Multiple Linear Regression Viewpoints

A Publication sponsored by the American Educational
Research Association's Special Interest Group on
Multiple Linear Regression: The General Linear Model

MLRV

Volume 33 • Number 2 • Spring 2007

Table of Contents

**Achieving Accurate Prediction Models:
Less is Almost Always More** **1**

John D. Morris Florida Atlantic University
Mary G. Lieberman Florida Atlantic University

**Regression Discontinuity:
Examining Model Misspecification** **6**

Randall E. Schumacker University of Alabama

**Is High School Performance and Standardized Test
Scores as Admission Criteria Enough Considering
the Institutional Cost of Misclassification?** **11**

Michael Bronsert University of Northern Colorado
Daniel Mundfrom University of Northern Colorado

**The Impact of Graphing Calculator Use
on Algebra I End of Course Examinations** **16**

Todd Sherron info2knowledge, LLC
Vicki Dimock Southwest Educational Development Laboratory
Rob Foshay Texas Instruments

Coors Field: A Pitchers Graveyard? **22**

Jay Schaffer University of Northern Colorado
Raj Chandran University of Northern Colorado

Multiple Linear Regression Viewpoints

Editorial Board

Multiple Linear Regression Viewpoints (ISSN 0195-7171) is published by the AERA Special Interest Group on Multiple Linear Regression: General Linear Model through the **University of Alabama at Birmingham**.

Subscription and SIG membership information can be obtained from:
Cynthia Campbell, MLR:GLM/SIG Executive Secretary
Department of Educational Technology, Research & Assessment
Northern Illinois University
DeKalb, IL 60115-2854.
ccampbell@niu.edu

MLRV abstracts appear in CIJE, the ERIC system, and microform copies are available from University Microfilms International, 300 North Zeeb Road, Ann Arbor, MI 48106. *MLRV* is listed in the *EBSCO Librarians Handbook.*

Achieving Accurate Prediction Models: Less is Almost Always More

John D. Morris **Mary G. Lieberman**
Florida Atlantic University

Accurate cross-validated prediction accuracy is posited as the ultimate criterion for prediction model performance. This study investigates and demonstrates, across a wide variety of data sets, the nearly ubiquitous benefit to classification model accuracy of optimal subset selection. Unlike popular "stepwise" methods often used (and abused) in the literature, this study considers only all-possible-subset cross-validated performance as the criterion of accuracy. The superiority of variable subsets is demonstrated for predictive discriminant analysis and logistic regression. Computer programs are also made available.

Among the techniques used for solving classification problems, logistic regression (LR) and predictive discriminant analysis (PDA) are two of the most popular (Yarnold, Hart & Soltysik, 1994). Unlike PDA, LR captures the probabilistic distribution embedded in a categorical outcome variable, avoids violations to the assumption of homogeneity of covariance matrices (in the case of the linear PDA model), and does not require strict multivariate normality. Therefore, when PDA assumptions are violated, we might expect greater cross-validated classification accuracy with LR than PDA.

Although several studies have compared the classification accuracy of LR and PDA, the results have been inconsistent. For example, some studies (Baron, 1991; Bayne, Beauchamp, Kane, & McCabe, 1983; Crawley, 1979) suggest that LR is more accurate than PDA for nonnormal data. However, several researchers (e.g., Cleary & Angel, 1984; Knoke, 1982; Krzanowski, 1975; Lieberman & Morris, 2003; Meshbane & Morris, 1996; Press & Wilson, 1978) found little or no difference in the accuracy of the two techniques with PDA often performing better than LR. Part of the reason these results are in dispute is that one may look at accuracy for all groups or separate-groups. As well, one may consider a cross-validated index of accuracy or the accuracy of reclassifying the calibration sample; these studies are not consistent in respect to the criterion of accuracy used. Specifically, examination of cross-validation accuracy in LR studies is uncommon, and when done is usually of the most basic (also non-unique and unstable) sort (hold-out sample). No computer packages support more appropriate resampling cross-validation methods (variously called PRESS, Lachenbruch *U*, leave-one-out, jackknife and bootstrap).

Whichever method (LR or PDA) is selected, one may consider subsets of all possible variables for purposes or parsimony, and/or to increase cross-validation accuracy of the model (Morris & Meshbane, 1995). The most usual method is to consider accuracy in classification of the sample upon which the model is created (internal) with the objective of parsimony. That is, realizing that some accuracy will be lost in reducing the number of predictor variables in classifying the calibration sample, but compromising that loss with the gain in parsimony afforded by the reduction in size of the prediction model. However, as in multiple regression, an increase in cross-validated prediction accuracy (the most appropriate criterion) is almost always available using a model composed of fewer than all variables available. Thus one may gain both parsimony and some degree of explanatory power for the model. In addition, although traditional methods considering the piecemeal change in performance of models in respect to prediction within the calibration sample have often been used (forward, backward, stepwise, or variants thereof), they are neither optimal, nor unique and are now generally in disfavor.

In the case of PDA an examination of the cross-validation accuracy of all 2^p-1 (where p is the number of predictor variables) subsets of variables has been recommended and utilized (Huberty, 1994; Huberty & Olejnik, 2006; Morris & Meshbane, 1995). In this case the method of cross-validation is the leave-one-out method. In the leave-one-out procedure (Huberty, 1994, *p*. 88; Lachenbruch & Mickey, 1968; Mosteller & Tukey, 1968) a subject is classified by applying the rule derived from all subjects except the one being classified. This process is repeated round-robin for each subject, with a count of the overall classification accuracy used to estimate the cross-validated accuracy. [Clearly the same round-robin procedure can be used to estimate either relative or absolute accuracy in the use of multiple regression and has appeared in that context, with perhaps the earliest reference due to Gollob (1967). In a system intended to select optimal multiple regression predictor variable subsets, Allen (1971) coined the procedure PRESS, and he appears to be the source most often cited in the multiple regression literature.]

Table 1. Data set, number of predictor variables (p), PDA hit-rate for all p variables, number of variables in the best performing subset(s), hit-rate for that subset, and the % change in hit-rate.

#	Data Set Source	p	Hit-rate for p Predictors	# Predictors in Best Subset(s)	Max Hit-rate	% Change
1	Rulon Grps 1 & 2	4	.809	3	**.831**[a]	2.72
2	Rulon Grps 1 & 3	4	.927	3	.934	.75
3	Rulon Gps 2 & 3	4	.830	3	**.836**	.72
4	Block - Grps 1 & 2	4	.679	2	**.743**	9.43
5	Block - Grps 1 & 3	4	.646	4	**.646**	0.00
6	Block - Grps 1 & 4	4	.603	1	**.667**	10.61
7	Block - Grps 2 & 3	4	.553	1	**.632**	14.29
8	Block - Grps 2 & 4	4	.600	2	**.640**	6.67
9	Block - Grps 3 & 4	4	.684	3	**.711**	3.95
10	Demographics	8	.581	3,6	.613	5.51
11	Dropout from 4th	10	.702	2,3,4,5	.787	12.11
12	Dropout from 8th	11	.739	5,6	**.803**	8.66
13	Fitness	10	.588	7	.616	4.76
14	Warncke -Grps 1 & 2	10	.482	2	**.607**	25.93
15	Warncke -Grps 1 & 3	10	.571	3,5,6	.657	15.06
16	Warncke -Grps 2 & 3	10	.402	1,5	**.575**	43.03
17	Bisbey 1& 2	13	.888	5,6,7,8,9,10	.914	2.93
18	Bisbey 2& 3	13	.839	2,4,5,6	**.983**	17.16
19	Talent - Grps 1 & 3	14	.578	7	.698	20.76
20	Talent - Grps 3 & 5	14	.772	8.9	.835	8.16
21	Talent - Grps 1 & 5	14	.746	5,7,8	.797	6.84

[a] Bold when > than LR.

In the case of PDA (and regression) a matrix identity due to Bartlett (1951) allows the task of the requisite N-1 matrix inversions to be accomplished with far less computational labor that would otherwise be necessary. However, this mathematical tool is irrelevant to the iterative method of LR optimization, thus N-1 LR optimizations must be completed for each of 2^p-1 subsets of predictor variables.

Unlike most LR studies that consider calibration sample statistics as the criterion for model fit (e.g., the Cox & Snell, or Nagelkerke R^2), the criterion for model accuracy is construed in this study, as is typically the case in PDA, as classification accuracy. That is, the proportion of correct leave-one-out cross-validated classifications (hit-rate) for the total sample and each separate group. Thus for a two-group problem, we may order the accuracy of our 2^p-1 candidate LR equations according to three different (total sample and each group) cross-validated classification accuracy criteria.

Method

Analyses from 21 two-group classification problems from Morris and Huberty (1987) were used to illustrate the method and computer program for PDA (Table 1) and LR (Table 2). Although not purported to represent all potential data structures, these data sets have been used in several classification studies as representing a wide variety of number of predictor variables, group separation, and covariance structures. As the number of predictors ranges from 4 to 14, the candidate 2^p-1 cross-validated subsets range from a very modest 15 to 16,383 for the 14 predictor variable problem. However, even in the case of the calculation and sorting of the16K+ cross-validated classification performances, the program executes (on a midrange laptop) in less than 30 seconds.

