Bonferroni Adjustments in Tests for Regression Coefficients

Daniel J. Mundfrom **Jamis J. Perrett** **Jay Schaffer**

Adam Piccone **Michelle Roozeboom**

University of Northern Colorado

A common application of multiple linear regression is to build a model that contains only those predictors that are significantly related to the response. In so doing, tests regarding the unique contribution of individual predictors to the model are often performed. It is not uncommon for practitioners to conduct each of these tests at the nominal $\alpha = 0.05$ level, without regard to the effect that this practice may have on the overall Type I error rate. This research investigated the utility of making a Bonferroni adjustment when conducting these tests of the partial regression coefficients. Simulated multivariate normal populations with various correlational structures, different numbers of predictors in the model, and differing numbers of "significant" predictors in the model were generated. Ten thousand samples, 5000 each of sizes 50 and 300, were drawn from each population condition and a multiple regression analysis was performed on each sample. In every case, the observed significance levels for the Bonferroni-adjusted tests were controlled below the nominal 0.05 level as expected, and in most cases substantially lower than the observed significance levels for the unadjusted tests.

Multiple Linear Regression (MLR) is a popular statistical procedure for investigating the nature of the relationships among several numerical characteristics. Typically, one of the characteristics is identified as the dependent or response variable and the remainder of the characteristics are called independent or predictor variables. Most introductory level statistics texts identify regression analyses as having two uses: 1) to estimate the average response for a sample of individuals having various values for each variable in a set of predictors, and 2) to predict the response for a "new" individual for whom only values of the predictors are measured/observed. In either case, a linear model, based on observed data is used to make the estimation or prediction.

In some applications, the researcher knows which variables should be used as predictors in the model and the purpose of the analysis is to predict the value of the response using previous information regarding the nature of the variables' relationships with each other. Data are collected on the predictor variables and the model is used to predict the value of the response variable for one or more "new" individuals. In other situations, the researcher is interested in determining which, if any, of several numerical characteristics are significantly related to a specific outcome. Data are collected on all the variables of interest—the dependent variable and all the independent variables—and an MLR analysis is performed to build a model that may later be used for prediction, i.e., the researcher determines which of these predictor variables displays a significant unique ability to explain variation in the response variable. While both of these applications of MLR are useful and appropriate, it is the latter situation which is the focus of this research.

When the purpose of the regression analysis is to determine which independent variables are unique contributors to the model, it is typical for the researcher to perform separate tests of the partial regression coefficients (i.e., the beta coefficients) for each predictor. Those predictors for which the test of the beta coefficient has a p-value that is less than the specified α-level are deemed to be making a unique contribution to the model and will be retained in the model as a predictor variable. On the other hand, those variables for which the test of the beta coefficient has a p-value that is larger than that specified α level are not identified as useful predictors and may be dropped from the model in the interest of parsimony. It is not an uncommon practice for each of these separate tests to be conducted at the nominal 5% significance level. The purpose of this research is to investigate whether conducting each of these tests at $\alpha = 0.05$ inflates the overall Type I error rate for the collection of all these tests and if a Bonferroni-type adjustment to the α level for each test would be appropriate to control the overall α-level closer to that nominal level.

Making adjustments to the significance level of a statistical test when multiple tests are conducted on the same data is a common statistical practice. Many procedures have been developed for making such adjustments. One of these procedures is the Bonferroni adjustment.

The Bonferroni adjustment is based on an inequality in probability theory that was derived by C. E. Bonferroni. The inequality states that if A_1, A_2, . . ., A_k represent k events and A_1', A_2', . . ., A_k' represent the corresponding complements, then

$$P\left(\bigcap_{i=1}^{k} A_i'\right) = 1 - \sum_{i=1}^{k} P(A_i).$$

An application of the Bonferroni inequality, the Bonferroni adjustment, is one of the commonly used methods for adjusting the significance levels of individual tests when multiple tests are performed on the same data. For example, consider three statistical tests being performed simultaneously, each at level α, such that

A = {a Type I error occurred in test 1}
B = {a Type I error occurred in test 2}
C = {a Type I error occurred in test 3}

so that $P(A) = P(B) = P(C) = \alpha$. Under these conditions, the probability that at least one Type I error occurs in the three tests, i.e., the overall significance level of the three tests, is inflated. The Bonferroni inequality provides an upper bound for the overall level of significance such that,

$$\begin{aligned} P(\text{at least one Type I error occurs}) &= 1 - P(\text{no Type I errors occur}) \\ &= 1 - P(A' \cap B' \cap C') \\ &< 1 - \{1 - [P(A) + P(B) + P(C)]\} \\ &= 1 - [1 - 3\alpha] \\ &= 3\alpha. \end{aligned}$$

The Bonferroni adjustment divides the nominal significance level, α, by the number of tests being performed simultaneously to prevent the overall level of significance from exceeding the nominal level, α. The adjusted level of significance, in general α/k for k tests, is used to conduct each of the k individual tests.

