NOTA TECNICA 2
AN EQUAL VARIANCE TEST

GEORGE G. DJOLOV

Abstract

This article introduces (and hopes to encourage thereby) the econometrics practitioner to (use) a homoscedasticity *test referred to in the field of statistics as the modified Levene test. Econometrics orthodoxy (from University to practice level) has focused mainly on three heteroscedasticity tests, namely the Goldfeld-Quandt (GQ), Breusch-Pagan-Godfrey (BPG), and the White (W) test. The difference between the aforementioned tests and the test elaborated on in this article is that the former are for regression whereas the latter is for ANOVA –analysis of variance– situations.*

Resumen

Este artículo introduce (y espera animar por eso) al practicante de econometría a usar una prueba de homoscedasticidad, que se trata en el campo de estadísticas como una prueba de Levene modificada. La ortodoxia de econometría (en la educación Universitaria) se ha enfocado principalmente en tres pruebas de heterocedasticidad, los test de Goldfeld-Quandt (GQ), Breusch-pagano-Godfrey (BPG), y el test de White (W). La diferencia entre las pruebas mencionadas y la prueba elaborada en este artículo es que lo son para una regresión y el último considera situaciones ANOVA –análisis de varianza–.

JEL Classification: *C10, C12.*

Key Words: *Levene Test, modified Levene Test.*

□ Postal address: Faculty of Commerce, Department of Business Economics (Room 230), New Commerce Building, University of the Witwatersrand, Private bag 3, WITS 2050, South Africa. E-mail: gdjolov@sapma.co.za
This article benefited from the review of Prof. J Galpin of the University of the Witwatersrand. Any flaws are the author's own.

INTRODUCTION

This article elaborates on an *equal-variance* test referred to in the field of statistics as the modified Levene – ML – test. The test is derived from its predecessor, the Levene – L – test. The ML test is an alternative to the popular Bartlett test[1] for testing the equality or homogeneity of variances but performs better than it for the cases involving departures from the assumption of normality to which the Bartlett test is very sensitive.

The GQ, BPG, and W tests are well familiar, and not the purpose of discussion in this paper other than to outline the way they are computed later on in the paper. However it should be noted that they are deployed for regression situations with the reliance on the normality assumption declining as one moves from the GQ to the W test. The property of declined reliance on the normality assumption is also shared by the modified Levene test.

The Levene test or for that matter its modified variant are used for ANOVA situations. Since in regression one tends to rely on ANOVA to ascertain the significance of a model, by extension ANOVA can also be used to provide results for the ML test. Alternatively the same can be achieved with a two-sample t-test which is a special case of ANOVA with two independent samples treated as two levels of the factor of interest.

THE LEVENE TEST

The Levene test (Levene, 1960) tests if the assumption of homogeneity of variances (HOV) is valid. In practice this surmounts to testing if k samples have equal variances.

The test which requires the split of the original data in k sub-groups involves the construction of Z variables reflecting the absolute deviations of the data for each group from its respective mean and then doing a t-test for the mean differences of the deviations so derived.

As the HOV assumption in regression is with respect to the error term the Levene test (or its modification) should be carried out with reference to the residuals from a fitted model.

The Levene test then tests:

Ho: $\sigma_1 = \sigma_2 = ... = \sigma_K$ (i.e. the spread of the error-terms across the k sub-groups is identical), vs.

Ha: $\sigma_i \neq \sigma_j$ (i.e. for at least one group pair, i and j, the spread is not identical).

The Levene test involves breaking up the residuals of a model in 2 (or more) sub-groups and for each sub-group constructing Z-variables that reflect the absolute deviations from that group's mean. Thus:

---

[1] For a discussion on the Bartlett test refer to Snedecor and Cochran (1989); see reference list.

[1] $$Z_{ij} = \left| A_{ij} - \overline{A_i} \right|$$

where:

A_{ij} refers to the values of a sub-group, and

$\overline{A_i}$ refers to the mean for the ith sub-group.

Following the construction of the above variables a t-test for their mean differences is performed. If the Levene test is statistically significant than the hypothesis of homogenous variances should be rejected. Alternatively an F-test can be carried out because the L statistic follows the F-distribution. In particular for a sample of size N divided into k sub-groups with N_i being the sample size of the ith sub-group, the L-statistic is given by:

(2) $$L = (N-k)/(k-1) * \left[\sum_{i=1}^{k} N_i (\dot{Z}_{iZ} - \dot{Z}_G)^2 \Big/ \sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - \dot{Z}_{iZ})^2 \right]$$

where:

Z_{ij} = the Z variable values for a sub-group.