Result and Conclusions

In the case of both PDA (Table1) and LR (Table 2), one can see that, in all cases, except #5 in PDA, selection of the best performing subset (of the 2^p–1 possibilities) offers a reduction in the number of predictor variables, often by more than half, thus parsimony is well served. One may also note that,

Table 2. Data set, number of predictor variables (p), LR hit-rate for all p variables, number of variables in the best performing subset(s), hit-rate for that subset, and the % change in hit-rate.

#	Data Set Source	p	Hit-rate for p Predictors	# Predictors in Best Subset(s)	Max Hit-rate	% Change
1	Rulon Grps 1 & 2	4	0.803	3	.815	1.49
2	Rulon Grps 1 & 3	4	0.914	3	.934	2.19
	Rulon Gps 2 & 3	4	0.824	3	.830	0.73
4	Block - Grps 1 & 2	4	0.692	1,2	.718	3.76
5	Block - Grps 1 & 3	4	0.620	3,4	.620	0.00
6	Block - Grps 1 & 4	4	0.577	1,2	.628	8.84
7	Block - Grps 2 & 3	4	0.566	1,2	.605	6.89
8	Block - Grps 2 & 4	4	0.587	2	.627	6.81
9	Block - Grps 3 & 4	4	0.684	3	.697	1.90
10	Demographics	8	0.591	4	**.620**[a]	4.91
11	Dropout from 4th	10	0.660	4	.787	19.24
12	Dropout from 8th	11	0.725	3	.782	7.86
13	Fitness	10	0.591	4	**.620**	4.91
14	Warncke -Grps 1 & 2	10	0.446	1	.580	30.04
15	Warncke -Grps 1 & 3	10	0.600	4	**.667**	11.17
16	Warncke -Grps 2 & 3	10	0.425	2	.563	32.47
17	Bisbey 1& 2	13	0.879	6,7,8,9,10	.914	3.98
18	Bisbey 2& 3	13	0.856	5,6,7	.924	7.94
19	Talent - Grps 1 & 3	14	0.621	5	**.733**	18.04
20	Talent - Grps 3 & 5	14	0.787	6,7,8,9	**.858**	9.02
21	Talent - Grps 1 & 5	14	0.740	5	.797	7.70

[a] Bold when > PDA.

particularly with larger models, multiple sets of predictors and size models often achieve maximum accuracy. In addition, one can see that due to the reduction in the number of predictor variables, cross-validation accuracy increased from less than 1% all the way up to more than 40%. Only in data set #5 (PDA & LR) did the reduced model perform the same as the full model. In the case of LR, still offering the same accuracy, but with increased parsimony, and in the case of PDA, offering no advantage. The mean increase in cross-validated hit-rate due to the reduction in the number of predictor variables over all 21 data sets was about 5% for LR and 10% for PDA. Thus one can have parsimony and increased accuracy. Through this procedure and computer programs, researchers will be able to make better decisions about optimally accurate classification model construction.

Although not the focus of this study, it is difficult to ignore potential comparisons between PDA an LR performance. As was stated, greater parsimony and accuracy is afforded in almost every case by selecting an optimally performing subset. As has been previously documented, cross-validation performance was often very close between PDA and LR. However, if one considers only the optimally performing subsets the advantage seems to go to PDA herein. PDA is best in 12 data sets, LR in 5, and performance is the same in 4.

Further consideration of the advantage of the availability of multiple optimally performing subsets should also be noted. Missing data is almost always a difficulty in dealing with real data. First, a model depending on a smaller number of variables has not only the philosophical advantage of parsimony, but may also afford the opportunity to accommodate missing data; there is more opportunity for the model to be applicable as the number of variables decreases. Moreover, if several equally performing (or nearly so) superior subsets are available, the opportunity to accommodate the missing data of an individual score vector is increased; one can use alternative models for alternate missing data configurations, unless, of course, that variable that is missing is including in all of the best subsets. Table 3 illustrates the top 20 (of 256 possibilities) subset accuracies for PDA prediction of dropout from high school from 8 predictors. One can see that the optimal subset contains 4 variables, but many subsets are close, such that performance is maintained. Such information can aid in handling missing data. That is, one might argue that if the four variables in the best performing model are available for a subject (SCHOOLS8, MATH8,

Table 3. Ranked 20 best (of 255) performing subsets, and total model.

	Variables Included in the Model:							
Hit-Rate	SCHOOLS8	REPEATS8	READING8	MATH8	LANG8	SCIENCE8	SOCST8	DSFS8
0.753	√			√		√		√
0.747	√			√				√
0.747	√							√
0.747	√				√	√	√	√
0.747	√			√	√	√	√	√
0.741	√		√		√		√	√
0.741	√			√		√	√	√
0.735	√		√		√	√		√
0.735	√			√			√	√
0.735	√	√	√				√	
0.735	√			√	√		√	√
0.735	√					√	√	√
0.735	√		√	√	√			√
0.735	√	√						√
0.735	√	√	√					
0.728	√	√				√	√	
0.728	√					√		√
0.728	√	√				√		√
0.728	√	√	√			√		√
0.728	√	√				√		
Total Model								
0.679	√	√	√	√	√	√	√	√

Note: SCHOOLS8: Accumulated # of schools attended by grade 8.
REPEATS8: Accumulated # of Grades repeated by grade 8.
READING8: 8th Grade Reading grade.
MATH8: 8th Grade Math grade.
LANG8: 8th Grade Language grade.
SCIENCE8: 8th Grade Science grade.
SOCST8: 8th Grade Social Studies grade.
DSFS8: Accumulated # of D and F grades over all subjects by grade 8.

SCIENCE8, DSFS8) it should be used. However, if for some reason MATH8 and SCIENCE8 (as well as REPEATS8, READING8, LANG8, and SOCST8) are missing from a case, then a model that demonstrates essentially the same performance is available using only SCHOOLS8 and DSFS8. In this case, the SCHOOLS8 is the number of schools the child had attended by the 8th grade and DSFS8 is the number of “D” and “F” grades the child had accumulated, whereas MATH8 and SCIENCE8 are grades in those specific subjects in the 8th grade. So, as an example, for a teacher, or school not reporting subject grades, but the more “global” accumulated variables of SCHOOLS8 and DSFS8 are retained in the county database, the alternate model could be used with the expectation of attaining essentially the same accuracy.

Note, however, in this case, that if SCHOOLS8 is not available, optimal accuracy appears improbable. It is a “don’t leave home without it” variable. The programs used herein (one for PDA and one for LR) for the examination and ordering of the 2^p–1 possible subsets of predictor variables by their leave-one-out accuracy are available from the senior author at: jdmorris@fau.edu. They are available as Intel based EXE files (compiled from FORTRAN).

References

Allen, D. A. (1971). *The prediction sum of squares as a criterion for selecting predictor variables* (Tech. Rep. No. 23). Lexington: University of Kentucky, Department of Statistics.

Baron, A. E. (1991). Misclassification among methods used for multiple group discrimination – The effects of distributional properties. *Statistics in Medicine*, *10*, 757-766.

Bartlett, M. S. (1951). An inverse matrix adjustment arising in discriminant analysis. *Annals of Mathematical Statistics*, *22*, 107-111.

Bayne, C. K., Beauchamp, J. J., Kane, V. E., and McCabe, G. P. (1983). Assessment of Fisher and logistic linear and quadratic discrimination models. *Computational Statistics and Data Analysis*, *1*, 257-273.

Cleary, P. D. & Angel, R. (1984). The analysis of relationships involving dichotomous dependent variables. *Journal of Health and Social Behavior*, *25*, 334-348.

Crawley, D. R. (1979). Logistic discriminant analysis as an alternative to Fisher's linear discriminant function. *New Zealand Statistics*, *14*(*2*), 21-25.

Gollub, H. F. (1967, September). *Cross-validation using samples of size one*. Paper presented at the annual meeting of the American Psychological Association, Washington, D. C.

Huberty, C. J., & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis*. New York: Wiley.

Huberty, C. J. (1994). *Applied discriminant analysis*. New York: Wiley.

Knoke, J. D. (1982). Discriminant analysis with discrete and continuous variables. *Biometrics*, *38*, 191-200.

Krzanowski, W. J. (1975). Discrimination and classification using both binary and continuous variables. *Journal of the American Statistical Association*, *70*, 782-790.

Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, *10*, 1-11.

Lieberman, M. L, & Morris, J. D. (2003, April). *Comparing classification accuracies between predictive discriminant analysis and logistic regression in specific data sets*. Paper presented at the meeting of the American Educational Research Association, Chicago.

Meshbane, A., & Morris, J. D. (1996, April). *Predictive discriminant analysis versus logistic regression in two-group classification problems*. Paper presented at he meeting of the American Educational Research Association, New York.

Morris, J. D., & Huberty, C. J (1987). Selecting a two-group classification weighting algorithm. *Multivariate Behavioral Research*, *22*, 211-232.

