Paper 271-2010

Visual Displays of Regression toward the Mean using SAS® SGplot

Ginger Lockhart Burrell, Ph.D.[1], Felix Thoemmes, Ph.D.[2], David P. MacKinnon, Ph.D.[3]
[1] Johns Hopkins Bloomberg School of Public Health
[2] Texas A & M University
[3] Arizona State University

Abstract

Regression toward the mean is a widespread phenomenon in statistics and may adversely affect researchers' substantive interpretation of findings. Therefore, it is worthwhile to evaluate the extent of regression toward the mean as a part of many applications of regression analysis. One way to assist with such an evaluation is with a *Galton squeeze diagram* (Campbell & Kenny, 1999), which provides a quick visual summary of regression toward the mean. This paper demonstrates a novel application of SAS® SGPLOT to create publication-ready Galton squeeze plots.

Introduction

More than 120 years ago, Sir Francis Galton described the phenomenon of regression to mediocrity, or, as it is known now, regression to the mean. In his landmark research on the relation between parents' and offsprings' heights, he found that very tall or very short parents tended to have children who were less extreme in height (Galton, 1886). Though he initially believed that this phenomenon was the result of a force of inheritance, it later became clear that what he observed was a characteristic of imperfectly correlated variables. That is, when two variables are imperfectly correlated, extreme scores on one variable move or *regress* toward the mean on the second variable (Campbell & Kenny, 1999). Because perfect correlation rarely occurs, regression toward the mean is an ever-present challenge in statistics. In nonexperimental studies, this phenomenon can have serious implications for substantive conclusions. For example, public policy researchers may be led to conclude that an intervention designed to reduce neighborhood crime was effective for the most crime-ridden neighborhoods, when in fact, the findings were the result of regression to the mean. The only way to avoid regression to the mean entirely is with a completely randomized design, but many areas of study, such as medicine, psychology, or public policy, entail costs and ethical considerations which prevent such a design. Thus, it is worthwhile to include an evaluation of regression to the mean in many applications of regression analysis.

One useful visual tool for assessing the extent and nature of regression to the mean is the *Galton squeeze diagram.* This diagram links the scores on one variable with the 'squeezed' or average of the second variable for each score on the first variable between two vertical axes. More formally, the axis on the left hand side of the Galton squeeze diagram plots individual scores on a variable measured at a previous time point, whereas the right hand side of the Galton squeeze diagram plots means at a later point in time for all possible pretest scores that existed in the dataset. Each unit in the dataset corresponds to one point at each of the two axes. Corresponding units at both points in time are connected with lines in the graph.

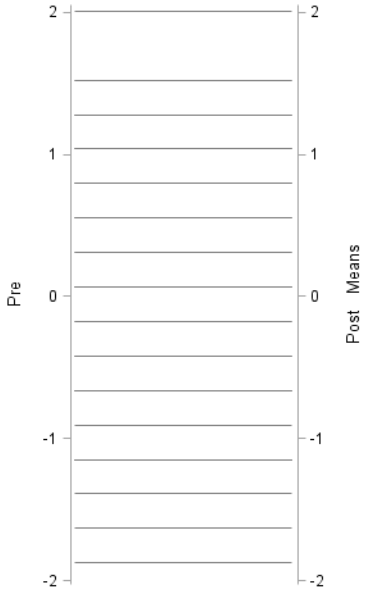
The formulaic expression of the content of the plot is:

Left Axis: x_{1i}, where x_1 is a variable at time 1, and i a subscript denoting units i to N.

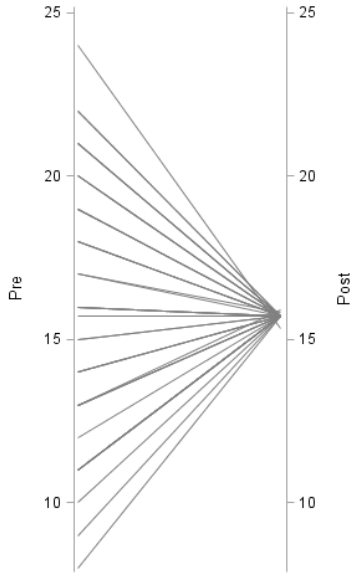
Right axis: $(\bar{x}_2 \mid x_{1i} = x)$ where x_2 denotes a mean on the second variable conditional on a certain value x on variable x_{1i}

For the purposes of this demonstration, we refer to the first variable x_{1i} as the 'pretest' and the second variable x_2 as the 'posttest'.