$\dot{Z}_{iZ}$ = group means of the Z variable defined for each sub-group, and

$\dot{Z}_G$ = grand total mean (i.e. the overall mean of the Z variable).

The L test rejects the hypothesis that the variances are homogenous if the L statistic exceeds $F(\alpha, k-1, N-k)$ which is the upper critical value of the F-distribution with k–1 and N–1 degrees of freedom at a significance level of α.

Because the scores of the absolute deviations from the group means can be expected to be highly skewed the normality assumption of the t or F-test (for the mean differences of the deviation scores) is usually violated. To deal with this, Brown and Forsythe (1974), have modified the Levene test; hence the modified Levene test.

The Modified Levene Test

Here instead of performing the t-test for mean differences on the absolute deviations from the mean, the analysis is performed on the absolute deviations from the group medians. Hence in expression [1] $\overline{A_i}$ is substituted with $\tilde{A}_i$ which is the median of the ith group. Thus:

$$Z_{ij} = \left| A_{ij} - \tilde{A}_i \right| \quad (3)$$

where:

A_{ij} refers to the values of a sub-group.

$\tilde{A}_i$ refers to the median for the ith sub-group.

The ML statistic conforms to the F-distribution in the same way as the L statistic and is identical to it with the difference that the mean terms or the $\dot{Z}$'s in expression [2] are now applied on the median adjusted Z_{ij} scores (of expression [3]).

The modification increases the test's ability to withstand violations from normality. The presence of the median, which compared to the mean is less affected by extreme values, grants symmetrical and non-symmetrical data alike an equally likely chance of being on either side of it. The implication of this is that all participating observations would be regarded (more or less) as equally important and thus as equally useful in the information they relinquish.

Put another way the modification provides the test with good robustness against non-normal data whilst allowing it to retain good power. This has been the major finding of Conover *et al*. (1981) who in conducting extensive simulations involving different distributions, sample sizes, means and variances found the ML test to be one of the most robust and powerful ones for testing the HOV assumption.

The robustness of the test refers to its ability to not detect falsely non-homogenous groups when the underlying data is not normally distributed and the groups are in fact homogenous. The power refers to the ability of the test to detect non-homogenous groups when the groups are in fact non-homogenous.

In practice using the constructed variables one can test the assumption of constant error variances using either a two-sample t-test procedure or a (two-group) One-way ANOVA one for its F-test. The ANOVA F statistic will have approximately the F distribution when the variances of the k groups are equal, and this will be true asymptotically when the sample sizes of all groups are equal and tend to infinity. If the p-value (of the t or F-test) is less than the level of significance (i.e. the level of risk a researcher is prepared to accept) than there is a difference in the variances of the groups and the decision is to reject Ho. Conversely if the p-value is greater than the level of significance Ho should not be rejected.

In summary the ML test:

- Functions on a not so strict assumption of normality.
- Is computationally easy to carry out. It can be computed either via an ANOVA procedure or be carried out via a two sample t-test; and
- Is not asymptotically restricted (i.e. it applies well to small and large samples).

APPLICATION

The ML, GQ, BPG, and W test were applied to a regression performed on a data set covering the period 1950-1999. Regressed was South Africa's real Gross Domestic Product (RGDP) on the country's real Gross Domestic Savings (RGDS), real Gross Domestic Investment (RGDI), real Personal Consumption Expenditure (RPCE), and real General Government Expenditure (RGGE)[2]. The variables real values were obtained by deflating their nominal ones by the Consumer Price Index (with a base of 1995 = 100). It should be borne in mind that this regression is done for illustrative purposes only in order to examine how the tests perform in the face of heteroscedasticity. It does not represent a formal model of the South African economy and should not be treated as such. The results from the tests are presented in Table 1 (page 16). Before discussing the results it is briefly described what the computations of the GQ, BPG, and W tests entail.

(a) GQ Test

The GQ test requires the identification of the independent variable considered to be associated with the heteroscedasticity.

Once this variable is identified the observations in the data set are ranked according to the values of that variable itself ranked in an ascending order. Once the data is ordered in this way k central observations are omitted from it with the remaining n–k observations being divided into two groups with (n–k)/2 observations in each group. Regarding the omissions there are differences of opinion as to how many k observations are to be omitted although it is accepted that too little or too many observations should not be omitted.