Morris, J. D., & Meshbane, A. (1995). Selecting predictor variables in two-group classification problems. *Educational and Psychological Measurement*, *55*, 438-441.

Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, *73*, 699-705.

Mosteller, F. & Tukey, J. W. (1968). Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.). *Handbook of social psychology* (Vol. 2, pp. 80-203). Reading, MA: Addison-Wesley.

Yarnold, P. R., Hart, L. A. & Soltysik, R. C. (1994). Optimizing the classification performance of logistic regression and Fisher's discriminant analysis. *Educational and Psychological Measurement*, *54*, 73-78.

Send correspondence to: John D. Morris
Florida Atlantic University
Email: jdmorris@fau.edu

Regression Discontinuity: Examining Model Misspecification

Randall E. Schumacker
University of Alabama

The Regression Discontinuity (RD) design looks similar to the non-equivalent group design, which uses analysis of covariance, but assumptions and advantages are much different. The major problem in analyzing data from the RD design is model misspecification. If the regression equation or statistical model does not reflect the data distribution, then biased estimates of the treatment effect will occur. For example, if the true pre-post relationship is curvilinear, but the regression equation only modeled linear regression effects, the treatment effects would be biased. However, a statistical approach is possible using a full model with all terms specified and then test restricted sub-models that omit individual parameters.

The basic RD Design is a two-group pretest-posttest model and is depicted as follows:

```
C  O  X  O
C  O     O
```

The RD design looks similar to the non-equivalent group design, which uses analysis of covariance, but assumptions and advantages are much different (Campbell, 1989; Loftin & Madison, 1991; Schumacker, 1992). The RD design does not have subject selection bias (pre-defined group membership) rather uses a pre-test measure to assign treatment or non-treatment status. The basic RD model would have an intercept term, pre-test measure, and dummy-coded group assignment variable regressed on a post-test measure. The pre-test measure does not have to be the same as the post-test measure.

The major problem in analyzing data from the RD design is model misspecification. If the regression equation or statistical model does not reflect the data distribution, then biased estimates of the treatment effect will occur. For example, if the true pre-post relationship is curvilinear, but the regression equation only modeled linear regression effects, the treatment effects would be biased. Consequently, it is a good idea to visually inspect the pre-post scatter plot to see what type of relationship exists. However, a statistical approach is possible using a full model with all terms specified and then test restricted sub-models that omit individual parameters. This will be illustrated in this paper.

There are five central assumptions when performing an RD analysis. A major concern is the model specification in the pre-post distribution being a polynomial function rather than a logarithmic or exponential function. The five central assumptions are:

1. The cutoff value must be absolute without exception. A subject selection bias is introduced and the treatment effect is biased if incorrect assignment to groups based on the cutoff value occurred (unless it is known to be random).
2. The pre-post distribution is a polynomial function. If the pre-post relationship is logarithmic, exponential or some other function, the model is misspecified and the treatment effect is biased. The data can be transformed to create a polynomial distribution prior to analysis to yield appropriate model specification.
3. There must be a sufficient number of pretest values in the comparison group to estimate the pre-post regression line.
4. The experimental and comparison groups must be formed from a single continuous pretest distribution with the division between groups determined by the cutoff value.
5. The treatment or program intervention must be delivered to all subjects, i.e., all receive the same reading program, amount of training, etc.

Model specification can be identified in three different ways or types: exactly specified, over specified, and under specified RD models. An exactly specified model has an equation that fits the “true” data. So if the “true” data is linear then a simple straight-line pre-post relationship with a treatment effect would yield unbiased treatment effects. The RD equation would include a term for the posttest Y, the pretest X, and the dummy-coded treatment variable Z with no unnecessary terms. When we exactly specify the true model, we get unbiased and efficient estimates of the treatment effect. If the RD equation is over specified it includes additional parameter estimates that are not required, i.e. interaction or curvilinear coefficients, and treatment effect would be inefficient. If the RD equation is under specified it leaves out important parameter estimates and the treatment effect would be biased.

The basic steps being proposed to statistical test the type of model when conducting an RD analyses would be as follows:

1. Subtract the cut-off score from the pretest score ($X_{\text{pre}} – X_{\text{cut}}$).
2. Visually examine the pre-post scatter plot for type of data relationship.
3. Determine if any higher-order polynomial terms or interactions are present.
4. Estimate the “full” RD regression equation.
5. Modify the RD equation by dropping individual non-significant terms.

Methodology

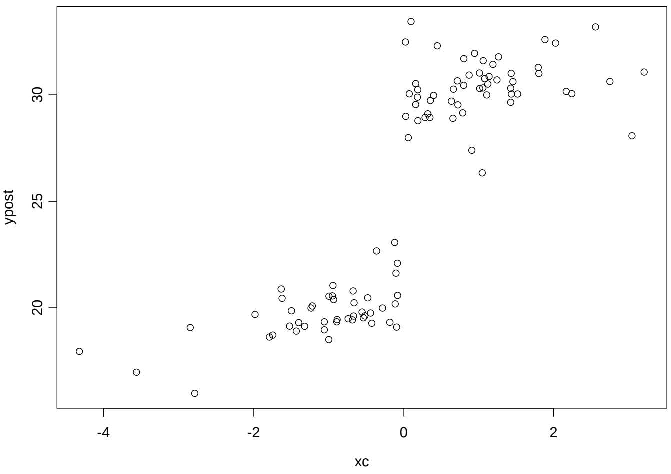
The “full” RD regression equation with subsequent “modified” or “restricted” regression models permit one to statistically determine the best fitting model for estimating treatment effects. A “full” regression discontinuity model could be as outlined below.

$$y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + \beta_4 X_i^2 + \beta_5 X_i^2 Z_i + e_i$$

The RD regression equation terms are defined as:

y_i = post test score outcome for ith subject
β_0 = regression coefficient for intercept
β_1 = linear pre test regression coefficient
β_2 = mean post test different for treatment group
β_3 = linear interaction regression coefficient between pre and group
β_4 = quadratic regression coefficient for pretest
β_5 = quadratic interaction regression coefficient for pre test and group
X_i = transformed pre test score for ith subject
Z_i = group assignment based on cut off score (0 = comparison, 1 = treatment)
e_i = residual score for ith subject.

Figure 1. Simulated Regression Discontinuity data



Data Simulation

The S-PLUS program that generated the simulated data and computed results for the RD analysis is footnoted (SPLUS, 2005). The *rnorm* function in S-PLUS generated 100 random normal data points (Chambers, Mallows, & Stuck, 1976). The post test scores (*Y*) and pre test scores (*X*) were created by adding residual error (*ey* or *ex*) to this random normal variable (*true*). Group assignment (*Z*) was determined based on subtracting a cut score of 20 from the pre test score (1–treatment, 0–comparison). This 10 point treatment gain was added to the post test score (*Y*). Optional *print* and *write* statements are included to either view or save the data in a file.

The least squares regression function, *lm*, was used to run the RD analyses where ypost = post test score; xc = transformed pre test score; z = group assignment; xz = linear interaction; xsq = quadratic pre test; and xsqz = quadratic interaction of pre test and group. The sequence of RD regression equations that were tested are as follows:

1. Full model: lm (y_{post} ~ xc + z + xz + xsq + xsqz)
2. No quadratic Interaction: lm (y_{post} ~ xc + z + xz + xsq)
3. No quadratic Interaction: lm (y_{post} ~ xc + z + xz)
4. Linear model: lm (y_{post} ~ xc + z)
5. No pre test model: lm (y_{post} ~ xc)

A visual inspection of the simulated data in Figure 1 indicates that we would expect the best fitting RD model to be the linear model. The scatter plot displays the y_{post} (post test scores) and xc (transformed pre test scores) variables. A ten point treatment effect is visible between the two groups. Recall that the treatment group had a mean of 30 and the comparison group had a mean of 20, which are visually present in the scatter plot.

Results

The full model results indicated that all regression coefficients were non-significant. The model misspecification (*over specified*) further indicated an inefficient treatment effect (*z* = -87.86), which we know is not true given the simulated data.

***RD Full Model* (F=384.3, df = 5, 95)**

```
Coefficients:
               Value Std. Error   t value  Pr(>|t|)
(Intercept)   -9.7719   57.9279    -0.1687    0.8664
         xc   -1.9094    5.2984    -0.3604    0.7194
          z  -87.8617  108.9808    -0.8062    0.4222
         xz    9.9789   10.5763     0.9435    0.3478
        xsq    0.0766    0.1455     0.5261    0.6001
      xsqz   -0.2570    0.2588    -0.9932    0.3232
```

The RD restricted model that dropped the quadratic interaction between squared pre test and group is still *over specified* because all of the regression coefficients were non-significant. The treatment effect was inefficient and over estimated at 19.15 (Z) compared to the known treatment effect of 10 points.

