Proceedings of the 1990 Winter Simulation Conference
Osman Balci, Randall P. Sadowski, Richard E. Nance (eds.)

SIMULATION AND REGRESSION: DEALING WITH THE ASSUMPTION OF A COMMON ERROR VARIANCE

Peter D. Welch

IBM Thomas J. Watson Research Center
Yorktown Heights, New York 10598

ABSTRACT

In the application of regression analysis to discrete event simulation the assumption of a common error variance is usually questionable. In this paper procedures are suggested for testing this assumption, for dealing with departures from it within the regression protocol, and for avoiding the difficulty by controlling the length an/or number of simulation runs at each experimental point.

1. INTRODUCTION

We will be discussing the application of regression models within the context of discrete event simulation. The range of such applications is very broad extending from simple polynomial fitting to the application of sophisticated experimental designs. In all cases we have

$$Y = X\beta + \varepsilon$$

where Y is the vector of responses, X is the matrix of independent variables, β is the vector of model coefficients and ε is the vector of errors. The errors are assumed to be uncorrelated, normal random variables with zero mean and an unknown common variance, σ^2. We assume that each element of Y is an estimate obtained from a simulation model with a fixed set of parameters. These parameters are reflected in the matrix X. We assume each element of Y to be an average over a single run or multiple runs but with no change in model parameters. The parameters of the model would change as we moved from one element of Y to another. We let n be the length of Y and let k be the index running from 1 to n. Now in this context, since the elements of Y are typically averages, the normal assumption is not troublesome. Further, since each element of Y comes from runs with different random number seeds, the assumption of zero correlation is again, typically, not a problem. The assumption of a common variance is, however, usually questionable. This paper discusses alternate ways of dealing with this difficulty. We let σ_k^2 be the variance of the kth element of ε. Under the assumption of the standard regression model, $\sigma_k^2 = \sigma^2$ for all k.

In non-simulation environments it is sometimes reasonable, through replications with the parameters remaining constant, to get internal (i.e. from within the data) estimates of the variances σ_k^2 which are independent of the model being fit. This is not always the case however because of the cost and other circumstances of experimentation. These replications provide for more powerful tests of the common variance assumption and provide a pure error sum of squares which can be used as the basis for a goodness of fit test for the regression model. In the simulation environment such internal estimates are usually readily available. Also, since the number and/or length of runs can be easily increased, it is possible to control the values of the estimates and hence the variances themselves. In other environments the cost and availability of experimentation are such that this is usually impractical.

Hence in the discussion below we will consider the problem of the assumption of a common variance for the following cases:

1) no internal estimates of the error variances σ_k^2 $k = 1, \ldots, n$,
2) internal estimates of the error variances but no attempt at control,
3) internal estimates of the error variance are used to control the length and/or number of runs to achieve a common error variance.

This specific problem is discussed at some length in Section 13.3 of Kleijnen [1987]. This is also an excellent reference for all aspects of the application of regression to simulation. In fact, Kleijnen coined the term "metamodel" which is often used to distinguish the regression model from the underlying simulation model.

2. NO INTERNAL ESTIMATES OF THE ERROR VARIANCE

As we pointed out above it is relatively simple and very beneficial in the simulation environment to generate internal estimates of the error variances σ_k^2 $k = 1, \ldots, n.$, However, the experimenter may not wish to make the effort to obtain internal estimates, or the system he is using may not provide the proper tools or flexibility to obtain such estimates, or there may not be enough data in each run to obtain valid estimates. In this case the methods available to test the assumption of a common variance are the standard regression techniques involving the analysis of residuals from the fitted model. The residuals are tested for normality, are plotted against the estimated responses (the fitted values) and are plotted against the independent variables (the columns of X). These methods are summarized in any modern regression text (see e.g. Chapter 3 of Draper and Smith [1981]).