Virtually every statistics textbook recommends some type of adjustment when pairwise comparisons of means are performed as a follow-up to a significant ANOVA (see, for example, Glass and Hopkins, 1996; Hinkle, Wiersma, & Jurs, 1998; Agresti and Finlay, 1997). It is rare, if ever however, that these same textbooks would recommend these same types of adjustments when conducting tests of main effects and interactions in a factorial ANOVA design or tests of the partial regression coefficients in a MLR analysis, yet these two situations also consist of multiple tests being conducted on the same sample of data. Hinkle, Wiersma & Jurs (1998), for example, provide a multiple regression example with four predictors in which each partial regression coefficient is tested at $\alpha = 0.05$ to determine if the corresponding predictor variable should be retained in the model. It would not be surprising to find similar examples in just about any text covering basic regression analysis.

Galambos and Simonelli (1996) discuss additional applications of the Bonferroni inequality beyond pair-wise comparisons of means as a post-hoc or a priori follow-up to a significant ANOVA. Simultaneous confidence intervals are presented for differences between pairs of means or variances, along with joint confidence intervals for partial regression coefficients, joint prediction intervals for n new observations, and the detection of outliers in multiple regression. Although Galambos and Simonelli (1996) discuss applications of the Bonferroni inequality in various fields of application, including number theory, extreme value theory, linear programming, and statistical methods, they do not discuss its use for variable selection in a MLR analysis.

For a multiple regression analysis with, say, seven predictor variables, it doesn't appear to be difficult to see that conducting seven separate tests, each at $\alpha = .05$, would inflate the overall Type I error rate for that analysis. Consequently, it also would not appear to be surprising that using a Bonferroni-type adjustment that would conduct each of those tests at $\alpha/7$, i.e., 0.05/7 = 0.0071, would provide some protection for that overall error rate. We see no compelling philosophical perspective that would preclude such adjustments from this type of application. Indeed, Korn and Graubard (1990) compared the power of Bonferroni-adjusted t-tests of the partial regression coefficients and the Wald statistic. Their results were mixed with each method outperforming the other in certain situations, leading them to conclude that,

"One possible interpretation . . . is not that the Bonferroni procedure works well, but that the Wald statistic works poorly . . ." (Korn and Graubard, 1990, p.274). More recently, Foster and Stine (2004) used a modified stepwise selection procedure incorporating a Bonferroni-type adjustment on tests of individual partial regression coefficients in an application using bankruptcy data. Using their method resulted in an earlier end to the stepwise variable selection procedure and that the resulting prediction model performed better than the conventional data mining techniques that were more typically used. Both of these studies indicate that the use of a Bonferroni adjustment to the significance level for tests of individual partial regression coefficients in variable selection when building a prediction model using a multiple regression analysis is a reasonable approach for providing some control over the overall Type I error rate.

Method

In this study, simulated data were used to create populations with known characteristics so that the effect on the overall Type I error rate from varying those characteristics could be determined. Data were generated using PROC IML in SAS. Populations were created that contained data from multivariate normal distributions and varied according to the number of predictor variables in the model, the magnitude of the zero-order correlations among the predictors, the magnitude of the zero-order correlations between the predictors and the response variable, and the proportion of predictors having non-zero correlations with the response variable.

The number of predictor variables in each population was varied across the values 2, 4, 6 and 8. Zero-order correlations among the predictor variables were varied across the values 0.1, 0.3, and 0.5, keeping these correlations low to eliminate any potential multicollinearity problems which could influence the results of individual tests regarding the partial regression coefficients. Correlations between the predictor variables and the response variable were varied across the values 0, 0.4, and 0.8. Because a Type I error can occur only when a non-significant predictor is identified as significant, only models that contained at least one non-significant predictor were investigated, i.e., all of the models investigated in this study had at least one predictor variable that had a correlation of 0 with the response variable. The proportion of predictor variables that had non-zero correlations with the response was varied across the values 0, 1/6, 1/4, 1/3, and 1/2. Only proportions that resulted in an integral value for the number of predictors with non-zero correlations were used in any particular model. For example, with two predictors, only the proportions 0 and 1/2 were used, i.e., we investigated cases with both variables uncorrelated with the response and with only one variable correlated with the response. With four predictors, only the proportions 0, 1/4, and 1/2, were used, i.e., we considered cases with all four predictors, one, and two predictors respectively being uncorrelated with the response. Similar restrictions were used in the scenarios with 6 and 8 predictor-variable models. In all the population conditions investigated, the proportion of predictor variables that had non-zero correlations with the response was restricted to no more than half of the variables in the model. This restriction was used so that the possible number of Type I errors that could occur in any sample did not become too small. We considered two different sample sizes—50 and 300—to see if the amount of data had any influence on the unadjusted and Bonferroni-adjusted tests of the individual partial regression coefficients.