***RD – drop quadratic interaction of pre test and group* (F = 480.2, df = 4, 95)**

```
Coefficients:
              Value Std. Error  t value Pr(>|t|)
(Intercept)  22.5800  47.8979     0.4714   0.6384
         xc   1.0476   4.3824     0.2390   0.8116
          z  19.1494  16.3454     1.1715   0.2443
         xz  -0.4933   0.8209    -0.6010   0.5493
        xsq  -0.0047   0.1203    -0.0392   0.9688
```

The RD model with both quadratic terms removed is still *over specified* and yielded a larger *F* value, however, the linear interaction (*xz*) between pre test and group was not statistically significant (*t* = -1.81; p = .07). The treatment effect was also inefficient and higher than the known true treatment effect value.

***RD – drop both quadratic interaction effects* (F = 646.9, df = 3, 96)**

```
Coefficients:
              Value Std. Error  t value Pr(>|t|)
(Intercept)  20.7031   0.2793    74.1277   0.0000
         xc   0.8761   0.1991     4.4010   0.0000
          z  19.7477   5.8067     3.4008   0.0010
         xz  -0.5234   0.2891    -1.8104   0.0734
```

The RD model with all interaction terms removed is an *exactly specified* model. This RD analysis modeled the "true" nature of the linear relationship between pre and post scores and yielded an intercept value of 20, which is close to the comparison group mean and a treatment effect of 9.26, which is close to the known treatment effect of 10 points, given the introduction of random error.

***RD – drop linear interaction* (F = 946.5, df = 2, 97)**

```
Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 20.4362  0.2400    85.1603  0.0000
         xc  0.6279  0.1460     4.2995  0.0000
          z  9.2592  0.3945    23.4706  0.0000
```

The RD model without the pre test score term removed is an *under specified* model. This RD analysis yielded a biased treatment effect that overestimated the “true” effect of 10 points. The F value is inflated and a key variable, the pre test score was omitted. Recall that *under specified* models leave out important variables, hence affect the model validity.

RD No pre test Model (F = 1591, df = 1, 98)

```
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  19.7607   0.1969   100.3519   0.0000
          z  10.5900   0.2655    39.8842   0.0000
```

Conclusions

The RD design was one of three designs approved for program evaluation by the Department of Education many decades ago, yet the technique is not widely used (Thistlethwaite & Campbell, 1960; McNeil, 1984; Trochim, 1984). The regression discontinuity design uses a least-squares equation to yield an intercept (baseline measure) and regression weight (treatment effect measure) in assessing program effectiveness. A positive or negative regression weight determines gain or loss due to treatment or intervention effect, which is also tested for statistical significance. However, if the regression model is misspecified then treatment effects are inefficient and biased estimates.

RD is a powerful alternative to using quasi-experimental designs with distinct advantages. Regression discontinuity has fewer assumptions in comparison to not meeting assumptions in quasi-experimental designs that use analysis of covariance, i.e., random sampling; normality of treatment levels; homogeneity of variance; independence of variance estimates; linear regression assumption; and homogeneity of regression lines. The analysis of covariance assumptions are seldom met, thus leading to erroneous interpretations of treatment effects (Campbell, 1989; Loftin & Madison, 1991).

The RD normal distribution assumption is not problematic and can be handled by robust regression methods or probit data transformation. The cut-off score misspecification is usually not a problem because state agencies or school districts mandate a cut-off score for high-stakes testing. The model misspecification can also be examined by including linear, polynomial, and interaction terms in the RD equation and then dropping non-significant terms. Other advantages include RD designs being able to explore treatment effect differences at different cutoff points, use different pre-test measures than post-test measures, do not require matching of subjects, and can use multiple comparison groups with different cutoff scores.

Educational researchers should therefore make increased use of the regression-discontinuity technique for program evaluation because you can use a different pre-test measure for the cut-off value, use different regression models that reflect the distribution of the data (linear, curvilinear, and interaction), and do not have to meet all of the assumptions in ANCOVA to yield stable estimates of treatment effects.

References

Chambers, J. M., Mallows, C. L. and Stuck, B. W. (1976). A Method for Simulating Stable Random Variables. *Journal of the American Statistical Association*, *71*, 340-344.

Campbell, K.T. (1989). *Dangers in using analysis of covariance procedures*. ERIC Document # ED 312298 (http://eric.ed.gov).

Loftin, L., & Madison, S. (1991). The extreme dangers of covariance corrections. In B. Thompson (Ed.), (1991). *Advances in educational research: Substantive findings, methodological developments* (Vol. 1, pp. 133-148). Greenwich, CT: JAI Press. (International Book Sellers Number: 1-55938-316-X)

McNeil, Keith (April, 1984). *Random Thoughts on Why the Regression Discontinuity Design Is Not Widely Used*. Paper presented at the American Educational Research Association Annual Meeting. New Orleans, LA.

Schumacker, Randall E. (April, 1992). *Factors Affecting Regression-Discontinuity*. Paper presented at the American Educational Research Association Annual Meeting. San Francisco, CA.

S-PLUS (2005). *S-PLUS 6 User’s Guide for Windows*. Insightful, Inc., Seattle, WA.

Thistlethwaite, D.L. & Campbell, D.T. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Educational Psychology*, *51*(*6*), 309-317.

Trochim, William M. K. (1984). *Research Design for Program Evaluation, the Regression Discontinuity Approach*. Sage Publications: Beverly Hills, CA.

Send correspondence to: Randall E. Schumacker
University of Alabama
Email: rschumacker@ua.edu

Is High School Performance and Standardized Test Scores as Admission Criteria Enough Considering the Institutional Cost of Misclassification?

Michael Bronsert **Daniel Mundfrom**
University of Northern Colorado

High school performance and aptitude test scores have been shown to have a marginal relationship with common measures of undergraduate student academic success. This minor relationship suggests a source of admission errors which could contribute to tuition revenue loss. This study's objective was to answer the following questions: (1) can discriminant functions be constructed that can correctly classify students as individuals obtaining a degree within a reasonable amount of time or individuals that withdrawal early, (2) how efficient are these discriminant functions, (3) do they differ by gender, and (4) what is the estimated institutional cost of misclassifying students. Results indicated that discriminant functions could be developed that correctly classified approximately 60% of students. These discriminant functions were also shown to have similar success rates for males and females. Finally, the estimated institutional cost of misclassifying students along with error rates of occurrence suggest a source of tuition revenue loss and that improved predictability of student potential for academic retention is needed.

The problem of undergraduate student attrition is an important economic issue due to loss of tuition revenue, cost of recruitment, and government grants tied to institutional performance (Simpson, 2005). This loss of potential revenue coupled with tighter school budgets and expanding pressures to do more with less places increased demand for accurate assessment of incoming students' potential (Murray, 1997). The information primarily utilized to make admission decisions are high school performance and standardized test scores which have been shown to be related to academic success (e.g., Aleamoni & Obler, 1978; Eimers & Pike, 1997; Mathiasen, 1984; Mouw & Khanna, 1993; Neely, 1977; Noble & Sawyer, 1987, 1997; Pike, 1991; Stumpf & Stanely, 2002). However, this relationship is marginal at best and usually requires the addition of other variables to improve the predictability of students' success to adequate levels. This minor relationship between high school performance and college success suggests a source of admission errors which could ultimately result in increased attrition rates and corresponding decreases in revenue.

The primary purpose of this study is to use admission data, more specifically high school performance and standardized test scores, to predict college success as defined by the attainment of a degree within a reasonable amount of time and to evaluate the institutional cost of misclassifying students. This study's objective was to answer the following questions: 1) can discriminant functions constructed from high school grade point average, high school rank, and American College Testing scores be developed that can correctly classify students into "Graduated" or "Dropped Out" categories, 2) how efficient are these discriminant functions in correctly classifying students into the two categories, 3) do these discriminant functions differ for male and female students and 4) what is the estimated institutional cost of misclassifying students into the wrong category.

Method

Subjects

Subjects of the study were limited to students who had enrolled at the University of Northern Colorado between the fall semester of 1998 and the fall semester of 2005. Individuals that were enrolled for three or less years and were not enrolled in the fall of 2005 were classified as "DROP" students. While individuals that obtain a degree within six years of matriculation were classified as "GRAD" students. All other individuals were considered in an academic transition state, i.e. each student will eventually either enter the GRAD or DROP category. Therefore, this category of students was not included in the analyses since their true membership has not yet been revealed. Complete data were found for 9,892 undergraduate students with 3,085 students belonging to the GRAD category split between 1,086 male and 1,999 female students. For the DROP category complete data were found for 6,807 students with 2,790 males and 4,017 females.

Procedure

Data were obtained from admission applications and included gender, high school grade point average (GPA), high school rank (RANK), American College Testing composite scores (ACT), years in which they enrolled, and whether or not they received a degree. Students lacking ACT test scores had their Scholastic Aptitude Test (SAT) scores converted into equivalent ACT test scores using the concordance table developed by Dorans, Lyu, Pommerich, and Houston (1997).

Three discriminant analyses were conducted, one for each of the following groups: total group of students, male students, and female students. In each case, quadratic discriminant functions were developed through the DISCRIM procedure of SAS® with priors equal and all individuals were subsequently classified using the Jackknife method into one of the two categories. All discriminant functions were developed using GPA, RANK and ACT test scores as discriminant variables and were considered statistically significant at an alpha level of .05 or less.