The purpose behind the omissions is to sharpen the difference between the small and the large variance group. If too few observations are omitted the test may fail to detect the presence of heteroscedasticity; i.e. its power is weakened.

Since one would like to retain the data largely in tact the omission of too many observations is not desired either. In this experiment the middle 5 observations were omitted.

Once the split is done and there are 2 groups regressions are performed on each group to obtain the GQ statistic given by:

$$GQ = \left[RSS_2 / df_2 \right] / \left[RSS_1 / df_1 \right] \tag{4}$$

In expression [4] RSS_2 and RSS_1 are the residual sum of squares for the regression of group 2 and 1 respectively and df_2 and df_1 are the degrees of freedom corresponding to RSS_1 and RSS_1.

The GQ statistic follows the F-distribution. The GQ test requires the normality assumption to be fulfilled; i.e. the residuals of the parent regression should be normally distributed, although it relies on it less strictly than the BPG test.

---

[2] Data Source: *1990-2000* Quarterly Bulletins of the South African Reserve Bank.

(b) BPG test

The limitations of the GQ test, in terms of identifying the correct independent variable with which to order the data, and the number of central observation to omit, can be avoided with the use of the BPG test.

Here the residual sum of squares of the parent regression is divided by the total number of observations, with the value so derived being divided against the residuals of the parent regression. These newly formed residuals are now regressed against the explanatory variables, with this regression's explained sum of squares being used to find the BPG statistic, which is given by:

[5] $$BGP = \frac{1}{2} x \text{ (Explained Sum of Squares)}$$

The BPG statistic asymptotically follows the chi-square distribution with n degrees of freedom where n is the number of regressors.

The BPG test is strictly confined to the normality assumption, i.e. the residuals of the parent regression should be normally distributed, compared to the GQ test.

(c) White's test – W test

The W test is an alternative test of heteroscedasticity, which is not demanding of the normality assumption being fulfilled. When there is lack of normality the test can be counted on to detect heteroscedasticity in contrast to the above tests. In performing the W test one obtains the residuals of the parent regression, and regresses their squares against the explanatory variables of the parent regression, their squared terms and their cross products.

From this auxiliary regression one obtains its R^2 or adjusted R^2 value to calculate the W statistic which asymptotically follows the chi-square distribution with degrees of freedom equal to the number of regressors, and given by:

[6] $$W = \text{number of observations } xR^2$$

The biggest disadvantage of the W test is its rapid consumption of degrees of freedom, which may result in one not being able to obtain a value for the test and ascertain the presence of heteroscedasticity, especially when one deals with non-normality, and this is the only test to use. If a model has several regressors, then introducing all the regressors, their squared and cross product terms can quickly consume degrees of freedom, which may leave an insufficient number of observations with which to perform the test.

(d) ML test

In this exercise the ML test was computed using the residuals of the parent regression and performing a one-way ANOVA on them once they were redefined in accordance with expression [3]. As has already been stated the ML test is useful in cases where the normality assumption fails, and as such can be viewed as an alternative to the W test.

Discussion of Findings

The residual plots of the predicted values and of the explanatory variables indicate the presence of heteroscedasticity in the parent regression, i.e. the one of RGDP on RPCE, RGGE, RGDI, and RGDS. The plots of the residuals versus the predicted values, RGDI, and RGDS, follow a fan (possibly a bowl) shape, whereas the plots against RPCE and RGGE follow a more complicated (possibly an oval) pattern where the residuals move from low to high to low value districts as one moves along the RPCE and RGGE axes. Overall the plots suggest that the regression is plagued by heteroscedasticity. What do the tests show?

The performance of the tests was judged at the 10% (lax) and 5% (strict) level of significance, which were set prior to the experiment being conducted.

The normality assumption was tested from its individual components (i.e. skewness and kurtosis) in addition to an overall test of normality, i.e. the omnibus one, which combines the distribution's skewness and kurtosis into a single measure to test for normality.

At this point it is instructive to recall the GQ test relies on the normality assumption less strictly than the BPG test whilst the W and ML test work under conditions where this assumption has failed.