When the correlation between the two variables is perfect, these conditional posttest means are equal to the scores on the pretest, resulting in parallel lines in the squeeze plot and no regression toward the mean:



Conversely, a correlation of zero would indicate that the best prediction of the posttest scores is the mean of the posttest. Thus, a squeeze diagram would show a convergence of conditional means to the overall sample mean:



As this plot shows, post test conditional means become more focused toward the overall sample mean as the correlation coefficient approaches zero.

This paper makes extensive use of the SAS SGPLOT command to create high-quality graphics of pair-link and Galton squeeze plots. To the best of our knowledge, no other program produces this type of graph in a publication-ready format. The example uses hypothetical data from David Kenny's website (2009; available at http://davidakenny.net/series/rtmprog.htm). The first set of SAS code creates a *pair-link* diagram (Campbell & Kenny, 1999), which shows the scores across two time points for each case without squeezing the post test scores. This diagram is similar to the 'spaghetti' plot used in growth models and simply connects the data points across the two time points for each individual datapoint. The second set of SAS code produces a *forward squeeze diagram* or simply *Galton squeeze diagram,* in which the individual scores of the pretest are connected to the conditional means of the post test. This diagram is described as 'forward' because it shows the extent of regression toward the mean as time moves forward. Galton (1886) also found that regression toward the mean occurs backwards in time, such that the conditional means of the *pretest* converge toward the mean. The SAS code for a *reverse squeeze diagram* or *Reverse Galton Squeeze diagram* is presented last.

Program

1. Set-up. The following SAS code prepares the dataset to construct the three diagrams. Cases are first sorted according to their pretest scores. For the two squeeze diagrams, the pretest and posttest variables are both standardized in order to delete the visual effect of change in means over time. Note that the variables for the pair–link diagram are not standardized.

```
data a;
infile 'C:\Documents and Settings\ginnir\My
Documents\Publications\Squeeze\kenny.dat';
input pre post;
run;

/*include ID variable*/;
data a;
set a;
id = _N_;
run;

PROC STANDARD DATA=a MEAN=0 STD=1 OUT=stand;
  VAR pre post;
RUN;

data stand;
set stand;
rename pre=stand_pre post=stand_post;
run;

data a;
merge a stand;
run;

proc sort data=a;
by stand_pre;
run;
```

2. Create conditional means and restructuring the data file. The following code uses PROC REG to estimate conditional means for the posttest scores by saving the predicted values from regression equations that have no predictor and are run separately for each value of the pretest variable using the BY statement.

```
proc reg data=a noprint;
model stand_post= ;
by stand_pre;
output out=apred predicted=postmeans;
run;
proc sort data=apred;
by stand_post;
proc reg data=apred noprint;
model stand_pre= ;
by stand_post;
output out=bpred predicted=premeans;
run;
```

3. Transpose the data file. The code below transposes the data file to the long form. This step is necessary to express the temporal order of the variables.

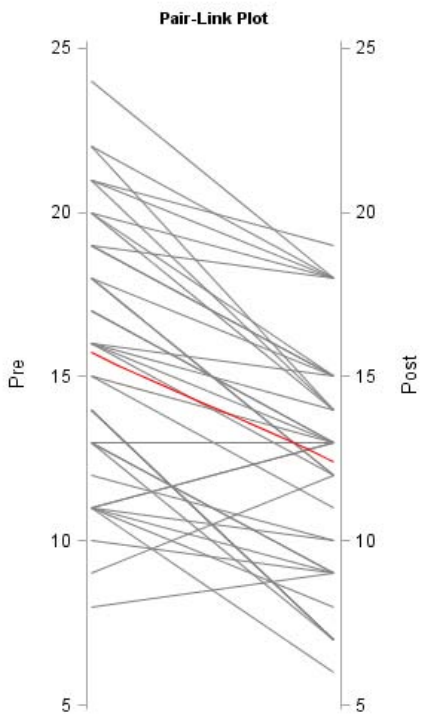
```
data blong;
set bpred;
time=1; pair=pre; gs=stand_pre; gsr=premeans ; output;
time=2; pair=post; gs=postmeans; gsr=stand_post; output;
keep id time pair gs gsr;
run;
```

4. Form plot template. This code sets up the template used for each of the plots and suppresses the default border.

```
proc template;
 define style nowall;
 parent=styles.statistical;
 style graphwalls from graphwalls /
   frameborder=off;
 end;
run;
```