Result and Conclusions

An analysis of the data indicated that statistically significant discriminant functions could be developed for Males [$F_{(3,\ 3872)} = 57.06$, $p < 0.0001$], Females [$F_{(3,\ 6012)} = 77.18$, $p < 0.0001$], and the Total Group [$F_{(3,\ 9888)} = 143.66$, $p < 0.0001$].

Total Group of Students

The discriminant function developed using all of the students were able to correctly classify into the appropriate category 2038 of the GRAD students at 66% accuracy and 3528 of the DROP students at 52% accuracy (Table 1). The total probability of misclassifying a student was .41 with subcategory probability error rates of .17 for GRAD students and .24 for DROP students (Table 2).

Male Students

The discriminant function developed for male students correctly classified 679 of the GRAD students at a success rate of 63% and 1589 of the DROP students at a success rate of 57% (Table 1). This function resulted in a total probability of misclassifying a student of .41 with probability error rates of .19 for GRAD students and .22 for DROP students (Table 2).

Female Students

The discriminant function developed for female students correctly classified 1321 of the GRAD students at 66% accuracy and 1993 of the DROP students at only a 50% success rate (Table 2). This function resulted in a total probability of misclassifying a student of .42 with a probability error rate of .17 for students belonging to GRAD category and .25 for those in the DROP category (Table 3).

Estimating Cost of Misclassification

The cost of misclassifying a particular student would depend on the type of admission decision error, the amount of time until that student either dropped out of school or would have obtained a degree, and the classification of the subsequent student that potentially is being displaced, i.e., whether or not the misclassified student displaces the acceptance of another DROP or GRAD student. The assumption of the latter criteria is that there are a finite number of available admission seats and that the acceptance of a particular student displaces one admission seat available to subsequent students.

University of Northern Colorado Case Study

With tuition revenues for the fiscal year of 2004-2005 being approximately $34.6 million for undergraduate students (UNC A, 2007) and a total undergraduate fall enrollment of 11,014 students (UNC B, 2007), the estimated annual tuition revenue per undergraduate student at the University of Northern Colorado was $3,141 (in 2004 dollars). The misclassification of a GRAD student into the DROP category would result in the student not being accepted into the institution despite the fact that that student would have persisted until graduation and consequently there would be a loss of tuition revenue from that student. The amount of this revenue loss would be determined by how long it would have taken that student to graduate. Approximately 86% of the undergraduate students that graduate from the University of Northern Colorado did so within four to five years of enrollment giving an

Table 1. Correct Classification by Discriminant Analysis in Each Group

	GRAD	DROP
Total Group	2038 (66%)	3528 (52%)
Male	679 (63%)	1589 (57%)
Female	1312 (66%)	1993 (50%)

Table 2. Error Rates and Total Probability of Misclassifying Students

	GRAD[a]	DROP[b]	Total[c]
Total Group	.17	.24	.41
Male	.19	.22	.41
Female	.17	.25	.42

[a] Probability of misclassifying GRAD student into DROP
[b] Probability of misclassifying DROP student into GRAD
[c] Total probability of misclassifying a student.

average academic career of 4½ years (UNC C, 2007) and an average loss of revenue to the institution of misclassifying a GRAD student as a DROP student of $14,135 per student (4½ years x $3,141 tuition revenue per year). Consequently, since the misclassified student was not granted acceptance, admission seat displacement issues are not relevant in this scenario and would not contribute to institutional loss.

The misclassification of a DROP student into the GRAD category would result in the student being accepted into the institution despite his or her future withdrawal and consequently the student would pay tuition as long as they were enrolled. However, the student would eventually withdraw preventing payment of future tuition revenues to the institution. The amount of future revenue loss would depend on the classification, i.e., DROP or GRAD, of the student that was displaced from being accepted following the initial admission decision error and the amount of subsequent years that the student would have enrolled had he or she been accepted. If the misclassified student were to displace the acceptance of a DROP student then the institution ultimately would not incur a loss from the misclassification since that student would have displaced the acceptance of another student who would have withdrawn early as well. However, if the misclassified student displaces the acceptance of a GRAD student the loss to the institution would be the amount of future tuition revenue lost once the student withdraws that would have been paid had the GRAD student not been displaced. Approximately 84% of undergraduate students that eventually withdrew from the University of Northern Colorado did so within the first two years following enrollment suggesting an attrition average of 1½ years (UNC C, 2007). Given an academic career length of 4½ years (from above) and an attrition period of 1½ years, the institution on average would lose out on three years of future tuition revenue following the withdraw of a student and an estimated cost of misclassifying a DROP student as a GRAD student would be approximately $9,423 per student (3 years x $3,141 tuition revenue per year) when a GRAD student is being displaced.

Discussion

The results demonstrated that the use of high school performance and college aptitude test scores can be used to develop discriminant functions that correctly classify students as degree receiving or early withdraw individuals. Overall, the discriminant functions were slightly better at correctly classifying students that belonged to the category that receives a degree which on the surface seems optimistic. However, this ability to correctly classify students is only marginally better than guessing in most cases and no better than guessing in one particular case, i.e., female individuals that withdrew early. The total probability of misclassifying students further support that high school performance and college aptitude test scores can be used to classify approximately 60% of students correctly. But once again, these error rates of correctly classifying students are only marginally better than guessing at 50%. Therefore, the percentages of correctly classified students and the rate of errors in classifying those students support the need for other discriminant variables to improve predictability of students' potential for academic success. Such improvement in predictability of students' academic success would be important especially considering that many institutions automatically accept and reject individuals based on composite scores made up of high school performance and aptitude test scores. Finally, the percentage of correctly classified students and error rates of misclassification were essentially the same for male and females along with the total group of students. The similar rate of errors and percentage correctly classified for all three groups suggests that gender is not essential in determining whether a student will eventually graduate or withdraw. However, other demographic or academic information might reveal differences in admission errors between individuals that graduate with individuals that eventually withdrew.

When considering the cost of making an incorrect admission decision the institution would incur the greatest revenue loss following the misclassification of an individual that would have persisted until graduating as a student that will eventually withdraw before obtaining a degree and subsequently, deny their acceptance. On the other hand, the misclassification of a student that will withdraw as a student that should persist until graduation would result in less revenue loss and in some situations, i.e., when a student that will eventually withdraw is being displaced, would not result in any loss of tuition revenue despite the occurrence of an error in student classification. Unfortunately, the actual classification of the student being displaced cannot be determined since he or she was never accepted and therefore, would prevent the actual amount of revenue loss to be determined following the misclassification of the initial student. Furthermore, when the cost of making an incorrect admission decision is considered along with the corresponding error rates for those decisions, the misclassification of a student that would persist until graduation as a student that will eventually withdraw, which would result in a larger revenue loss, would also have a lower probability of occurring. While, the misclassification of a student that will withdraw as a student that will graduate would have a greater probability of occurring, it would also have a lower financial impact on the institution. This result suggests that despite the greater probability of making an error in misclassifying a student that withdrawals early as a student that persists until graduating, this error would ultimately have less affect on the institution's "bottom line" than the other error of misclassifying a student that persists until graduating as a student that withdrawals early. However, the occurrence of either error in admission decisions would ultimately result in some amount of tuition revenue loss suggesting the need for improved accuracy in classification of students.

Finally, an institution able to adjust enrollment numbers more efficiently will be able to attenuate their loss of tuition revenue due to admission decision errors. For example, institutions routinely accept more individuals than actually enroll to prevent empty admission seats and admission seats left vacant following a student's withdrawal can be filled with new applicants the following term. However, other sources of institutional revenue such as recruitment cost, student fees, state grants based on institutional performance, and auxiliary services would also contribute to the overall loss in revenue regardless of enrollment efforts (Simpson, 2005; Swail, 2004). These other sources of revenue loss suggest that cost of attrition cannot necessarily be totally "recruited away" and that increased accuracy of students' potential can be one source to reduce the cost associated with attrition.

References

Aleamoni, L. M, & Oboler, L. (1978). Act versus SAT in predicting first semester GPA. *Educational and Psychological Measurement*, *38*, 393-399.

Dorans, J. N., Lyu, C. F., Pommerich, M., & Houston, W. M. (1997). Concordance between ACT assessment and recent SAT I sum scores. *College and University*, *73*, 24-31.

Eimers, M. T., & Pike, G. R. (1997). Minority and nonminority adjustment to college:Differences or similarities? *Research in Higher Education*, *38*, 77-97.

Mathiasen, R. L. (1984). Predicting college academic achievement: A research review. *College Student Journal*, *18*, 380-386.

Mouw, J., & Khanna, R. (1993). Prediction of academic success: A review of the literature and some recommendations. *College Student Journal*, *27*, 328-336.

Murray, M. H. (1997). Better learning through curricular design at a reduced cost. *Journal of Engineering Education*, *86*, 309-313.

Neely, R. (1977). Discriminant analysis for prediction of college graduation. *Educational and Psychological Measurement*, *37*, 965-970.