Initially, 10 populations of size 100,000 were generated for each of the population conditions identified above. A multiple regression analysis was performed on each population to determine which predictors were actually significantly related to the response. For each sample size (i.e., 50 and 300), 500 samples were generated from each of the 10 populations in each of the population conditions, for a total of 5000 samples of size 50 and 5000 samples of size 300 for each condition. A multiple regression analysis was performed on each sample and each of the partial regression coefficients was tested at both the unadjusted α-level, 0.05, and the adjusted α-level, $0.05/k$, where k is the number of predictors in the model. A Type I error was defined as interpreting one predictor as significant in the sample that was not identified as significant in the parent population of that sample. The proportion of samples that lead to a Type-I error for any of the predictors in the model was computed as the actual significance level of the test. Actual significance levels for both the adjusted tests and the unadjusted tests were computed and compared.

Table 1. *Actual significance levels for Bonferroni-adjusted and unadjusted tests of the partial regression coefficients for various numbers of predictors, various*

numbers of correlated predictors, and two different sample sizes

number of predictors	number of non-zero correlations with response	sample size	actual significance levels unadjusted	Bonferroni-adjusted
2	0	50	0.0930	0.0476
2	0	300	0.0940	0.0481
2	1	50	0.0511	0.0261
2	1	300	0.0489	0.0244
4	0	50	0.1744	0.0476
4	0	300	0.1792	0.0479
4	1	50	0.1388	0.0364
4	1	300	0.1398	0.0373
4	2	50	0.0993	0.0252
4	2	300	0.0908	0.0264
6	0	50	0.2450	0.0462
6	0	300	0.2565	0.0479
6	1	50	0.2131	0.0381
6	1	300	0.2214	0.0403
6	2	50	0.1827	0.0322
6	2	300	0.1794	0.0324
6	3	50	0.1447	0.0193
6	3	300	0.1340	0.0293
8	0	50	0.3065	0.0468
8	0	300	0.3249	0.0479
8	1	50	0.2810	0.0415
8	1	300	0.2944	0.0425
8	2	50	0.2494	0.0374
8	2	300	0.2603	0.0356
8	3	50	0.2092	0.0312
8	3	300	0.2178	0.0289
8	4	50	0.1732	0.0176
8	4	300	0.1740	0.0180

Results

The actual significance levels for both the unadjusted tests and the Bonferroni-adjusted tests are displayed in Table 1, for both of the sample sizes investigated. Separate values are reported for each of the four values for the number of predictors in the model and for the separate values of the number of predictors with non-zero correlations with the response. The actual significance levels are aggregated over the three different values of the correlations between the predictor variables and over the two values for the non-zero correlation between the predictors and the response. In every case, the actual significance levels for the Bonferroni-adjusted tests were less than the nominal 0.05 level as expected. In every case except the one with only two predictors and only one of them having a non-zero correlation with the response, the unadjusted significance levels were substantially larger than the nominal 0.05 level. It is clear from the table that as the number of predictors in the model increases, the unadjusted actual significance level also increased, with the increase being smaller with a larger number of the predictors having a non-zero correlation with the response.

For the Bonferroni-adjusted significance levels, in every case with all of the predictors having no correlation with the response, the actual significance level was close to, but slightly smaller than, the nominal 0.05 level. As the number of predictors with a non-zero correlation with the response increases, the Bonferroni-adjusted significance levels get smaller, indicating, as expected, that the Bonferroni adjustment over-corrects so that the observed significance level is less than the nominal value. Although variations exist between the actual significance levels for both the unadjusted and adjusted cases for the

