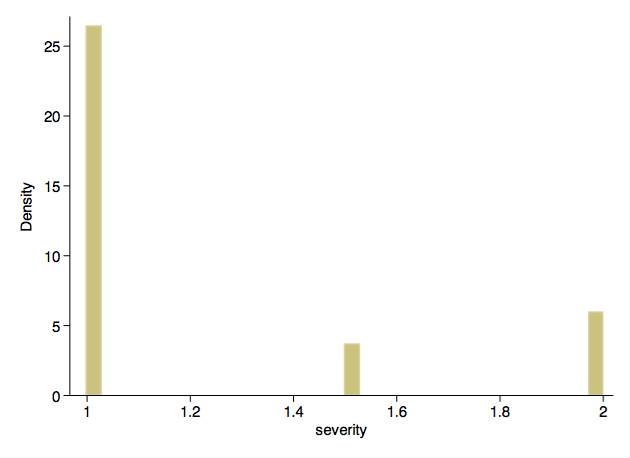
# EDAV Project 2: Regression with Principal Components

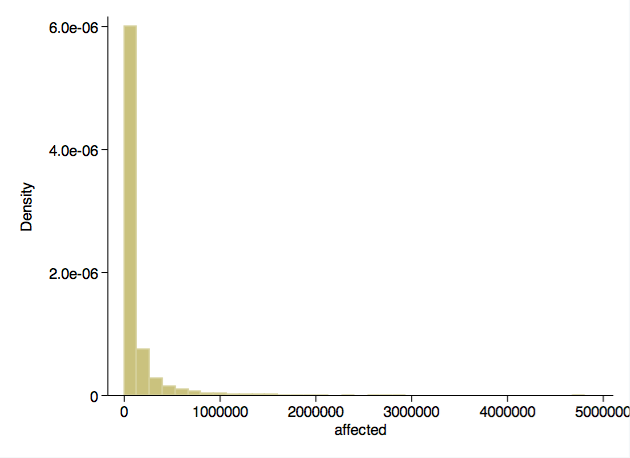
## Aim: Visualize relationship between flood characteristics and number of people displaced

### 1. Visualize distribution of variables

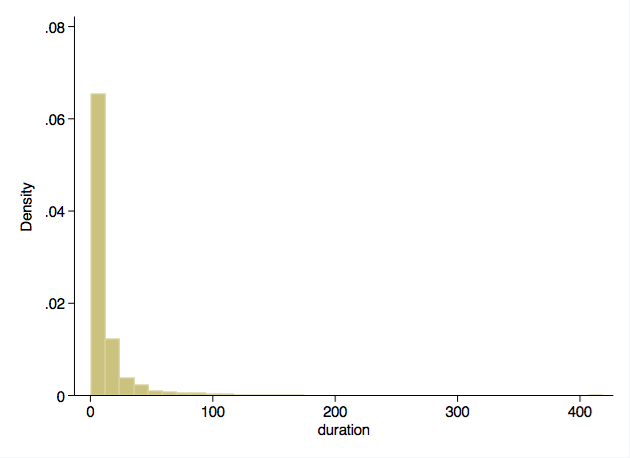
* Severity:



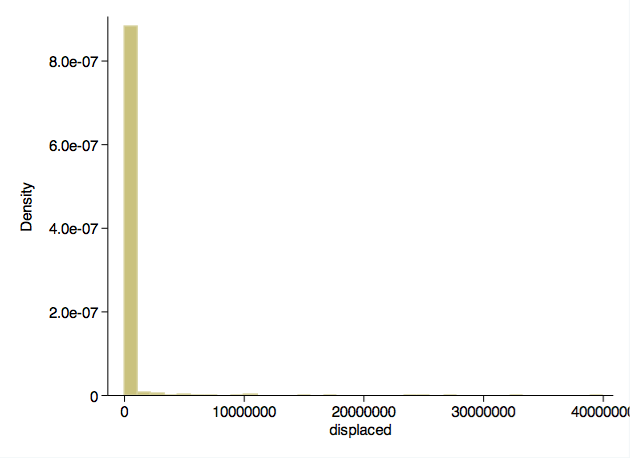
* Affected area:



* Duration:



* Number displaced:



### 2. Center and scale

Except for severity, other variables appear skewed, so log-transform them and then center and s cale them.