Even though the residuals of the parent regression are not normally distributed they have a symmetrical dispersion. The p-value of the skewness test, p-value = 0.27, is above either level of significance whilst the p–value of the omnibus or overall normality test, p–value = 0.02, is below either level of significance. In the former case one can not reject the null hypothesis of the residuals being symmetrically distributed where as in the latter case one can reject the null hypothesis of the residuals being normally distributed. The hat diagonal plot (of the externally studentised residuals) shows that there are no observations that are both outlying and influential. Shortly put the experiment is outlier free.

The implication of the residual distribution is that one may expect:

(a) The GQ test to detect the presence of heteroscedasticity since even though the residuals are not normally distributed their distribution is symmetrical.

(b) The BPG to fail to detect the presence of heteroscedasticity given that test relies strictly on the normality assumption being fulfilled. Failure of the normality assumption would weaken if not eliminate the test's ability to detect heteroscedasticity.

(c) The W test may not perform as desired. Even though the normality assumption has failed, the symmetrical one is satisfied. This may weaken the test's ability to detect heteroscedasticity.

(d) The ML test may not perform as desired for reasons the same as the W test.

Looking at Table 1 one can see that the GQ test statistic exceeds its critical value, such that the null hypothesis of homoscedasticity can be rejected (at $\alpha = 0.05$, and 0.10), indicating that the regression suffers from heteroscedasticity. The GQ test performs as expected.

TABLE 1
TEST RESULTS - SUMMARY

Test	Value	Critical Value		Conclusion	
		$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$
GQ	7.71	2.23	1.86	Reject Ho	Reject Ho
BPG	$1.89E^{-29}$	9.49	7.78	Accept Ho	Accept Ho
W	22.49	23.68	21.06	Accept Ho	Reject Ho
ML	3.97	4.00	2.79	Accept Ho	Reject Ho

Note: Refer to appendix section for calculations and for the power of the ML test.

On the other hand we can see the sensitivity of the BPG test to the normality assumption. The BPG test performs as stipulated. The BPG statistic's value is practically zero and is well below the critical value (for both the 10% and 5% level of significance). With the use of the BPG test the presence of heteroscedasticity can not be detected.

The W test performs as stipulated. The test is able to detect heteroscedasticity at the 10% level of significance but fails to do so at the 5% level even though the normality assumption has failed. This failure in consistency by the test can be attributed to the residuals being symmetrically dispersed.

Here one should note the point made earlier about the W test and the degrees of freedom it consumes, which for this case were 14. If there were additional variables included in the parent model the W-test may well have been impossible to carry out.

The point here is that for small or medium sized samples where regression involves a number of regressors, the W test may not be possible to calculate, i.e. it may leave an insufficient number of observations on which to perform the auxiliary regression.

The ML test performs as stipulated. The test is able to detect heteroscedasticity at the 10% level of significance but fails to do so at the 5% level even though the normality assumption has failed. This failure in consistency by the test, as in the W test, can be attributed to the symmetrical dispersion of the residuals. The conclusion that should be drawn here is that for the W and ML test to perform as envisaged failure of the normality assumption has to be accompanied by failure of symmetry too.

The power of the ML test is 63% (page 22) meaning that it would detect the presence of heteroscedasticity (given that it exists) in 63% of the time or in 63 out of every 100 trials. This is better than flipping a coin but may not be as high as say having a power of 80%, 90%, or even higher, which can be attributed to the same reason as for the test's failure in consistency.

Overall the tests conform to expectations. The ML test's contribution is that it represents another heteroscedasticity diagnostic that can be useful to employee in cases where there is no evidence of normality and symmetry. The test

is computationally easy to carry out and in relation to the W test one need not be concerned with how rapidly degrees of freedom are consumed.

Conclusion

The modified Levene test as well as the GQ, BPG, and W tests, represent a numerical confirmation of what one observes on the residual plots. Therefore no homo or heteroscedastic test should be regarded as the definitive guide in diagnosing the presence or lack of HOV. The residual and probability plots should still be examined.

