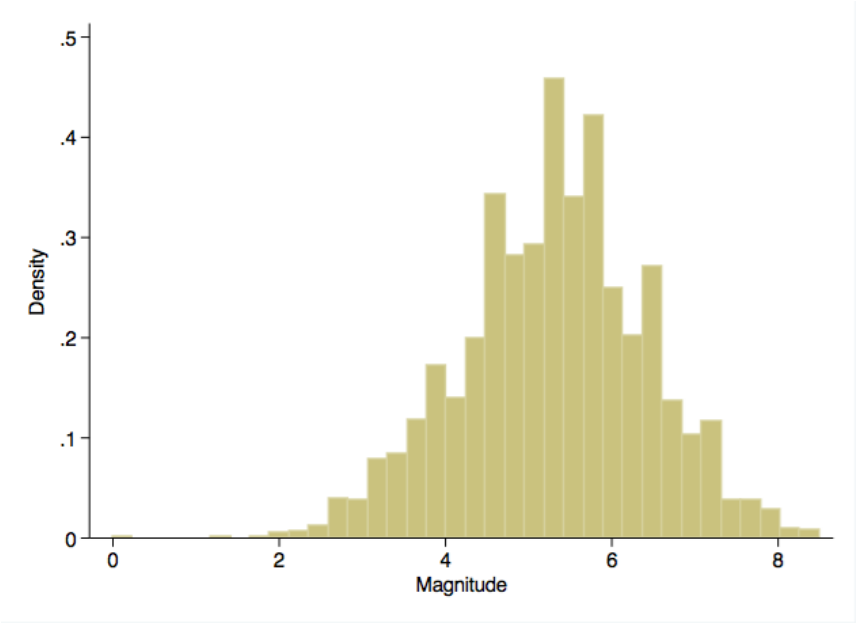
EDAV Project 2: Regression with Principal Components

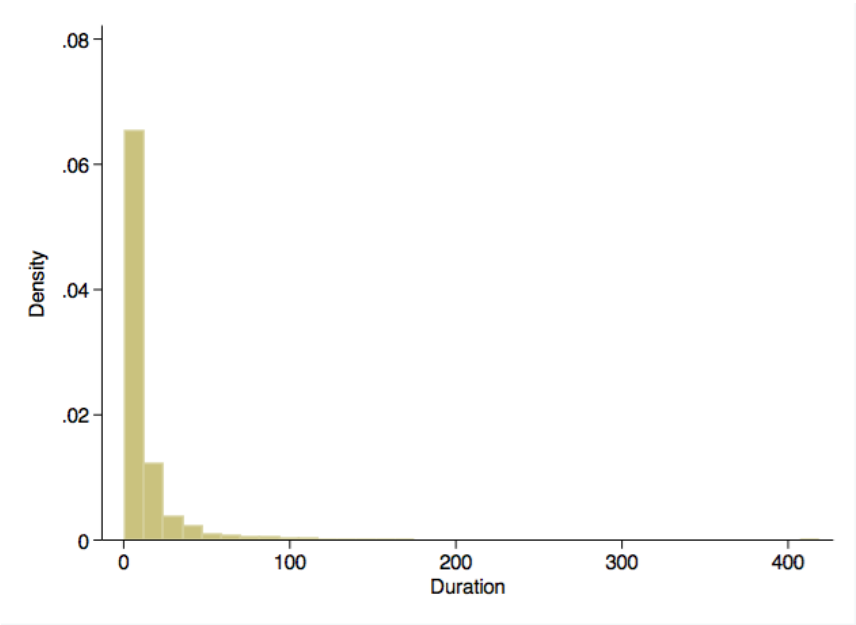
Aim: Visualize and analyze relationship between flood characteristics and number of people displaced