### 3. Linear regression

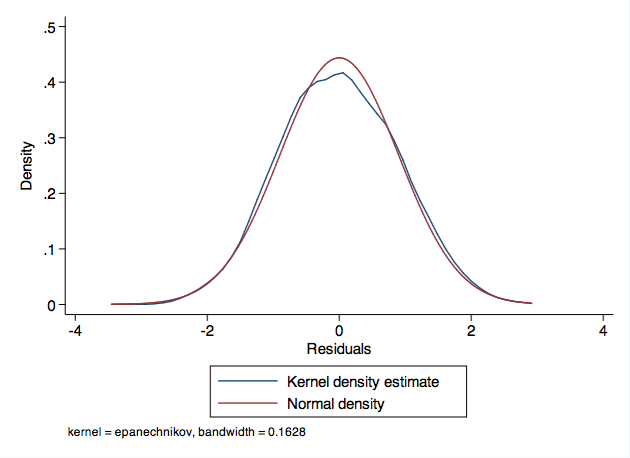
Perform multiple linear regression with “displaced” as dependent variable and other three varia bles as independent variables.

This model seems to explain ~19% of number of people displaced.

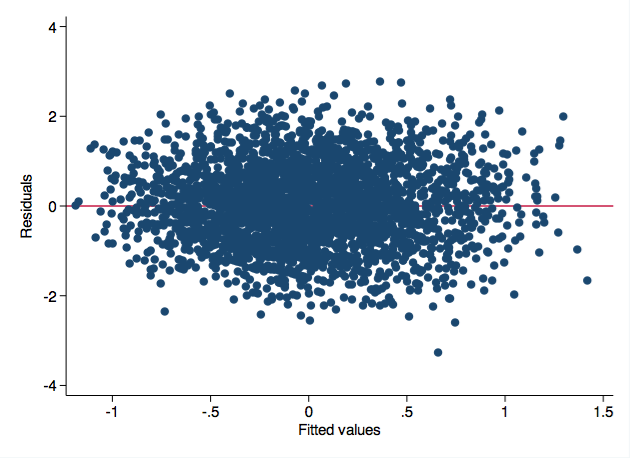
1 . regress std\_log\_displaced severity std\_log\_affected std\_log\_duration  
   
 Source | SS df MS Number of obs = 3034  
 -------------+------------------------------ F( 3, 3030) = 239.40  
 Model | 581.158345 3 193.719448 Prob > F = 0.0000  
 Residual | 2451.84164 3030 .809188658 R-squared = 0.1916  
 -------------+------------------------------ Adj R-squared = 0.1908  
 Total | 3032.99998 3033 .999999994 Root MSE = .89955  
   
 ----------------------------------------------------------------------------------  
 std\_log\_displa~d | Coef. Std. Err. t P>|t| [95% Conf. Interval]  
 -----------------+----------------------------------------------------------------  
 severity | .1240879 .0421517 2.94 0.003 .0414392 .2067367  
 std\_log\_affected | .1801437 .0188169 9.57 0.000 .1432485 .217039  
 std\_log\_duration | .3070196 .0184443 16.65 0.000 .2708549 .3431842  
 \_cons | -.2192853 .0545864 -4.02 0.000 -.3263154 -.1122551  
 ----------------------------------------------------------------------------------

Check assumptions of linear regression:

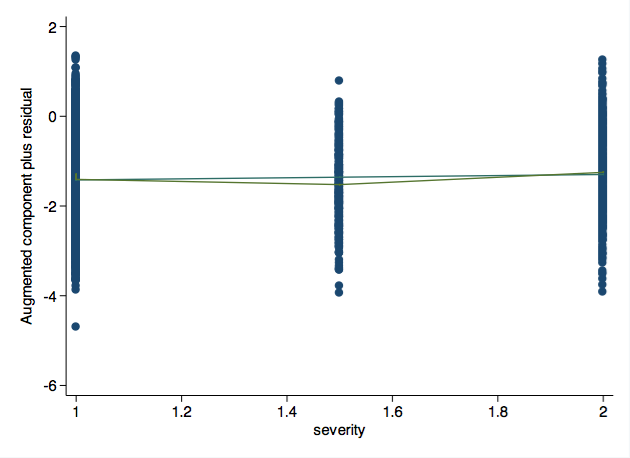
* Assumption of normality of residuals appears satisfied.