If the residuals pass these tests, a reasonable application of regression can be assumed. If they fail any of the tests there are two possible remedies. If they exhibit a funnel like pattern when plotted against the estimated responses, fitting a model to an appropriate transformation of the responses can be considered. Frequently a logarithmic transformation is used because frequently there is a percentage type error. Transforming the responses may not be desirable because the model required for the transformed data may be considerably more complex than the model for the untransformed data. This can be particularly serious if the original model is simple and has appealing a priori meaning. If, however, transforming the responses leads to a simpler model the experimentor has the best of both worlds. The other alternative is weighted least squares. However, in this situation it would probably be quite difficult to generate a reasonable set of weights. Weighted least squares assumes that $\sigma_k^2 = \sigma^2/w_k$ where σ^2 is unknown but the w_k are known. In practice, however, the w_k usually have to be estimated from the data. To do this a plot of the residuals against one of the dependent variables has to suggest a simple relationship between the variance of the residuals and the dependent variable. Then this relationship has to be quantified by fitting a least squares model (e.g. fitting a straight line) to estimates of the σ_k^2 as a function of the dependent variable. If, as in this case, you don't have replicated observations this requires generating approximate replications by grouping observations according to the value of the dependent variable. For an example, see Draper and Smith [1981]. *pp.*112 – 115. For a general discussion of weighted least squares see Section 2.10 of Draper and Smith [1981]. In weighted least squares there is no transformation of the

model and all the usual statistical tests can be rigorously applied. However, any orthogonality built into the model will, in general, be destroyed. This is not serious in simple polynomial fitting but would complicate the application of orthogonal experimental designs.

3. INTERNAL ESTIMATES OF THE ERROR VARIANCE BUT NO ATTEMPT AT VARIANCE CONTROL

We next consider the case where the experimenter obtains internal estimates of the error variances but makes no attempt based on them to control the number and/or length of the runs. We will concentrate the discussion on the case where independent replications (in the simulation sense) have been obtained either directly or through the method of batch means. We assume that for each experimental point we have the same number, R, of replications. Having a common number of replications is reasonable and easily achievable. We will see below that it has advantage as concerns the testing of the assumption of a common variance. Hence we assume we have R uncorrelated random variables with a common mean and a common variance at each experimental point. The variance, of course, may change from point to point. This is what we wish to test. The responses, the elements of Y, are the averages of these R quantities.

We let s_k^2 be the sample variance of these R variables at the kth point. Then s_k^2/R is the estimate of σ_k^2. The quantity $(R-1)s_k^2/\sigma_k^2$ has a χ^2 distribution with R-1 degrees of freedom. Under the assumption of a common variance the s_k^2's should all have the same distribution. There is a standard formal test of this assumption, Bartlett's Test. This test should be applied by the experimenter but it should be supplemented with the graphics tests we will next describe. First, under the assumption of a common error variance, $(R-1)s_k^2/\sigma^2$ $k=1 \ldots n$ is an independent, identically distributed sample from a χ^2 distribution with R-1 degrees of freedom. Hence if we let $s^2 = \sum^n s_k^2/n$, $(R-1)s_k^2/s^2$ $k=1 \ldots n$ should have approximately a χ^2 distribution with R-1 degrees of freedom and hence should plot as points scattered around the theoretical straight line on χ^2, with R-1 degrees of freedom, probability paper; that is, on a plot of the order statistics of the sample versus the corresponding quantiles of the χ^2 distribution. Goodness of fit tests can also be made. Finally the s_k^2's should be plotted against the responses and the independent variables, in a manner parallel to the analysis of residuals in regression, to see if there are any significant trends. For a good discussion of the general problem of testing for the equality of n theoretical variances see Section 11.6 of Hald [1952]. If a method other than replications were used to estimate the σ_k^2's then a χ^2 approximation to its distribution with the same coefficient of variation could be used in applying the techniques described above.

If the s_k^2 pass all these tests then the assumption of a common variance would be accepted. Furthermore s^2 would be the best estimate of σ^2 and $n(R-1)s^2/\sigma^2$ would have a χ^2 distribution with n(R-1) degrees of freedom. This could be compared to the error sum of squares of the residuals in a goodness of fit test of the regression model.