1. Visualize distribution of variables

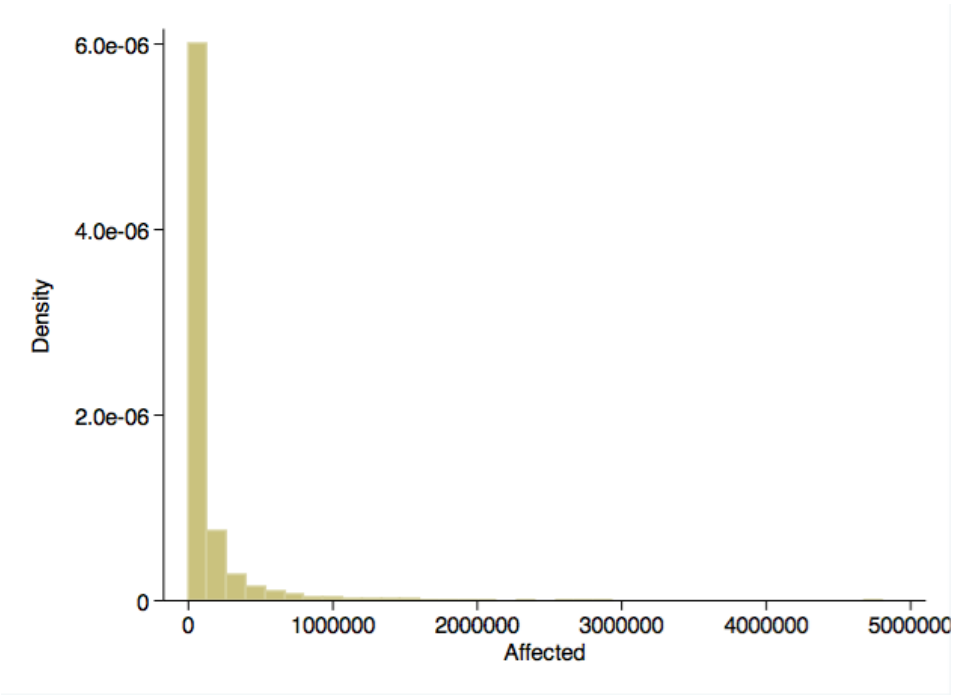
- Magnitude:



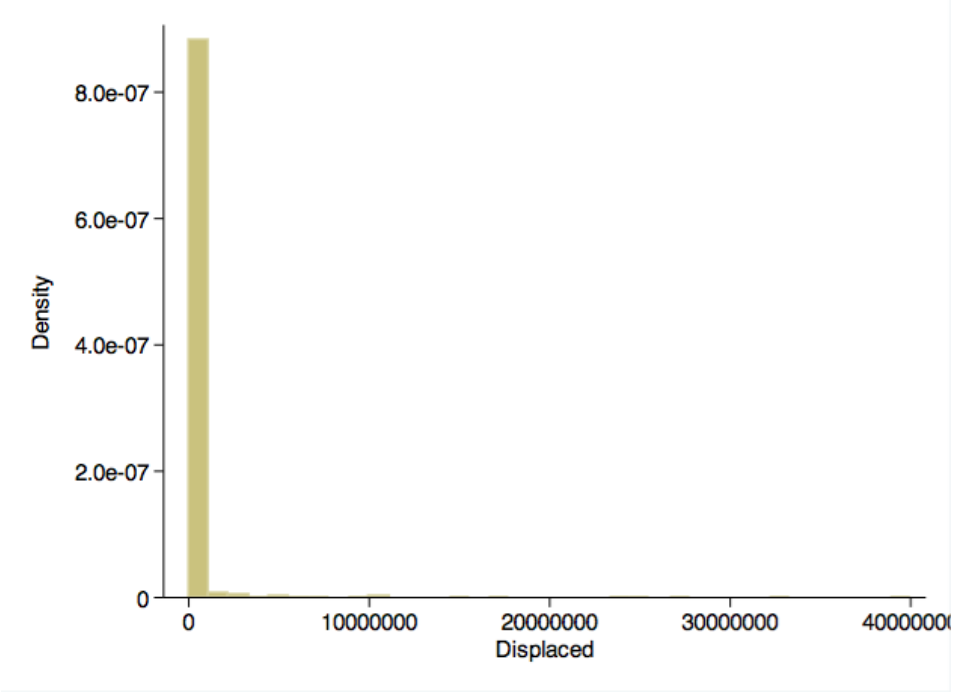
- Duration:



- Affected area:



- Number displaced:



2. Center and scale

Except for magnitude, other variables are not normally distributed, so log-transform them and then center and scale all variables.