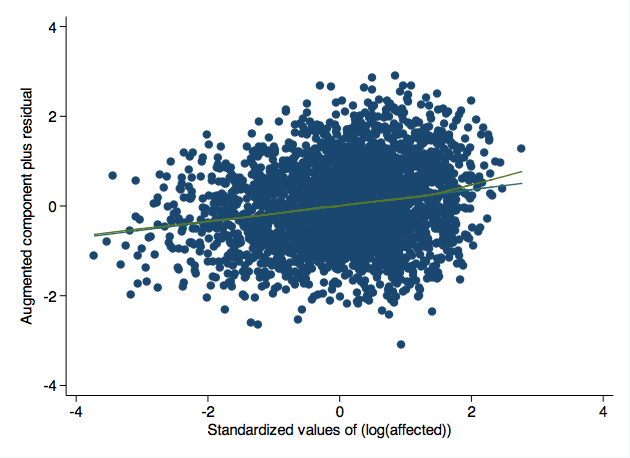


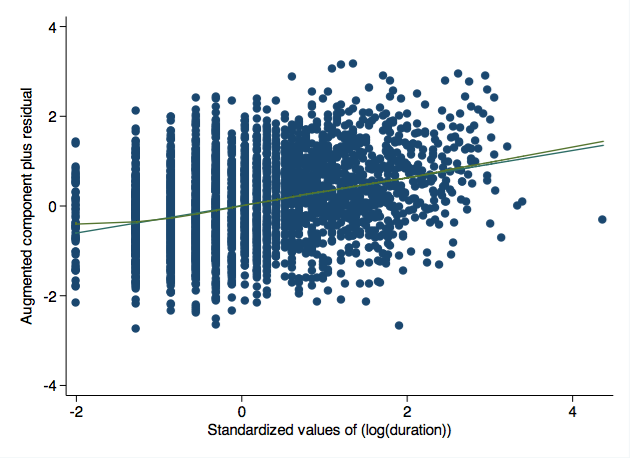
* Assumption of homoscedasticity of residuals appears satisfied.



* No evidence of significant collinearity (VIF <10).
* 2 . vif  
    
   Variable | VIF 1/VIF   
   -------------+----------------------  
   std\_log\_du~n | 1.30 0.766732  
   std\_log\_af~d | 1.29 0.774600  
   severity | 1.06 0.945632  
   -------------+----------------------  
   Mean VIF | 1.22
* Assumption of linearity appears satisfied.







### 4. Perform principal components analysis on “affected”, “severity”, and “duration”

Estimate principal components.

3 . pca severity std\_log\_affected std\_log\_duration  
   
 Principal components/correlation Number of obs = 4312  
 Number of comp. = 3  
 Trace = 3  
 Rotation: (unrotated = principal) Rho = 1.0000  
   
 --------------------------------------------------------------------------  
 Component | Eigenvalue Difference Proportion Cumulative  
 -------------+------------------------------------------------------------  
 Comp1 | 1.58649 .701889 0.5288 0.5288  
 Comp2 | .884599 .355685 0.2949 0.8237  
 Comp3 | .528914 . 0.1763 1.0000  
 --------------------------------------------------------------------------  
   
 Principal components (eigenvectors)   
   
 ----------------------------------------------------------  
 Variable | Comp1 Comp2 Comp3 | Unexplained   
 -------------+------------------------------+-------------  
 severity | 0.4067 0.9125 0.0443 | 0   
 std\_log\_af~d | 0.6412 -0.3196 0.6977 | 0   
 std\_log\_du~n | 0.6508 -0.2554 -0.7150 | 0   
 ----------------------------------------------------------

Component 1 explains ~53% of variance, so compute score of that component (pc1) and regress on that alone.