Noble, J., & Sawyer, R. (1987). *Predicting grades in specific college freshman courses from ACT test scores and self-reported high school grades*. ACT Research Report Series, 87-20. Iowa City, IA.

Noble, J., & Sawyer, R. (1997). *Alternative methods for validating admission and course placement criteria*. AIR Professional File, 63, 1-9.

Pike, G. R. (1991). The effect of background, coursework, and involvement on students' grades and satisfaction. *Research in Higher Education*, *32*, 15-30.

Simpson, O. (2005). The cost and benefits of student retention for students, institutions and governments. Studies in Learning, *Evaluation Innovation and Development*. *2*, 34-43.

Stumpf, H., & Stanely, J. C. (2002). Group data on high school grade point averages and scores on academic aptitude tests as predictors of institutional graduation rates. *Educational and Psychological Measurement*, *62*, 1042-1052.

Swail, W. S. (2004). *The art of student retention: A handbook for practitioners and administrators*. Educational Policy Institute. Texas Higher Education Coordinating Board. 20th Annual Recruitment and Retention Conference, Austin, TX.

UNC A. (2007). Retrieved July 30, 2007, from http://www.unco.edu/acctservices/ftp/budget/UNCFY07_BDB.pdf

UNC B. (2007). Retrieved July 30, 2007, from http://www.unco.edu/acctservices/instanalysis/enrlrpts/archive/fiscalyear/20042005.pdf

UNC C. (2007). Retrieved July 30, 2007, from http://www.unco.edu/acctservices/instanalysis/pdf/retent.pdf

Send correspondence to: Michael Bronsert
University of Northern Colorado
Email: bron3924@blue.unco.edu

The Impact of Graphing Calculator Use on Algebra I End of Course Examinations

Todd Sherron info2knowledge, LLC
Vicki Dimock Southwest Educational Development Laboratory
Rob Foshay Texas Instruments

This study examined the impact of the use of graphing calculators on standardized end of course examinations in Algebra I courses. Researchers sought to answer questions regarding the relationships among the use of graphing calculators on standardized assessments and student achievement, levels of access, and classroom use of graphing calculators. The researchers recruited participation in the study by high schools in two states. Students took a pre- and post- version of a state standardized end-of-course examination without using a graphing calculator then took a second post-test using a graphing calculator. Researchers examined data with descriptive statistics and multiple linear regression, to investigate differences and relationships between mathematics achievement, graphing calculators, and student and teacher variables. Researchers found that students demonstrated higher levels of math performance when a graphing calculator was used. There was a positive correlation between the residual gain scores and students using a classroom set of graphing calculators.

In a meta-analysis of 54 studies of the use of any type of calculator in the classroom, Ellington (2003a) reports, “when calculators were included in testing and instruction, students in grades K-12 experienced improvement in operational skills as well as in paper-and-pencil skills and the skills necessary for understanding mathematical concepts” (p. 456). These findings were for classes of mixed ability students and were not sufficient to generalize to low or high ability classes. Use of calculators for longer periods of time (greater than 9 weeks) appeared to yield more positive effects.

In addition, Ellington’s (2003b) preliminary findings in a meta-analysis of studies of the use of graphing calculators suggest positive effects of the use of graphing calculators on students’ procedural skills, conceptual skills, combined skills, and skills retention. In all four areas, students using graphing calculators outperformed the students who did not have access to graphing calculators on mathematics achievement tests. In three studies, students using graphing calculators retained what they learned better than their non-graphing calculator counterparts. On mathematics tests of conceptual skills and overall math achievement, students who used graphing calculators during instruction outperformed the students who did not use graphing calculators during instruction. Comparison of the retention studies and the studies that lasted long term (16 or more weeks) with the short term studies (less than 16 weeks) revealed that students benefit from using graphing calculators for an extended period of time.

Several states (e.g., Texas, North Carolina, Mississippi, Maryland, and New York) now require the use of graphing calculators in their curriculum standards and on their standardized state assessments. Other states allow, but do not require, the use of graphing calculators on state assessments. In an examination of the use of graphing calculators in Texas high schools and the use of those calculators on the Texas Assessment of Knowledge and Skills (TAKS) (Dimock and Sherron, 2005), a linear regression analysis indicated that holding all else constant, scale scores on the TAKS test were 28 points higher in schools where teachers reported the use of graphing calculators for homework. A second significant positive correlation was found between scale scores and students supplying their own calculators. In schools where this was the case, the average scale scores were 36 points higher. Due to the chronology of the introduction of the TAKS test and the timing of this study, the ability to compare test data both with and without the use of graphing calculators on this test was not possible.

The current study examined the use of graphing calculators on a standardized end-of course examination with students enrolled in Algebra I courses in a state that requires the use of graphing calculators on state assessments and those enrolled in Algebra I in a state that does not require the use of graphing calculators on state assessments. The study sought to answer the following research questions:

1. Does the use of a graphing calculator on an Algebra I End of Course Exam by students influence student achievement as measured by that test?
2. Are there relationships among student achievement scores on an Algebra I End of Course Exam and level of students’ access to, and use of, graphing calculators?

Method and Procedure

To answer these questions regarding the potential relationships among the use of graphing calculators on standardized assessments and student achievement, as measured by those assessments, and the possible relationships of levels of access and classroom use of graphing calculators with scores on this assessment, the researchers recruited participation in the study by high schools in Texas, a state that requires the use of graphing calculators on state assessments, and Arkansas, a state that allows but does not require the use of graphing calculators on state assessments. Teachers who were teaching Algebra I courses in these schools agreed to participate, as indicated by signed informed consent agreements. Parents of the students enrolled in these teachers classrooms signed informed consent for their children to participate.

A repeated measures design was used in which students took two forms of a standardized Algebra I End of Course examination without using a graphing calculator and a third form of that assessment using a graphing calculator. As a pre-test, Form A of an Algebra End Of Course exam was administered to Algebra I students at the beginning of the 2005–2006 school year. Since items can vary in their sensitivity, future usage should identify what percent are graphic calculator friendly and whether items are aligned in the curricula. This test was not used for purposes of determining student placement in a course or to determine any school ratings for adequate yearly progress or other high stakes accountability purposes. Testing situations where students know it doesn't count may affect their motivation and performance, so random assignment was used in the study. Students post-tested on a second form of the Algebra I End-of-Course examination without the use a graphing calculator and a third form of the examination with a graphing calculator. Students were randomly assigned to take one of the two versions of the posttest with and one without graphing calculators. Thus, each student took both versions, but were randomly assigned to whether they used the graphing calculator with Form B or Form C. This should help to eliminate any item or form bias in the student responses.

At the end of the 2006 academic school year, a survey was also administered to teachers and students to collect information regarding variables such as socio-economic status of students in the school, teachers' experience and professional development in the use of graphing calculators, and students' access to and use of graphing calculators in class. As an incentive in the study, teachers received stipends to participate in the study.

Researchers applied a multiple linear regression to examine data for differences and relationships between mathematics achievement, student graphing calculator use, and teacher variables to answer the research questions.

Data Preparation

Teachers administered the tests and then submitted the tests to researchers for scoring. All tests were administered in the regular classroom setting in a 45 minute class period. Correct answers received one point and incorrect answers were scored as zero. Missing values were also scored as incorrect response. A total score for each Test (Test Time1, Test Time 2, and Test Time 3) was created by summing the items q1-q40. Survey data was also coded and entered in to a data file. All data were screened and verified before analyses were performed.

Sample

Researchers recruited schools in two states that had different policies regarding the use of the graphing calculators on state assessments. Data were collected from 362 students, 13 teachers, and 4 schools. There were no significant differences between states with regard to the residualized gain scores. Therefore, the sample is reported in aggregate. Ninety-five percent of the students participating in the study were in the ninth grade. Four percent were tenth graders and 1% was in the eleventh grade. Students enrolled in a Texas high school made up 57% of the participants. Ten of the teachers were from Texas and three were not. The majority of the teachers taught four to six Algebra I classes per day. Eleven of the thirteen teachers indicated they taught 9th grade only. Five teachers indicated they teach in an urban setting and five reported they teach in rural setting. Only one teacher indicated s/he teaches in a suburban setting.

Results

Test Score Results

Table 1 illustrates the aggregate mean scores for all three tests across all students. Test Time 3, the test on which graphing calculators were used, had a significantly higher mean score than either of the two tests taken without a graphing calculator.

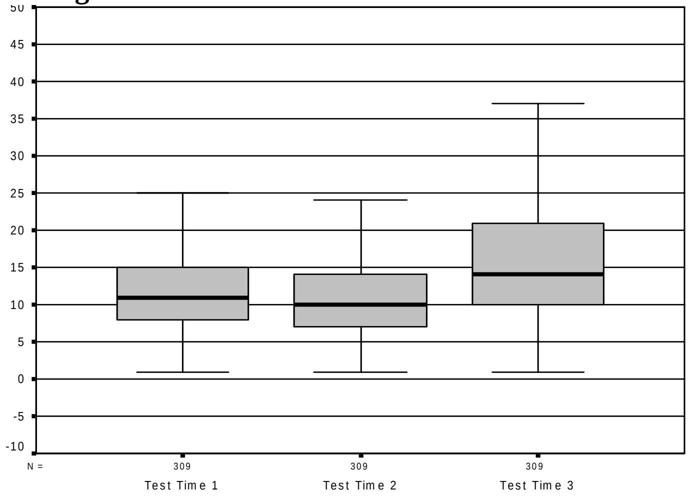
A simply but highly informative graphical method for displaying the spread of scores in a distribution is a box plot. This graphical summary illustrates both the central tendency and the dispersions of scores. The measure of central tendency used in the box plot is the median (although close in value to the mean); the measure of dispersion, which is illustrated by the length of the box, is the inter-quartile range which contains 50% of values (see Figure 1).