3. Linear regression

Perform multiple linear regression with “displaced” as dependent variable and other three variables (“magnitude”, “duration”, “affected”) as independent variables.

This model seems to explain ~20% of number of people displaced (model R2 = 0.19).

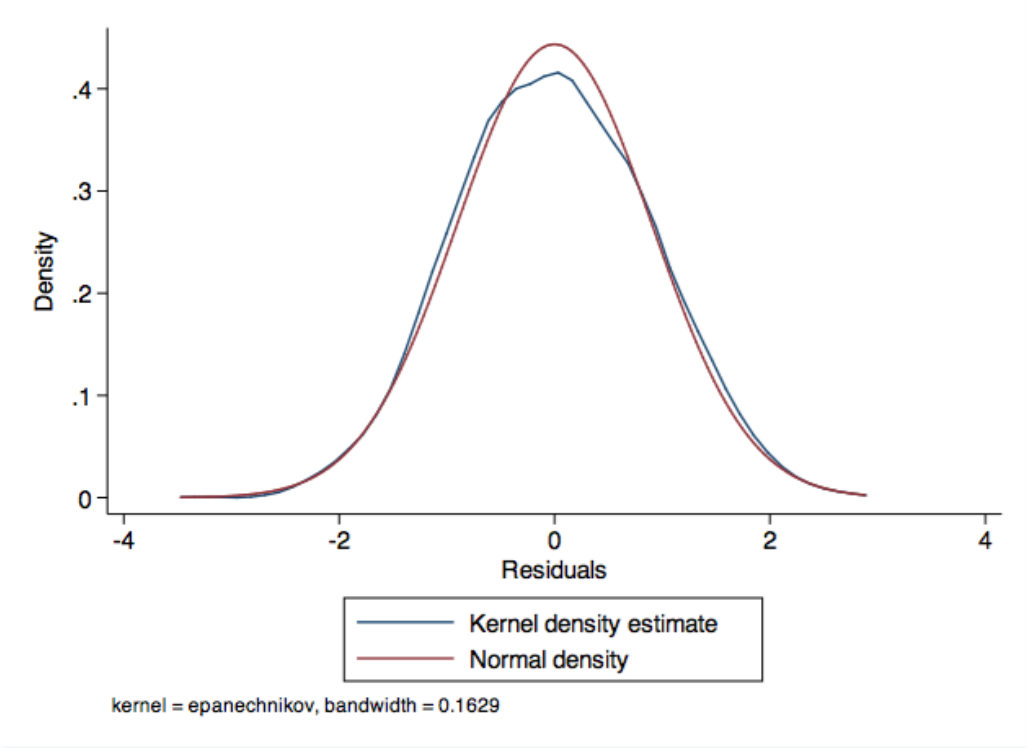
```
Model 1.  regress std_log_displaced std_magnitude std_log_duration std_log_affected

      Source |       SS       df       MS              Number of obs =    3034
-------------+------------------------------           F(  3,  3030) =  238.20
       Model |  578.799987     3  192.933329            Prob > F      =  0.0000
    Residual |  2454.19999  3030  .809966995           R-squared     =  0.1908
-------------+------------------------------           Adj R-squared =  0.1900
       Total |  3032.99998  3033  .999999994           Root MSE      =  .89998

------------------------------------------------------------------------------
std_log_displa~d |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
   std_magnitude |   .2641184   .1101812     2.40   0.017     .0480809    .4801559
std_log_duration |   .2125168   .0464031     4.58   0.000     .1215321    .3035015
std_log_affected |  -.0117362   .0844968    -0.14   0.890    -.1774131    .1539406
           _cons |  -.0675865   .0165535    -4.08   0.000    -.1000437   -.0351292
------------------------------------------------------------------------------
```

Remove “affected” since not significantly associated with dependent variable. R2 remains unchanged.