As seen, model R2 is similar to earlier model with separate terms for “magnitude” and “duration” (R2 ~0.18 versus ~0.19).

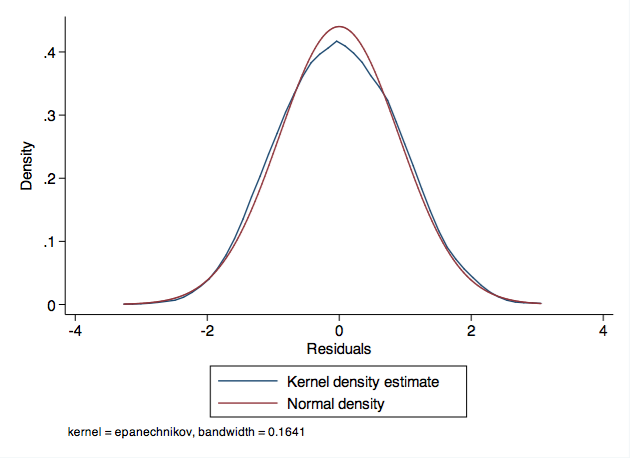
4 . regress std\_log\_displaced pc1  
   
 Source | SS df MS Number of obs = 3034  
 -------------+------------------------------ F( 1, 3032) = 660.36  
 Model | 542.436732 1 542.436732 Prob > F = 0.0000  
 Residual | 2490.56325 3032 .821425874 R-squared = 0.1788  
 -------------+------------------------------ Adj R-squared = 0.1786  
 Total | 3032.99998 3033 .999999994 Root MSE = .90633  
   
 ------------------------------------------------------------------------------  
 std\_log\_di~d | Coef. Std. Err. t P>|t| [95% Conf. Interval]  
 -------------+----------------------------------------------------------------  
 pc1 | .3320503 .0129215 25.70 0.000 .3067145 .3573861  
 \_cons | -.0640488 .0166419 -3.85 0.000 -.0966794 -.0314183  
 ------------------------------------------------------------------------------

Regression model with “pc1” appears to be equal to or superior R2 to regression models with “du ration” alone or “magnitude” alone.

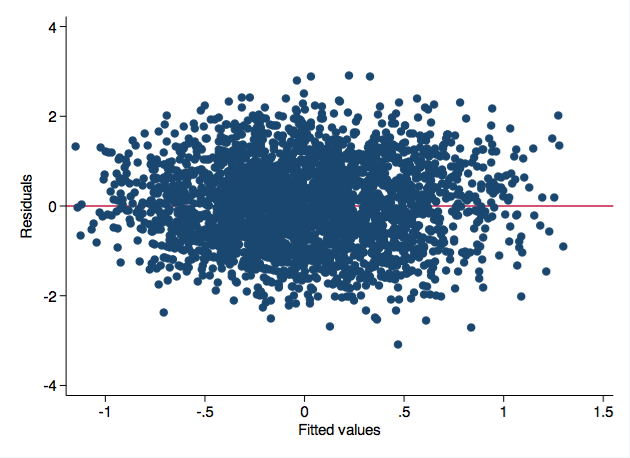
5 . regress std\_log\_displaced severity  
   
 Source | SS df MS Number of obs = 3034  
 -------------+------------------------------ F( 1, 3032) = 67.93  
 Model | 66.4655162 1 66.4655162 Prob > F = 0.0000  
 Residual | 2966.53446 3032 .978408464 R-squared = 0.0219  
 -------------+------------------------------ Adj R-squared = 0.0216  
 Total | 3032.99998 3033 .999999994 Root MSE = .98915  
   
 ------------------------------------------------------------------------------  
 std\_log\_di~d | Coef. Std. Err. t P>|t| [95% Conf. Interval]  
 -------------+----------------------------------------------------------------  
 severity | .3714916 .0450724 8.24 0.000 .283116 .4598672  
 \_cons | -.4636299 .0590483 -7.85 0.000 -.5794086 -.3478511  
 ------------------------------------------------------------------------------  
   