If the s_k^2 do not pass the test then the experimenter has the same alternatives as were described earlier with respect the analysis of residuals: he can transform the response variable or he can apply weighted least squares. However, there is more information and it is in a much more usable form than in the residuals from the fitted model. It is much easier to see and quantify trends in plots than it is to see and quantify changing patterns of spread. For example, if a logarithmic transformation will achieve a constant variance then the s_k should be in constant proportion to the Y_k. Concerning weighted least squares, as we remarked earlier, the major problem is the estimation of the weights, w_k. If a pattern of trend can be discerned in the s_k^2 when plotted against one of the independent variables then curve fitting (or more generally, least squares) can be applied to quantify the trend and generate the weights. The weights would be the inverses of the values read off the fitted curves at the values of the independent variables. This is what is done in the example previously cited on pp. 112-115 of Draper and Smith [1981]). Using weights $w_k = 1/s_k^2$ is proposed in Kleijnen [1987] when R is large. Large is defined to be greater than 25 on the basis of studies by Nozari [1984]. In this case we are essentially assuming that $s_k^2 = \sigma_k^2$. The goodness of fit test for the regression model would involve comparing the estimate of σ^2 from the error sum of squares with unity. It would be a χ^2 test rather than an F test. The two techniques, transformation of the responses and weighted least squares, would continue to have the difficulties cited earlier.

4. INTERNAL ESTIMATES OF THE ERROR VARIANCE BUT WITH ATTEMPTS AT VARIANCE CONTROL

We will next discuss the possibility of attempting to insure that there is a common error variance by exercising control over the number and/or length of runs. This could be done through a variety of sequential procedures. We will briefly discuss two. The first is designed to produce a set of responses for which the common variance assumption is reasonable. The second is designed to generate responses with a constant, predetermined error variance by producing responses with a common constant estimated error variance. In these approaches rather than adjusting to a problem with the error variances by changing the model or the regression procedure we eliminate the problem by varying the lengths and/or number of runs at each of the sample points. Such approaches will become increasingly reasonable as the cost of computing declines.

Consider the following protocol. An initial experiment of R independent replications at each experimental point is conducted. We assume R is large enough (say $R > 10$) so that reasonable estimates, s_k^2, of the σ_k^2 can be obtained. Using the methods discussed above, the assumption of a common variance is tested. If it is accepted the experimenter proceeds to the analysis of the regression model. If it is rejected a procedure paralleling weighted least squares is followed. However, instead of deriving weights, w_k, to correct for unequal variances, the experimenter derives numbers of replications, n_k, designed to achieve the assumption of a common variance. To avoid throwing away data, a target value, σ_T^2, less than or equal to the minimum over k of the s_k^2 is set and numbers of replications are determined by $n_k = s_k^2/\sigma_T^2$. At the kth point $n_k - R$ more replications would be taken. The final estimated responses would be the averages over all the n_k replications. This technique corresponds to the extreme form of weighted least squares but has the advantage that it could be easily automated. Alternative procedures could base n_k on the other schemes for determining weights described earlier; i.e. those based on trends observed in s_k^2 plotted against the dependent variables. In all these cases there would be an easily derivable estimate of the amount of computing time required to complete the experimental runs. If this were too great the experimenter could apply weighted least squares. If the method of batch means is employed this procedure and the one to be discussed next require that the simulation software can stop sets of runs for statistical analysis and start them again where they left off as if they had not been interrupted. We will argue below that such sequential capability is destined to become an essential feature of simulation software.

A second alternative would be to place the experimentation (i.e. the simulation runs) under strict sequential control and to terminate runs associated with each experimental point when the estimate of σ_k^2 first reaches some target value, σ_T^2. This value could either be be set a priori or could be set after a first stage of experimentation. Let V_k be the variance of an individual replication at the kth point. Let N_k be the point at which the sample variance first achieves a value less than or equal to σ_T^2. Then, in this method, Y_k is the average over the N_k replications (N_k is a random variable). Let σ_k^2

be the variance of the resulting Y_k. The major problem is that σ_T^2 underestimates σ_k^2. However, it underestimates it by a factor which approaches one as σ_T^2/V_k becomes small. This phenomena is reflected in the literature of sequential methods of fixed width confidence interval generation, where the sampling is continued until the sample standard deviation reaches a predetermined level. The sample standard deviation must be increased if a confidence interval based on the normal distribution is to provide the correct coverage (see e.g. Anscombe [1953]). In Figure 1 we have plotted σ_k^2/σ_T^2 as a function of σ_T^2/V_k. This curve was obtained from a simulation study of i.i.d. normal random variables. 95% confidence intervals are indicated. These intervals are individual and do not hold jointly. Now if we knew V_k we could correct for this factor. But, of course, if we knew V_k we wouldn't have any of these problems. Hence the only thing we can do is make sure that σ_T^2/V_k is in the region where the correction is insignificant. We see from Figure 1 that this requires σ_T^2/V_k to be approximately .01 which means, on the average, 100 replications or more. (In many instances in practice, this may very well be prohibitive. Although, as we will stress below, computational resources are becoming less and less of a constraint.) In this case we have estimates of σ_k^2 for all k which are essentially equal to a predetermined constant σ_T^2. We do not have estimates which have a χ^2 distribution with R-1 degrees of freedom. The F test of goodness of fit would become a χ^2 test and the t confidence intervals on the coefficients would become normal confidence intervals. There would be no need to consider transforming the model and no destruction of the orthogonality of a design.