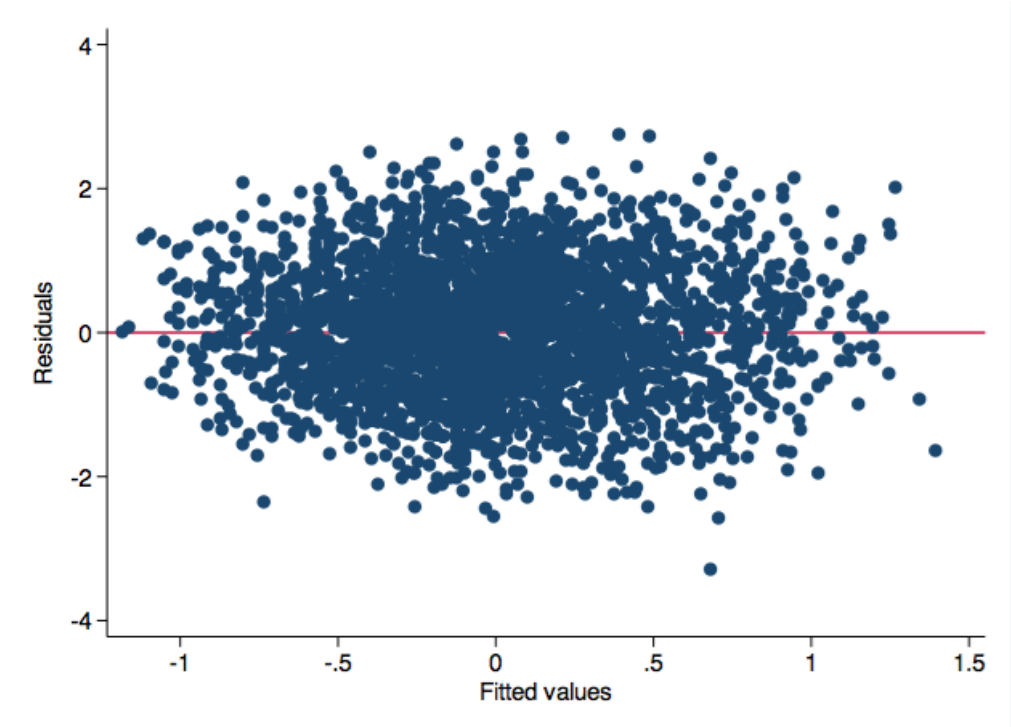
```
Model 2.  regress std_log_displaced std_magnitude std_log_duration

      Source |       SS       df       MS              Number of obs =    3034
-------------+------------------------------           F(  2,  3031) =  357.40
       Model |  578.784361     2   289.39218            Prob > F      =  0.0000
    Residual |  2454.21562  3031  .809704922           R-squared     =  0.1908
-------------+------------------------------           Adj R-squared =  0.1903
       Total |  3032.99998  3033  .999999994           Root MSE      =  .89984

------------------------------------------------------------------------------
std_log_displa~d |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
   std_magnitude |   .2491953   .0244173    10.21   0.000     .2013191    .2970716
std_log_duration |   .2180322   .0240046     9.08   0.000     .1709653    .2650992
           _cons |  -.0675127   .0165423    -4.08   0.000    -.0999479   -.0350774
------------------------------------------------------------------------------
```

Check assumptions of linear regression:

- Assumption of normality of residuals appears satisfied.