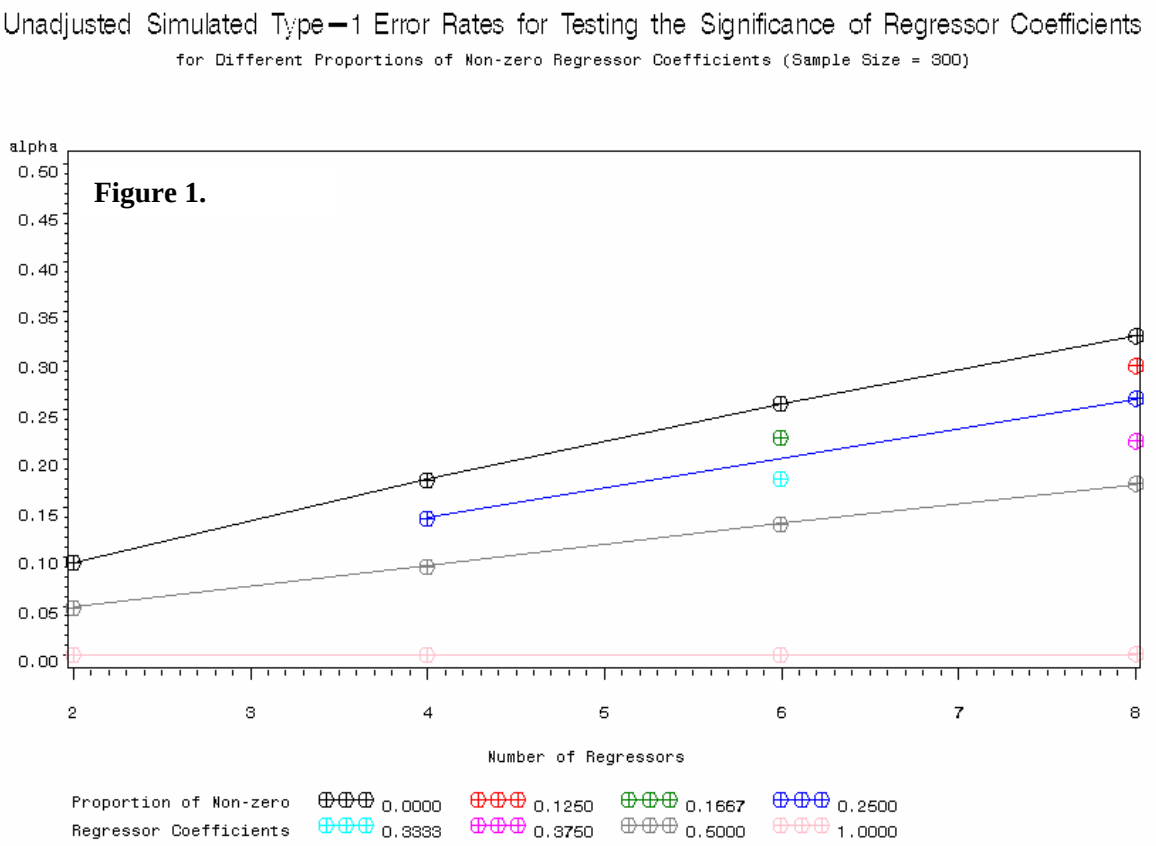

Figure 1.