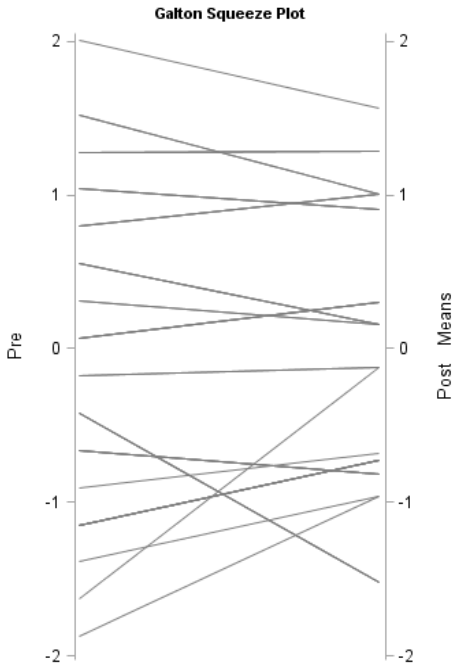
5. Pair-link diagram. These statements use PROC SGPLOT to produce the pair-link diagram. Here, the X axis is suppressed and two Y axes are constructed. Because the links are plotted in both directions, the second set of lines is suppressed by setting THICKNESS=0. Also included is a red line showing the change in the mean,

```
title h=.7 "Pair-Link Plot";
proc sgplot data=blong noautolegend;
xaxis display=none;
yaxis label='Pre';
y2axis label='Post';
series x=time y=pair /group=ID
lineattrs = (COLOR=GRAY PATTERN=1 THICKNESS=1);
series x=time y=pair /y2axis
lineattrs = (THICKNESS=0);
reg x=time y=pair / nomarkers
lineattrs = (COLOR=RED PATTERN=1 THICKNESS=1);
run;
```



6. Galton squeeze diagram. This code produces a forward Galton squeeze diagram. The commands are similar to those used for the pair-link plot, but standardized variables are used and the Y axis represents the means on the post test for each score of Y.

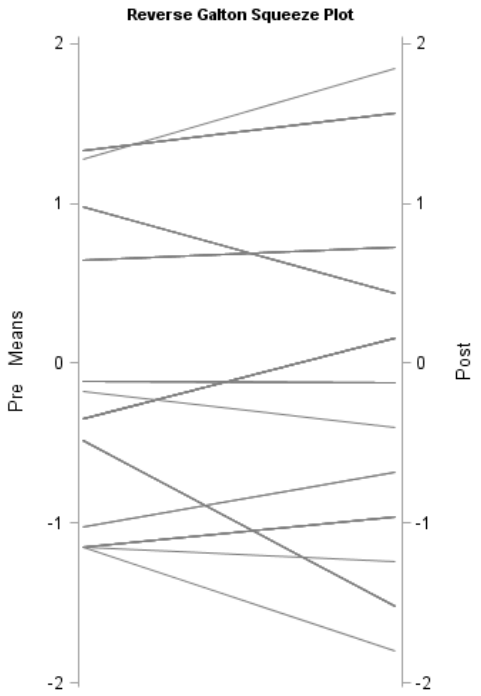
```
title h=.7 "Galton Squeeze Plot";
proc sgplot data=blong noautolegend;
xaxis display=none;
yaxis label='Pre';
y2axis label='Post   Means';
series x=time y=gs /group=ID
lineattrs = (COLOR=GRAY PATTERN=1 THICKNESS=1);
series x=time y=gs /y2axis
lineattrs = (THICKNESS=0);
run;
```



Note the convergence of lines toward the mean (0) of the post test. Also note that the number of lines decreased from the pair-link plot. Identical pre-test values all converge to the same post-test mean and are represented with a single line. The number of lines is identical to the number of unique values at pre-test.

7. Reverse Galton squeeze diagram. This code produces the reverse Galton squeeze diagram, which shows regression toward the mean moving backward in time. Here, the left axis represents the means of the pretest conditional on the scores of the post test.

```
 title h=.7 "Reverse Galton Squeeze Plot";
proc sgplot data=blong noautolegend;
xaxis display=none;
yaxis label='Pre   Means';
y2axis label='Post';
series x=time y=gsr /group=ID
lineattrs = (COLOR=GRAY PATTERN=1 THICKNESS=1);
series x=time y=gsr /y2axis
lineattrs = (THICKNESS=0);
run;
```



Note the convergence toward the mean of the pretest.

Conclusion

Regression toward the mean occurs whenever variables are not perfectly correlated and can lead to incorrect conclusions, especially in nonrandomized studies. Galton squeeze plots generated by SAS Sgplot can be useful tools for evaluating the extent of regression toward the mean because they produce clear, which can help users make interpretations about their data.

References

Campbell, D.T., & Kenny, D.A. (1999). *A primer on regression artifacts.* New York: The Guilford Press.

Galton, F. (1886). Regression toward mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland, 15,* 246-263.

Contact Information

Your comments and questions are valued and encouraged. Please contact the authors at:

Ginger Lockhart Burrell
gburrell@jhsph.edu

Felix Thoemmes
felix.thoemmes@tamu.edu

David P. MacKinnon
david.mackinnon@asu.edu