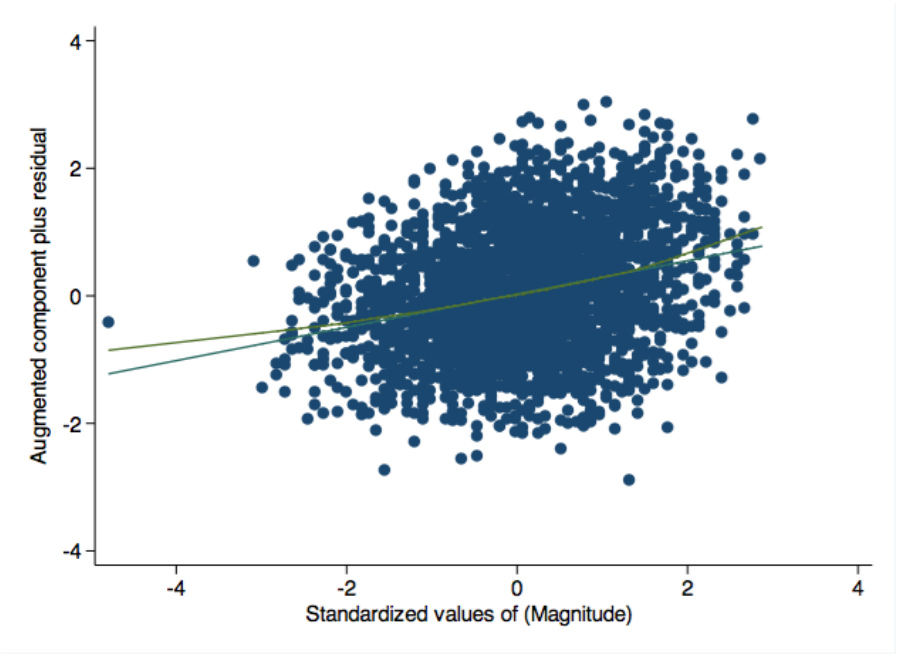
- Assumption of homoscedasticity of residuals appears satisfied.



- No evidence of significant collinearity (VIF <10).

Variable	VIF	1/VIF
std_log_du~n	2.21	0.452957
std_magnit~e	2.21	0.452957
Mean VIF	2.21	

- Assumption of linearity appears satisfied.



4. Perform principal components analysis on “magnitude” and “duration”

Estimate principal components.

```
pca std_magnitude std_log_duration

 Principal components/correlation                  Number of obs    =      4312
                                                   Number of comp.  =         2
                                                   Trace            =         2
     Rotation: (unrotated = principal)             Rho              =    1.0000

     --------------------------------------------------------------------------
        Component |   Eigenvalue   Difference         Proportion   Cumulative
     -------------+------------------------------------------------------------
           Comp1 |      1.73842      1.47684             0.8692       0.8692
           Comp2 |      .261578            .             0.1308       1.0000
     --------------------------------------------------------------------------

 Principal components (eigenvectors)

     ------------------------------------------------
         Variable |    Comp1     Comp2 | Unexplained
     -------------+--------------------+-------------
     std_magnit~e |   0.7071    0.7071 |           0
     std_log_du~n |   0.7071   -0.7071 |           0
     ------------------------------------------------
```

Component 1 explains ~87% of variance, so compute score of that component (pc1) and regress on that alone. R2 of Model 3 is similar to the earlier Model 2 with separate terms for “magnitude” and “duration”.

```
Model 3.  regress std_log_displaced pc1

       Source |       SS       df       MS              Number of obs =    3034
 -------------+------------------------------           F(  1,  3032) =  714.48
        Model |   578.41338     1   578.41338           Prob > F      =  0.0000
     Residual |   2454.5866  3032  .809560224           R-squared     =  0.1907
 -------------+------------------------------           Adj R-squared =  0.1904
        Total |  3032.99998  3033  .999999994           Root MSE      =  .89976

 ------------------------------------------------------------------------------
 std_log_di~d |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
 -------------+----------------------------------------------------------------
          pc1 |   .3297514   .0123365    26.73   0.000     .3055626    .3539401
        _cons |  -.0676347   .0165297    -4.09   0.000    -.1000453   -.0352241
 ------------------------------------------------------------------------------
```

Regression model with “pc1” (Model 3 above) appears to have superior R2 (0.19) than regression models with “duration” alone (0.16) (Model 4 below) or “magnitude” alone (0.17) (Model 5 below).