References

1. Breusch, T. and Pagan, A. (1979). "A Simple Test for Heteroscedasticity and Random Coefficient Variation", *Econometrica*, Vol. 47, 1287-1294.
2. Brown, M.B. & Forsythe, A.B. (1974). "Robust Tests for the Equality of Variances", *Journal of the American Statistical Association*, Vol. 69, 364-367.
3. Conover, W.J.; Johnson, M.E. and Johnson, M.M. (1981). "A Comparative Study of Tests for Homogeneity of Variances, With Applications to the Outer Continental Shelf Biding Data", *Technometrics*, Vol. 23, 351-361.
4. Godfrey, L. (1978). "Testing for Multiplicative Heteroscedasticity". *Journal of Econometrics*, Vol. 8, 227-236.
5. Goldfeld, S. and Quandt, R. (1972). *Non-linear Methods of Econometrics.* North-Holland, Amsterdam.
6. Gujarati, D. (1995). *Basic Econometrics.* 3rd Ed. McGraw-Hill Book Co, Singapore.
7. Levene, H. (1960). *In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling.* I. Olkin *et al*., Eds. Stanford University Press, Stanford Calif., 278-292.
8. Ott, L. (1993). *An Introduction to Statistical Methods and Data Analysis.* 4th Ed. Wadsworth Publishing Company, Belmont, Calif.
9. Pindyck, R. and Rubinfeld, D. (1998). *Econometric Models and Economic Forecasts.* 4th Ed. McGraw-Hill Book Co, Singapore.
10. Snedecor, G.W. and Cochran, W.G. (1989). *Statistical Methods.* 8th Ed. Iowa State University Press.
11. White, H. (1980). "A Heteroscedasticity Consistent Covariance Matrix Estimator and a Direct Test of Heteroscedasticity", *Econometrica*, Vol. 48, 817-818.

APPENDIX

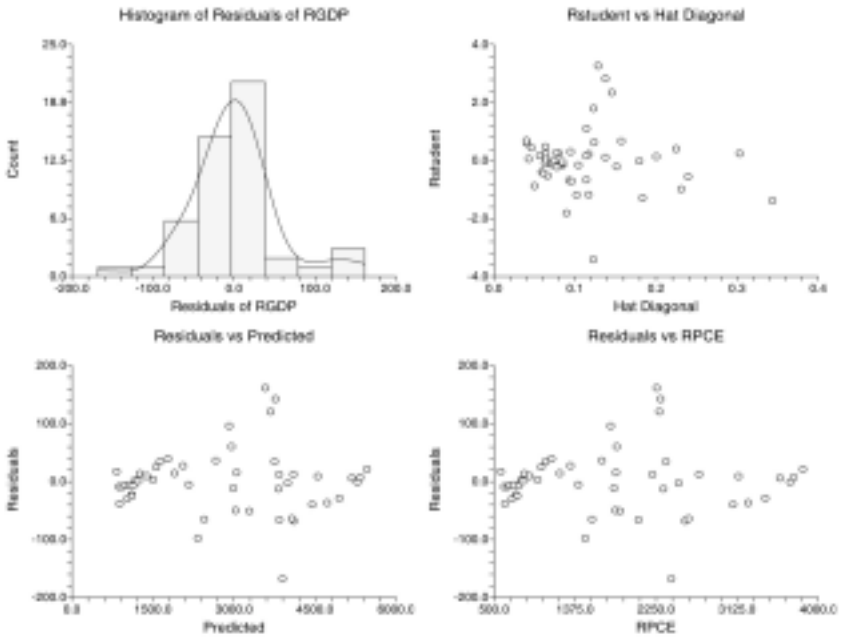
PARENT REGRESSION
Dependent Variable: RGDP

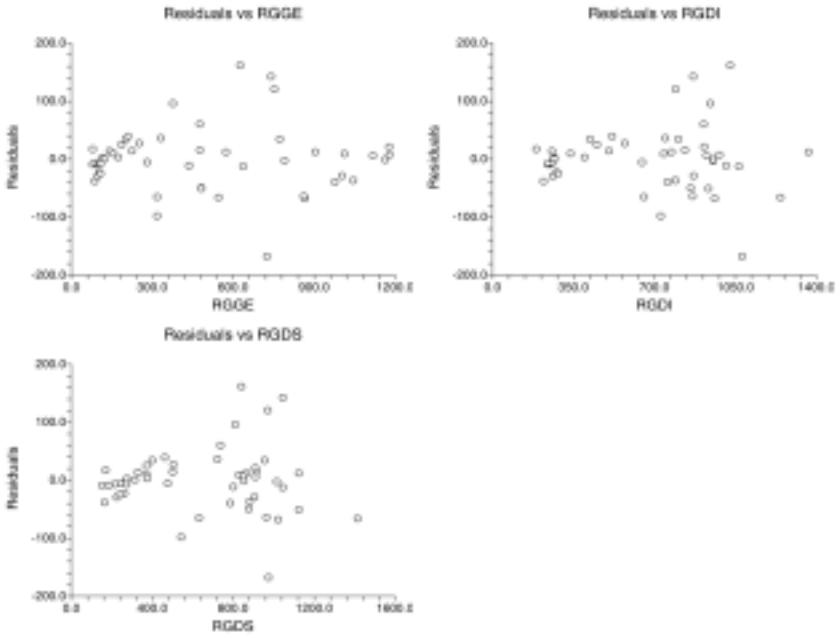
Model
7.355296+ 1.045294*RPCE+ .3958012*RGGE+ .3396563*RGDI+ .7336847*RGDS