Table 1. Mean Test Scores

Mean Test Score	*N*	Mean	SD
Test 1 (without graphing calculators)	309	12.4	5.6
Test 2 (without graphing calculators)	309	11.9	6.9
Test 3 (with graphing calculators)	309	15.7	8.0

To answer the research questions a Linear Regression Analysis was performed to explore the individual contribution of the student variables. For this analysis, the dependent variable was a residualized gain score. That is, students' scores for Test Time 1 and Test Time 2 were regressed onto Test Time 3, calculating residualized or regressed gain scores. These scores were calculated by predicting posttest scores from the pretest scores on the basis of the correlations between Test Time1, Test Time 2 and Test Time 3 (posttest), and then subtracting these predicted scores from the posttest scores to obtain residual gain scores. The effect of the pretest scores is removed from the posttest scores; that is, the residual scores are posttest scores purged of the pretest influence. The residualized gain score has a mean of 0 and a standard deviation of 1. The minimum predicted value was 7 and the maximum was 35.

Figure 1. Box Plot for Test Time



Model Specification

To apply the regression procedure, researchers selected mathematics achievement (residualized gain score) as the dependent variable (Y) to be predicted and explained by independent variables representing availability, type of training, familiarity with a graphic calculator, class time, percentage of instructional activities, and types of use of graphing calculators in the classroom. Researchers specified a model with the following student variables as independent variables:

- Own/Lease my own graphing calculator (T2S2)
- I do not have my own graphing calculator but I use one that is part of the teacher's classroom set while I am in class (T2S4)
- I am eligible for the free or reduced lunch program at my school (T2S5)
- Self-taught without using manual – explored graphing calculator features on my own (q2)
- Self taught using manual (q3)
- Learned how to use as we go in the math course/s I am taking or have taken (q4)
- Graph a function (q5)
- Graph more than one function on the same screen (q6)
- Graph an inequality (q7)
- Graph a scatter plot (q8)
- Create a table (q9)
- Write a program (q10)
- Use the TRACE feature (q11)

- Use the ZOOM feature (q12)
- Use the WINDOW feature (q13)
- Use the INTERSECT feature (q14)
- Use the MAXIMUM and MINIMUM features (q15)
- Connect graphing calculators to motion calculators to motion detectors, computers, or other graphing calculator (q16)
- Teacher Presentation or explanation (q17)
- Whole class discussion (q18)
- Small group work (q19)
- Individual work (q20)
- percent of the instructional activities in your Algebra 1 class involve a graphing calculator (q21)
- To investigate graphs (e.g., to perform stretches, shifts, reflections) (q22)
- To find graphical solutions for different kinds of equations, functions, and relations (q23)
- To check answers (q24)
- To perform direct manipulations of graphs and numerical data (zooming, scaling, scrolling) (q25)
- To create tables (q26)
- To do the more difficult calculations (q27)
- To find maxima, minima, vertices, x- and y- intercepts, and other points on the graph of a function (q28)

The model below was specified and estimated. Note: The base group were students who: (1) did not report any training; (2) were unfamiliar with a graphing calculator; (3) did not estimate class time activities; and (4) did not report types of use.

$$\begin{aligned}
MathGainScore_i &= \beta_1 + \beta_2 T2S2_i + \beta_3 T2S4_i + \beta_4 T2S5_i + \beta_5 q2_i + \\
&\beta_6 q3_i + \beta_7 q4_i + \beta_8 q5_i + \beta_9 q6_i + \beta_{10} q7_i + \beta_{11} q8_i + \beta_{12} q9_i + \beta_{13} q10_i + \\
&\beta_{14} q11_i + \beta_{15} q12_i + \beta_{16} q13_i + \beta_{17} q14_i + \beta_{18} q15_i + \beta_{19} q16_i + \beta_{20} q17_i + \\
&\beta_{21} q18_i + \beta_{22} q19_i + \beta_{23} q20_i + \beta_{24} q21_i + \beta_{25} q22_i + \beta_{26} q23_i + \beta_{27} q24_i + \\
&\beta_{28} q25_i + \beta_{29} q26_i + \beta_{30} q27_i + \beta_{31} q28_i + \varepsilon_i
\end{aligned}$$

Under this analysis, 14% of the variance in the dependent variable (Residual mathematics achievement gain score) was accounted for and 4 of the 30 independent variables were statistically significant with *p*-values in the $p < 0.003$ to $p < 0.05$ range. An experiment wide error rate is plausible since a $p < .05$ level implies that a 1 in 20 chance of significance exists. Consequently, one or two variables may be significant by chance alone. See Table 2 for parameter estimates.

Parameter Interpretation

Interpretation of the parameter estimates is as follows:

There is a positive correlation between the residual gain score and students using classroom set of graphing calculators ($t = 2.065$, $p < 0.004$). In other words, the average student math residual gain scale scores increases by 0.369 residual points if the student remarked they used a classroom set of graphing calculators ($\beta_1 + \beta_3$).

Variable (q2) Self-taught without using manual – explored graphing calculator features on my own was statistically significant ($t = 2.35$, $p < 0.019$). That is, the average student math residual gain scale scores increases by 0.372 points if the student remarked they self-taught without using manual – explored graphing calculator features on my own ($\beta_1 + \beta_5$).

As the variable (q6) familiarity of graphing more than one function increases by 1 unit, math residual gain scale scores increases by 0.194 points, holding all else constant ($t = 2.02$, $p < 0.045$).

The variable q20 (Individual work) was statistically significant ($t = 3.03$, $p < 0.003$). That is, as the variable q20 Time spent on Individual work increases by 1 unit, math residual gain scale scores increases by 0.175 points, holding all else constant.

Table 2: Parameter Estimates

Model	Unstandardized Beta	Std Error	Standardized Beta	t	Sig
(Constant)	-1.788	.380	-	-4.70	.00
T2S2	.284	.154	.136	1.84	.07
T2S4	**.369**	**.178**	**.163**	**2.07**	**.04***
T2S5	-.008	.125	-.044	-.69	.49
Self-taught w/out manual	**.372**	**.158**	**.152**	**2.35**	**.02***
Learned as we go	.253	.168	.099	1.51	.13
Graph a function	.006	.116	.006	.06	.95
Graph more than one function	**.194**	**.096**	**.212**	**2.02**	**.05***
Graph inequality	.001	.079	.016	.22	.83
Graph a scatter plot	-.001	.070	-.013	-.17	.87
Create Table	-.003	.072	-.004	-.05	.96
Write a program	-.109	.080	-.093	-1.37	.17
Use the Trace	-.005	.070	-.059	-.71	.48
Zoom	.004	.083	.047	.54	.59
Window	-.003	.078	-.004	-.05	.96
Intersect	-.009	.071	-.104	-1.33	.18
MaxMin	.004	.072	.049	.66	.51
Connect to motion detector	-.003	.069	-.035	-.51	.61
Teacher pres/explain	-.008	.067	.085	1.28	.20
Whole class discussion	-.008	.062	-.101	-1.43	.15
Small group	-.009	.059	-.107	-1.61	.11
Individual work	**.175**	**.058**	**.189**	**3.03**	**.003***
% of instructional activities	.008	.070	.079	1.20	.23
Investigate graphs	.004	.142	.002	.03	.97
Find graphical solutions	.004	.176	.018	.27	.79
Ck answers	.233	.180	.088	1.29	.20
Perform direct manipulations	.104	.135	.054	.77	.44
Create tables	-.101	.166	-.046	-.61	.54
Find max/min	.008	.166	.038	.52	.60

Note. * Statistically significant

In conclusion, four of the 30 independent student variables positively correlated to the math residual gain scores. These four variables explained 14% of the variance in the math residual gain scale scores with 86% unexplained variance due to other factors (variables). Analysis revealed that: (1) The average student math residual gain scores increases if the student remarked they used a classroom set of graphing calculators; (2) The average student math residual gain scale scores increases if the student remarked they were self-taught without using manual – explored graphing calculator features on their own; (3) As familiarity of graphing more than one function increases by 1 standard deviation, math residual gain scale scores increase; and (4) As the amount of time spent on individual work increases by 1 standard deviation, math residual gain scale scores increase.

Conclusions

The purpose of this study was to investigate the impact of graphing calculator use on Algebra I end of course examinations. This study sought to answer two research questions:

1. Does the use of a graphing calculator on an Algebra I End of Course Exam by students influence student achievement as measured by that test?