Model 4. regress std_log_displaced std_log_duration

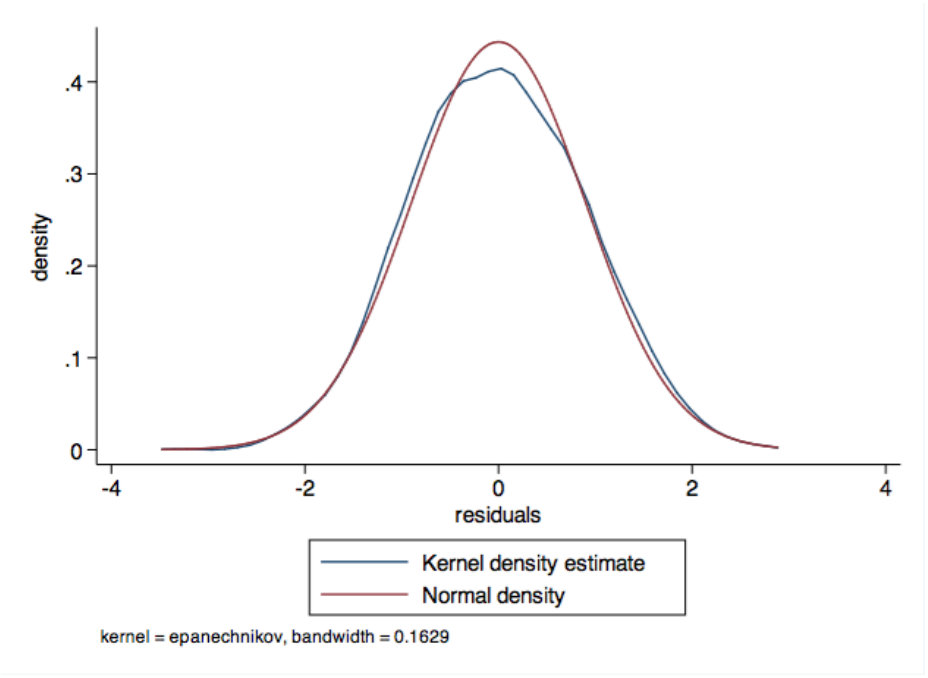
```
      Source |       SS       df       MS              Number of obs =    3034
-------------+------------------------------           F(  1,  3032) =  590.56
       Model |  494.448928     1  494.448928            Prob > F      =  0.0000
    Residual |  2538.55105  3032  .837252986           R-squared     =  0.1630
-------------+------------------------------           Adj R-squared =  0.1627
       Total |  3032.99998  3033  .999999994           Root MSE      =  .91502

------------------------------------------------------------------------------
std_log_displa~d |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
std_log_duration |   .3992275   .0164281    24.30   0.000     .3670161    .4314389
           _cons |   -.063597   .0168168    -3.78   0.000    -.0965705   -.0306235
------------------------------------------------------------------------------
```