Residuals - Normality Tests

Assumption	Value	Probability	Decision(10%)
Skewness	1.1115	0.266349	Accepted
Kurtosis	2.5754	0.010013	Rejected
Omnibus	7.8680	0.019565	Rejected

Plots Section





GQ TEST (RANKING DONE BY RGDS)

OLS REGRESSION 1 (Count 22)

Analysis of Variance Section

Source	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level
Intercept	1	4.064054E+07	4.064054E+07		
Model	4	4184784	1046196	1379.9969	0.000000
Error	17	12887.95	758.1148		
Total(Adjusted)	21	4197673	199889.2		

Root Mean Square Error	27.53388	R-Squared	0.9969
Mean of Dependent	1359.153	Adj R-Squared	0.9962
Coefficient of Variation	2.025812E-02	Press Value	35215.14
Sum |Press Residuals|	601.5283	Press R-Squared	0.9916

OLS REGRESSION 2 (Count 23)

Analysis of Variance Section

Source	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level
Intercept	1	3.868012E+08	3.868012E+08		
Model	4	1.221187E+07	3052967	522.0702	0.000000
Error	18	105260.6	5847.81		
Total(Adjusted)	22	1.231713E+07	559869.5		

Root Mean Square Error	76.47097	R-Squared	0.9915
Mean of Dependent	4100.908	Adj R-Squared	0.9896
Coefficient of Variation	1.864733E–02	Press Value	159893.8
Sum |Press Residuals|	1401.699	Press R-Squared	0.9870

BPG TEST

OLS REGRESSION (Count 50)

Dependent Variable: Constructed BPG Test Variable

Model
0+ 0*RPCE+ 0*RGGE+ 0*RGDI+ 0*RGDS

Analysis of Variance Section

Source	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level
Intercept	1	2.416826E–31	2.416826E–31		
Model	4	3.772375E–29	9.430938E–30	0.0000	1.000000
Error	45	1.664628E–02	3.699173E–04		
Total(Adjusted)	49	1.664628E–02	3.3972E–04		

Root Mean Square Error	1.923324E–02	R-Squared	0.0000
Mean of Dependent	6.952446E–17	Adj R-Squared	0.0000
Coefficient of Variation	0	Press Value	2.142971E–02
Sum |Press Residuals|	0.6924834	Press R-Squared	–0.2874

W-TEST

OLS REGRESSION (Count 50)

Dependent Squared Residuals (parent Regression)

Model
–11986.56+ 47.37957*RPCE–55.08334*RGGE–21.33363*RGDI–27.329*RGDS–1.796328E–02*RPCE2–.2203785*RGGE2+ .1153379*RGDI2+ .1737958*RGDS2+ 8.962686E–02*RPCERGGE+ 1.436439E–02*RPCERGDI–2.727322E–02*RPCERGDS+ .091984*RGGERGDI+ 5.469556E–02*RGGERGDS–.304424*RGDIRGDS

Analysis of Variance Section

Source	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level
Intercept	1	4.511029E+08	4.511029E+08		
Model	14	8.719064E+08	6.227903E+07	2.0439	0.043309
Error	35	1.066472E+09	3.047064E+07		
Total(Adjusted)	49	1.938379E+09	3.955875E+07		

Root Mean Square Error	5520.022	R-Squared	0.4498
Mean of Dependent	3003.674	Adj R-Squared	0.2297
Coefficient of Variation	1.837757	Press Value	3.917399E+09
Sum |Press Residuals|	289657	Press R-Squared	–1.0210

ML TEST

ONE-WAY ANOVA (Count 50)

Analysis of Variance Table

Source Term	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level	Power (α =0.10)
A: Redefined Residuals	1	6309.731	6309.731	3.97	0.051990*	0.625555
S(A)	48	76267.8	1588.912			
Total (Adjusted)	49	82577.52				
Total	50					

* Term significant at alpha = 0.10.