Both of the cases discussed in this section require the ability to stop and reschedule a set of simulation runs based on statistical considerations. This is now done in some systems to generate confidence intervals of fixed width or fixed relative width (a fixed percentage of the point estimate). As we will discuss below, in the future such experimental flexibility will become critical to the effective and efficient application of simulation.

5. SUMMARY

We have discussed the problem of the common variance assumption in the applications of regression models within the context of discrete event simulation. In Sections 2 and 3 we discussed passive approaches which assume that a set of simulation runs are made and are then subjected to statistical analysis. In Section 4 we discussed more dynamic approaches which take advantage of the ready availability of additional data in the simulation context. There is no other area of applied statistics where you can garner additional data with greater ease than in simulation. Whatever cost constraints might exist are being rapidly removed by the burgeoning availability of computing power. Hence in the future we will see closer and closer coupling of statistical calculations and simulation run control. Sequential procedures will be applied not only to model fitting but also to generating model coefficient confidence intervals of fixed widths, to model and parameter selection, to adaptive response surface determination, etc. Only by the close coupling of statistics and run control can efficient and effective application of simulation be made. There is a deep need to investigate these issues and to build appropriate high level controls into simulation packages.

ACKNOWLEDGEMENT

I would like to thank Ed MacNair for performing the calculations leading to Figure 1 and for help with the rigors of formatting the paper.

REFERENCES

Anscombe, F.J.(1953), "Sequential Estimation", *JRSS Series B*, 15, 1-21,

Draper N.R. and Smith H. (1981), *Applied Regression Analysis*, Wiley, New York

Hald A.(1952), *Statistical Theory with Engineering Applications*, Wiley, New York

Kleijnen J.P.C. (1987), *Statistical Tools for Simulation Practitioners* , Marcel Dekker, New York

Nozari A. (1984), "Generalized and Ordinary Least Squares with Estimated and Unequal Variances", *Communications in Statistics, Simulation and Computation*, 13, no.4, 521-537

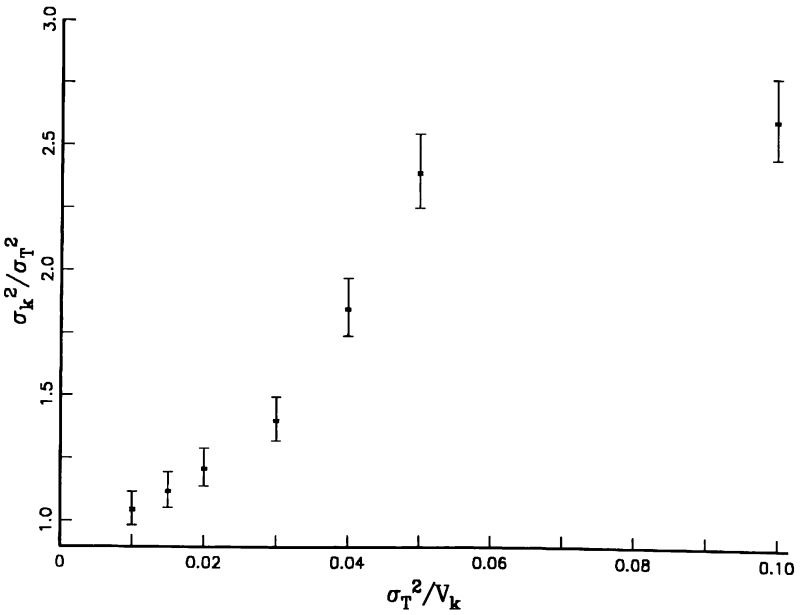

Figure 1. σ_k^2/σ_T^2 as a Function of σ_T^2/V_k with 95% Confidence Intervals