**Automating the Selection of C (middle observations to be eliminated) for the Goldfeld-Quandt Test in SAS[[1]](#footnote-1)**

**A Goldfeld-Quandt test can be done in SAS as follows**

a) Order the data according to the magnitude of the proportionality variable, Z. You must determine which variable, Z, is likely to cause the heteroscedasticity, and you must determine which observations of Z will result in small error variances or large error variances. If we assume the error variance increases with Z, then it is just necessary to order the data from small to large values of Z. The SAS command to sort a data set according to the values of a variable is:

PROC SORT;  
BY Z;  
RUN;

This reorders the entire data set, for all variables, with small values of Z first.  
b) Omit the middle d observations, where d is one third of the sample size, n. Do this by creating two new data sets, each with part of the sample. Let the first subsample of small values be observations 1 through n1 and the second subsample of large values be observations n2 through n, corresponding to the reordered variables.

DATA B;  
SET A;  
OBSNUM = \_N\_;  
IF OBSNUM<=n1;  
RUN;  
DATA C;  
SET A;  
OBSNUM = \_N\_;  
IF OBSNUM>=n2;  
RUN;

c) Fit two separate regressions, one on each subsample. Use the PROC statement to identify which data set is being used. If the DATA= option is omitted, SAS uses the most recent data set.

PROC REG DATA=B;  
MODEL Y=X1 X2;  
RUN;  
PROC REG DATA=C;  
MODEL Y=X1 X2;   
RUN;

These regressions must be exactly the same form as the original regression on the full sample; same variables involved.

d) Note the residual sums of squares (SSR) for each regression. It is printed in the Analysis of Variance table, in the row labeled "ERROR" and column labeled "SUM OF SQUARES".

e) The F-statistic is then the ratio of the residual sum of squares for the subsample of large values, SSR3, to the residual sum of squares for the subsample of small values, SSR1.

F = SSR3 / SSR1

with K and ("n"-K-1) degrees of freedom in the numerator and denominator. If this exceeds the critical value of F, the null hypothesis of constant error variances would be rejected.

Goldfeld-Quandt Test for Heteroscedasticity - An Illustration  
The test is applied to an estimate where petroleum consumption (PCON) is a function of the number of motor vehicles registered (REG) and the gasoline tax rate (TAX). Residuals have the pattern of increasing with POPulation, and the proportionality variable that is the basis of this test is population, Z = POP.  
First the data set is sorted according to POP. Then the data set is divided into two subsamples, with the smallest 17 observations in data set B and the largest 18 observations in data set C. The middle 15 observations have been eliminated. Use a variable the counts the observations, here named OBSNUM. For the Goldfeld-Quandt procedure, create this variable only after the SORT has been done.  
DATA B;  
SET A;  
IF OBSNUM1<=17;  
RUN;  
DATA C;  
SET A;  
IF OBSNUM2>=33;  
RUN;

Run the regression on the smallest observations and note the residual sum of squares.  
PROC REG DATA=B;  
MODEL PCON=REG TAX;  
RUN;

Calculate the F-ratio, which is defined for this test as the ratio of the residual sum of the largest observations to that of the smallest, RSS3 to RSS1.

F = SSR3 / SSR1 = 2108821 / 28956 = 72.83.

Since the critical value of F at the 5% level of significance with v1=2 and v2=17 is 3.59, we reject the null hypothesis of equal variance and conclude that there is heteroscedasticity.

1. Source: http://economics.kenyon.edu/Sas%20Help.htm [↑](#footnote-ref-1)