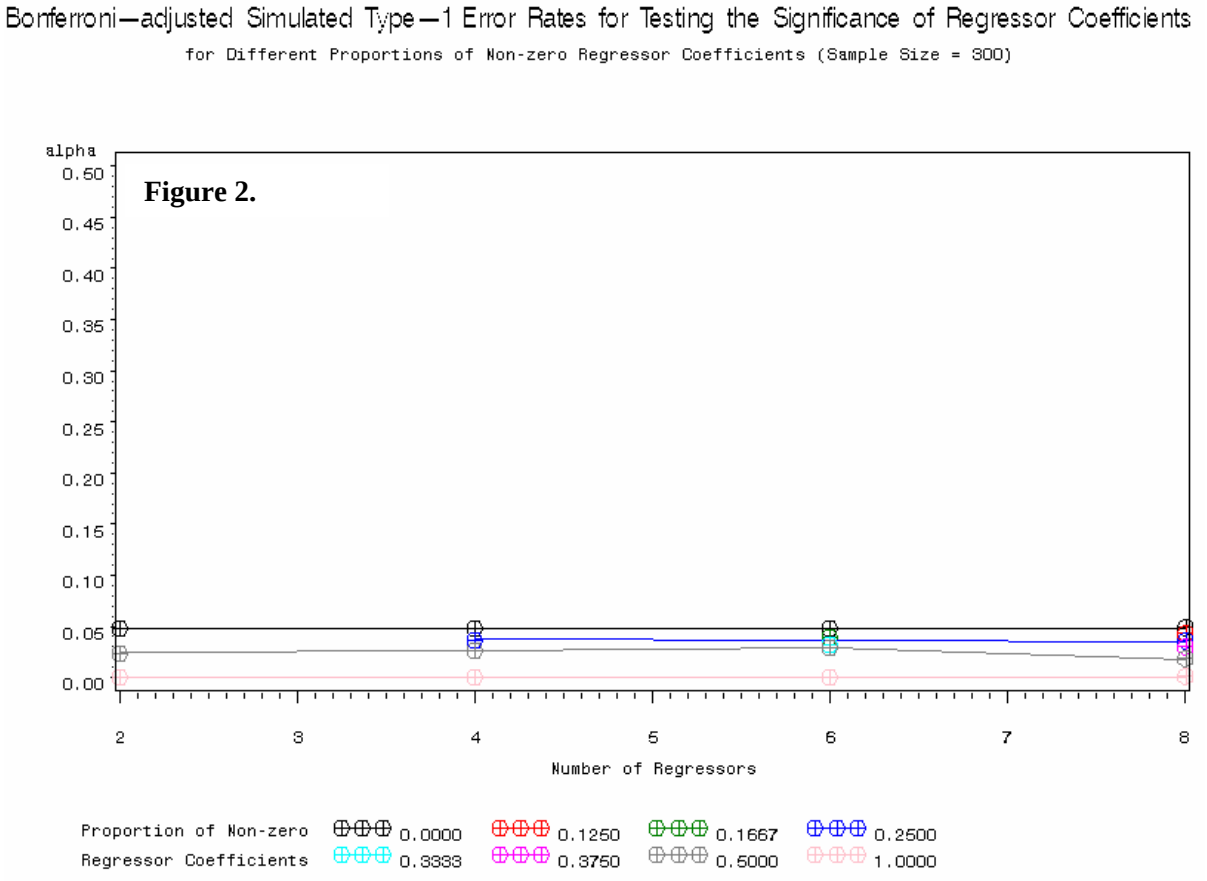

Figure 2.

two sample sizes, these variations are small, indicating that the size of the sample had little meaningful effect on the observed significance levels.

Figures 1 and 2 display the information in Table 1 graphically for the unadjusted actual significance levels and the Bonferroni-adjusted actual significance levels, respectively, for the sample size = 300 cases. Similar graphs for sample size = 50 were nearly identical to those in figures 1 and 2, indicating no differences in these observed significance levels due to sample size. The graphs corresponding to sample size = 50 have been omitted. In both figures, the data are aggregated according to the proportion of predictors having non-zero correlations with the response. Actual data values occurred only at the discrete values of the number of predictors equal to 2, 4, 6, and 8. The lines are drawn to indicate the linear trends that correspond to the observed significance levels for both the adjusted and unadjusted tests as the number of predictors increases and the number of predictors having non-zero correlations with the response decreases.

Conclusion

It is not common practice among applied researchers to use adjusted t-tests for variable selection in regression analysis. Rather, it is much more common to see each individual test conducted at the nominal level, usually using $\alpha = 0.05$. This research indicates that when unadjusted t-tests are used for individual variable selection, the associated overall Type-I error rate may be inflated by as much as 2 to 6 times the nominal α-level depending upon the number of predictors in the model and the number of predictors that have a non-zero correlation with the response. Consequently, one or more variables are identified as "significant" predictors of the response that are not actually needed in the model, i.e., the amount of unique variance in the response explained by these variables is negligible. A more conservative approach, one that controls the overall α-level, would be to use the Bonferroni-adjusted approach to conduct these tests. As shown here, tests based on this adjustment are overly conservative, especially as the number of predictors having non-zero correlations with the response increases. However, if the goal of the research is to identify only those predictors that are actually related to the response, these results seem to indicate that using the Bonferroni adjustment would be preferred.

References

Agresti, A. & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd Ed.). Upper Saddle River, NJ: Prentice Hall.

Foster, D. P. & Stine, R. A. (2004). Variable selection in data mining: Building a predictive model for bankruptcy. *Journal of the American Statistical Association, 99*, 303-313.

Galambos, J. & Simonelli, I. (1996). *Bonferroni-type inequalities with applications*. New York: Springer-Verlag New York, Inc.

Glass, G. V & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd Ed.) Needham Heights, MA: Allyn & Bacon.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1998). *Applied statistics for the behavioral sciences* (4th Ed.). Boston, MS: Houghton Mifflin.

Korn, E. L. & Graubard, B. I. (1990). Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni t statistics. *The American Statistician, 44*, 270-276.

Send correspondence to: Daniel J. Mundfrom
University of Northern Colorado
Email: daniel.mundfrom@unco.edu