Model 5. regress std_log_displaced std_magnitude

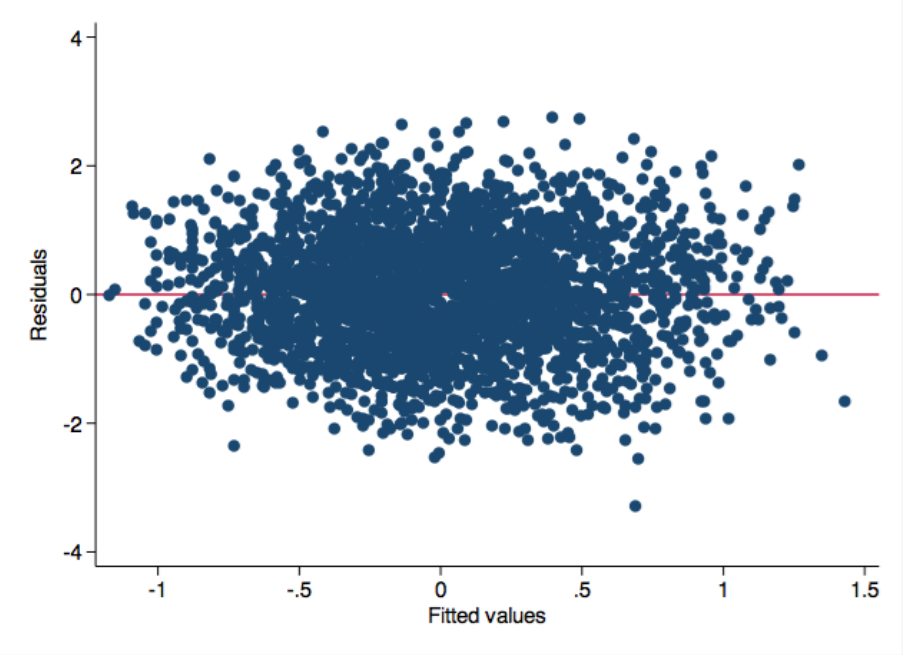
```
      Source |       SS       df       MS              Number of obs =    3034
-------------+------------------------------           F(  1,  3032) =  615.76
       Model |  511.983959     1  511.983959            Prob > F      =  0.0000
    Residual |  2521.01602  3032  .831469664           R-squared     =  0.1688
-------------+------------------------------           Adj R-squared =  0.1685
       Total |  3032.99998  3033  .999999994           Root MSE      =  .91185

------------------------------------------------------------------------------
std_log_dis~d |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
std_magnitude |   .4132298   .0166528    24.81   0.000     .3805779    .4458817
        _cons |  -.0543578   .0166988    -3.26   0.001    -.0870999   -.0216158
------------------------------------------------------------------------------
```

Check assumptions of linear regression for model with principal component alone (Model 3):

- Assumption of normality of residuals appears satisfied.



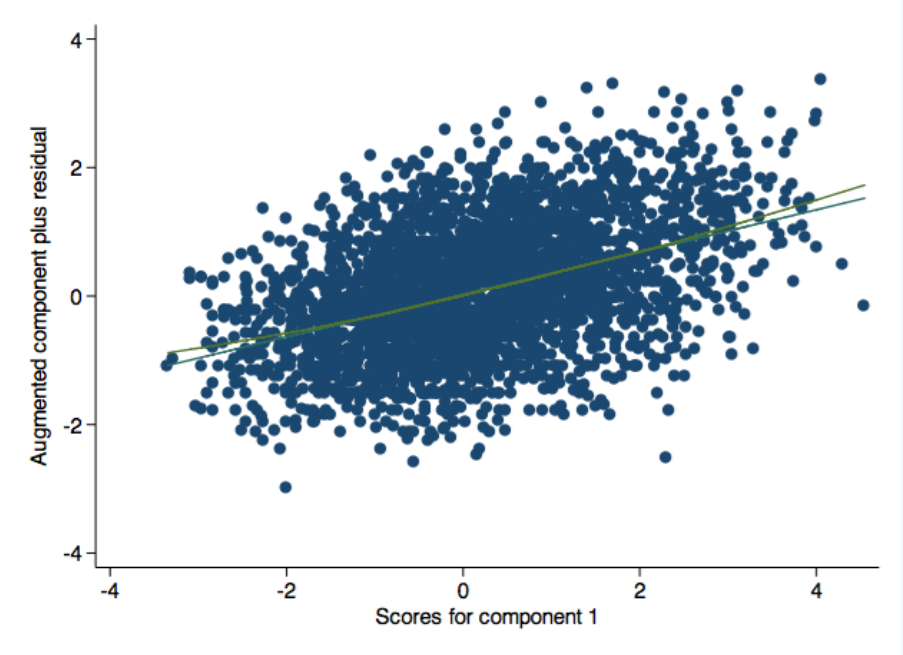
- Assumption of homoscedasticity of residuals appears satisfied.



- No evidence of significant collinearity (VIF <10).