   
 6 . regress std\_log\_displaced std\_log\_affected  
   
 Source | SS df MS Number of obs = 3034  
 -------------+------------------------------ F( 1, 3032) = 374.66  
 Model | 333.563804 1 333.563804 Prob > F = 0.0000  
 Residual | 2699.43618 3032 .890315361 R-squared = 0.1100  
 -------------+------------------------------ Adj R-squared = 0.1097  
 Total | 3032.99998 3033 .999999994 Root MSE = .94357  
   
 ----------------------------------------------------------------------------------  
 std\_log\_displa~d | Coef. Std. Err. t P>|t| [95% Conf. Interval]  
 -----------------+----------------------------------------------------------------  
 std\_log\_affected | .3362418 .0173714 19.36 0.000 .3021809 .3703027  
 \_cons | -.0289544 .0171955 -1.68 0.092 -.0626704 .0047615  
 ----------------------------------------------------------------------------------  
   
   
 7 . 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  
 ----------------------------------------------------------------------------------

Check assumptions of linear regression for model with principal component alone:

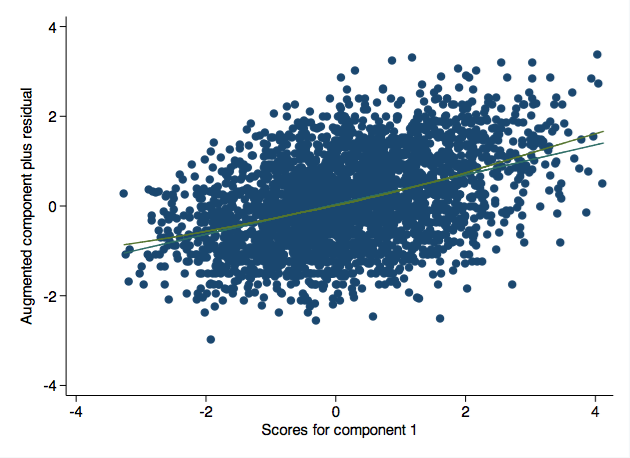
* Assumption of normality of residuals appears satisfied.



* Assumption of homoscedasticity of residuals appears satisfied.

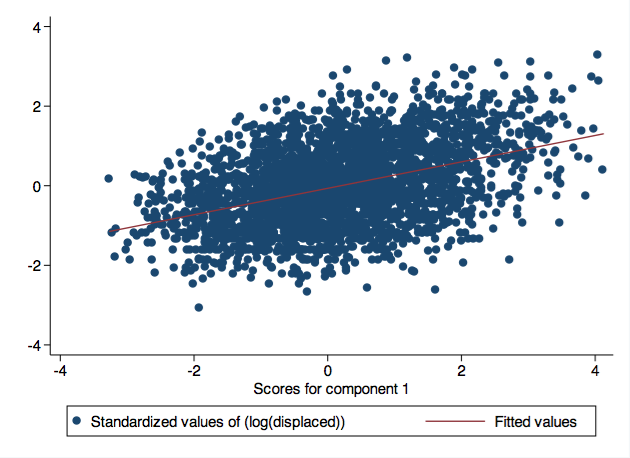


* No evidence of significant collinearity (VIF <10).
* 8 . vif  
    
   Variable | VIF 1/VIF   
   -------------+----------------------  
   pc1 | 1.00 1.000000  
   -------------+----------------------  
   Mean VIF | 1.00
* Assumption of linearity appears satisfied.



### 5. Visually assess relationship between multiple flood characteristics and # displaced

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



### 6. Summmary

Principal components analysis allowed dimension reduction to one dimension, thereby allowing di rect visualization.

Prediction using one principal component was equally predictive as regression limited to one tr aditional independent variable and additionally allowed direct visualization between predictor and outcome. However, the disad vantage of this approach is that the first principal component is more difficult to conceptually understand than a traditional p redictor such as severity.