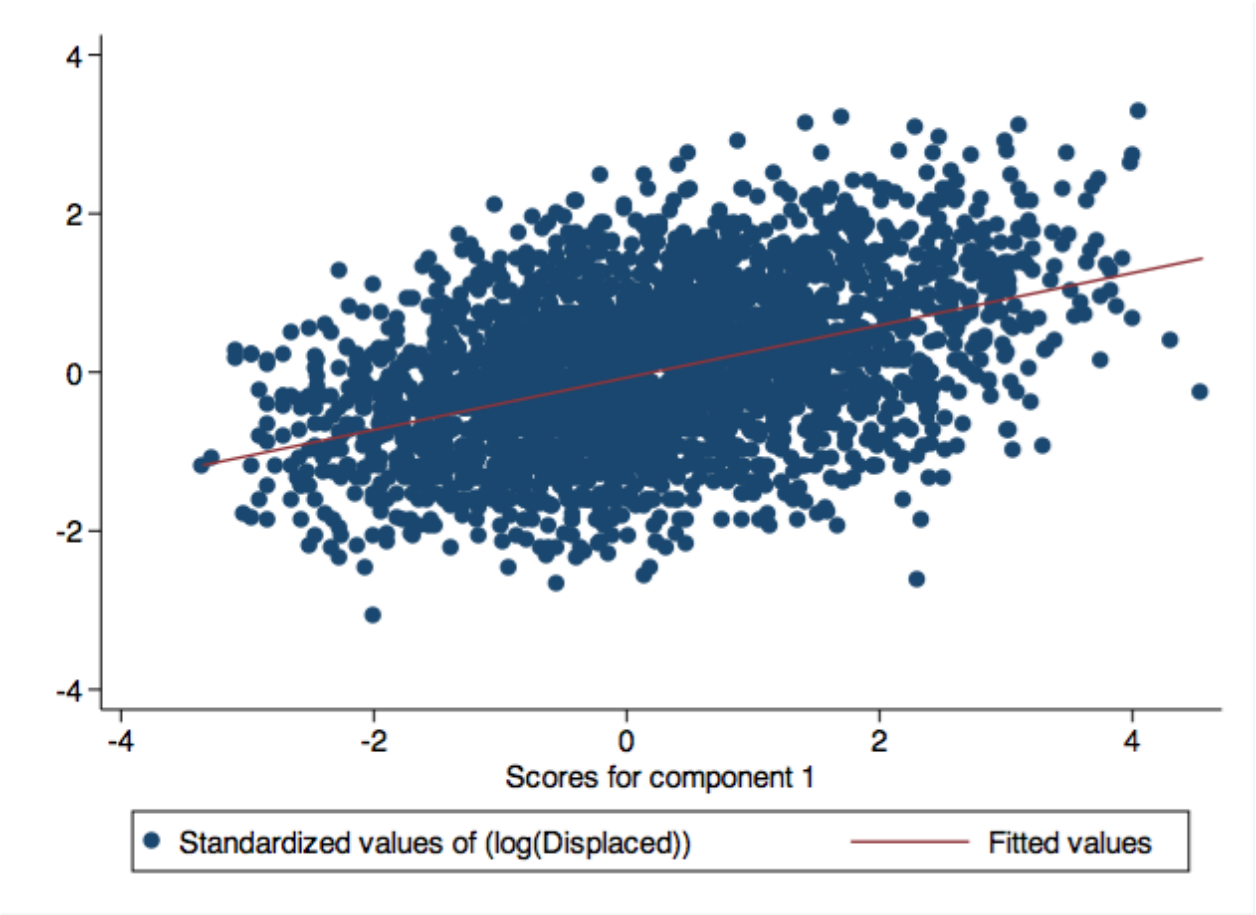
Variable	VIF	1/VIF
pc1	1.00	1.000000
Mean VIF	1.00	

- Assumption of linearity appears satisfied.



5. Visually assess relationship between multiple flood characteristics and the number of people displaced

Such a visualization would be difficult to perform with traditional linear regression in the presence of multiple independent variables.



6. Summmary

Principal components analysis allowed dimension reduction to one dimension, thereby allowing direct visualization between predictor and outcome.

Regression using one principal component was superior to regression limited to one traditional independent variable.