2. Are there relationships among student achievement scores on an Algebra I End of Course Exam and level of students' access to graphing calculators?

Analysis did reveal that students scored higher on standardized assessments when a graphing calculator was used. Further, regression analysis indicated positive relationships among student achievement scores on an Algebra I End of Course Exam and level of students' access to graphing calculators. That is, (1) The average student math residual gain scores was higher if the student remarked s/he used a classroom set of graphing calculators, (2) The average student math residual gain scale scores was higher if the student remarked s/he was self-taught without using a manual, that is, s/he explored graphing calculator features on their own, (3) As familiarity of graphing more than one function increased by 1 unit, math residual gain scale scores increased, (4) As the amount of time spent on individual work increased by 1 unit, math residual gain scale scores increased.

Discussion

The lack of other significant variables may offer insight into the nature of the mathematics instruction on the campus. For example, socio economic status (SES) which is commonly reported when predicting student achievement was not a statistically significant predictor variable. It may also suggest the level of use or student familiarity with the graphing calculator. Nonetheless, this study adds evidence to the body of research suggesting that the use of a graphing calculator makes a significant and practical impact in mathematics achievement in Algebra I classes. Our findings are consistent with the findings of Ellington's (2003) meta-analysis indicating that the use of graphing calculators in testing significantly improved performance. Her findings were for classes of mixed ability students and were not sufficient to generalize to low or high ability classes.

The regression analysis demonstrated significant positive relationships among student achievement scores on an Algebra I End of Course Exam, use of a graphing calculator on that examination, and level of students' access to graphing calculators. Average student math gain scores were higher if the student remarked s/he used a classroom set of graphing calculators. The average student gain scale scores was higher if the student remarked s/he had explored graphing calculator features on her/his own by using the manual. As the mean score for student familiarity with graphing more than one function increased by 1 unit, student math gain scale scores increased. Finally, as the reported amount of time spent on individual work increased by 1 unit, math gain scale scores increased. Thus, factors that may impact student performance as measured by standardized assessments are student access to graphing calculators, student knowledge of how to use graphing calculator functions, and the use of graphing calculators on standardized assessments.

References

Dimock, V. & Sherron, T. (2005). *A Study of the Impact of Graphing Calculators*. Dallas, TX: Texas Instruments.

Ellington, A. (2003). A meta-analysis of the effects of calculators on students' achievement and attitude levels in precollege mathematics classrooms. *Journal for Research in Mathematics Education*. *34*(5), 433-463.

Send correspondence to: Todd Sherron
info2knowledge, LLC
Email: todd@toddsherron.com

Coors Field: A Pitchers Graveyard?

Jay Schaffer **Raj Chandran**
University of Northern Colorado

Batting and pitching statistics for the Colorado Rockies have long been considered inflated by sports writers and fans. Schaffer and Heiny (2006) documented a Coors Field effect on slugging percentage. This research examines the Coors Field effect on pitching statistics, ERA and on-base percentage.

In their 14 years of existence, the Colorado Rockies have not yet distinguished themselves as a "good team". They have only made the playoffs once, as a wildcard team, in the strike shortened season of 1995. It would be easy to blame Rockies pitching, since their statistics are perennially at the bottom of the league. In fact, the Rockies ranked 26 out of 28 teams in opponent batting average in the only season they made the playoffs. The Rockies have always had to make up for their weak pitching with their offensive prowess, aided in part by the elevation of their ballpark, Coors Field. Schaffer and Heiny (2006) demonstrated a significant slugging percentage advantage by playing half of their games in the thin air of Coors Field.

Is the elevation of Coors Field the bane of Rockies and visiting pitchers? The answer of nearly every sportswriter would be an emphatic "*Yes*!" Johnathan Leshansky of athomeplate.com sums up pitching at Coors Field: "Of course pitching at Coors is like trying to defuse a bomb when you have a bad case of the shakes. You might do alright for a while, but the odds are that eventually something is going to blow up in your face." In fact, the general manager for the Rockies, Dan O'Dowd even stated, "I'm not sure even if we had Randy Johnson, Curt Schilling, or those kinds of guys in our rotation. I'm sure they'd be good, but I don't think they'd be as good as they are pitching elsewhere."

According to www.baseball-reference.com, the air pressure in Denver is about 15% lower than at other parks near sea level. Reduced air pressure decreases aerodynamic forces on the baseball by the same amount. Thus, there is less movement on breaking pitches, making them easier to hit, and less drag on balls in flight allowing baseballs to fly further.

In a study performed by Schaffer and Heiny (2006), the effect of elevation on slugging percentage was examined. By performing a repeated measures ANOVA, Schaffer and Heiny concluded that a significant effect of elevation existed on slugging percentage. This study will use the same analyses as Schaffer and Heiny (2006), but examine earned run average and on base percentage.

In order to help bridge the gap between Coors Field being a hitters park versus an average ballpark, the Rockies began putting their baseball's in a humidor. The unorthodox practice began in 2002 and has been ongoing since. The method was so well received that now all 30 MLB teams keep their baseballs in a climate controlled environment. In 2006, the Rockies posted their smallest disparity in earned run average, posting a 4.72 ERA at home and a 4.59 ERA on the road. Part of this receding gap may be due to Rockies pitchers adjusting to the effects of elevation, though many believe that the implementation of the humidor is the main cause.

The effect of high elevation on the flight of the baseball has been studied by Sterling Professor Emeritus of Physics at Yale University and "Physicist to the National League" Robert K. Adair (2002). He wrote, "Since the retarding force on the ball is proportional to the density of the air, the ball will travel farther in parks at a high altitude. A 400-foot drive by Sammy Sosa or Mark McGwire at Shea Stadium, near sea level, on a windless summer day would translate to a 404-foot drive in Atlanta on the Georgia Piedmont at 1,050 feet, the highest park in the majors before 1994. The same home run could be expected to go about 403 feet in Kansas City and 403 feet at the Metrodome in Minneapolis or Wrigley Field in Chicago. These differences are not so great as to modify the game, but Sosa could expect his long drive to travel about 420 feet at mile-high Denver. And if the major leagues are further internationalized someday, say to Mexico City, at 7,800 feet, Sosa's blow could sail nearly 430 feet. Old home run records will be swept away unless the fences are moved out in the high parks."

Adair mentions that moving the fences back is not the only solution. He continues, "Even if the fences are adjusted, the high-altitude stadiums will still be a batter's boon and a pitcher's bane. With fences moved back, there will be acres of outfield for balls to fall into for base hits, and, though the pitcher's fastball will be about six inches quicker in Denver, the curve will bite about 20% less, which is more important.

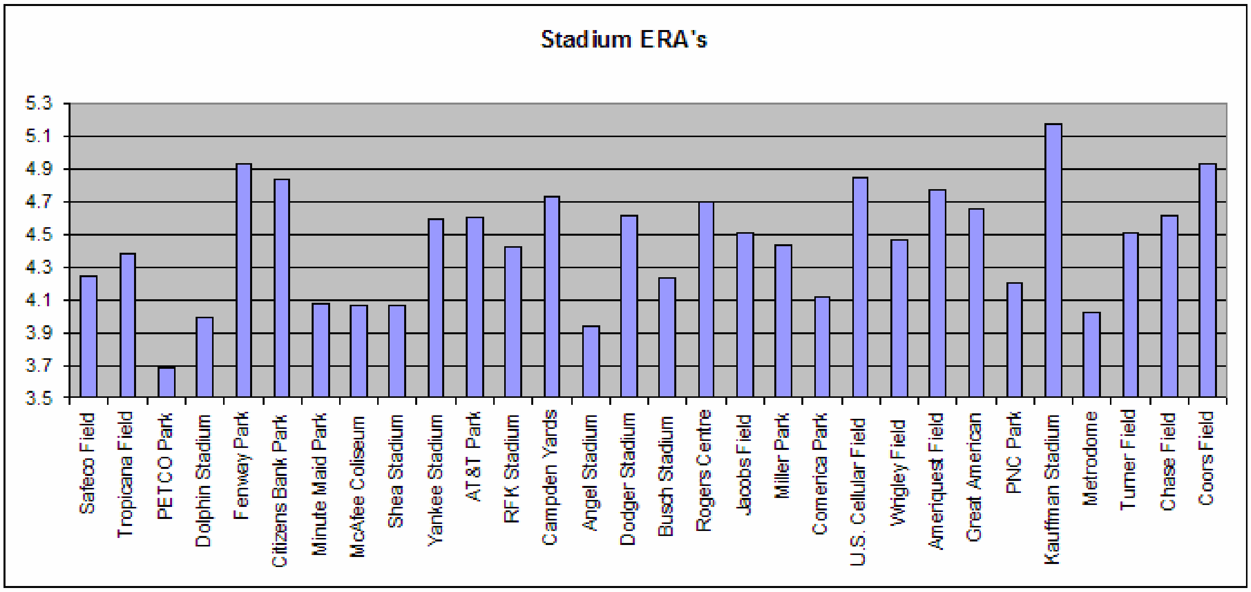


Figure 1. ERA for each Stadium

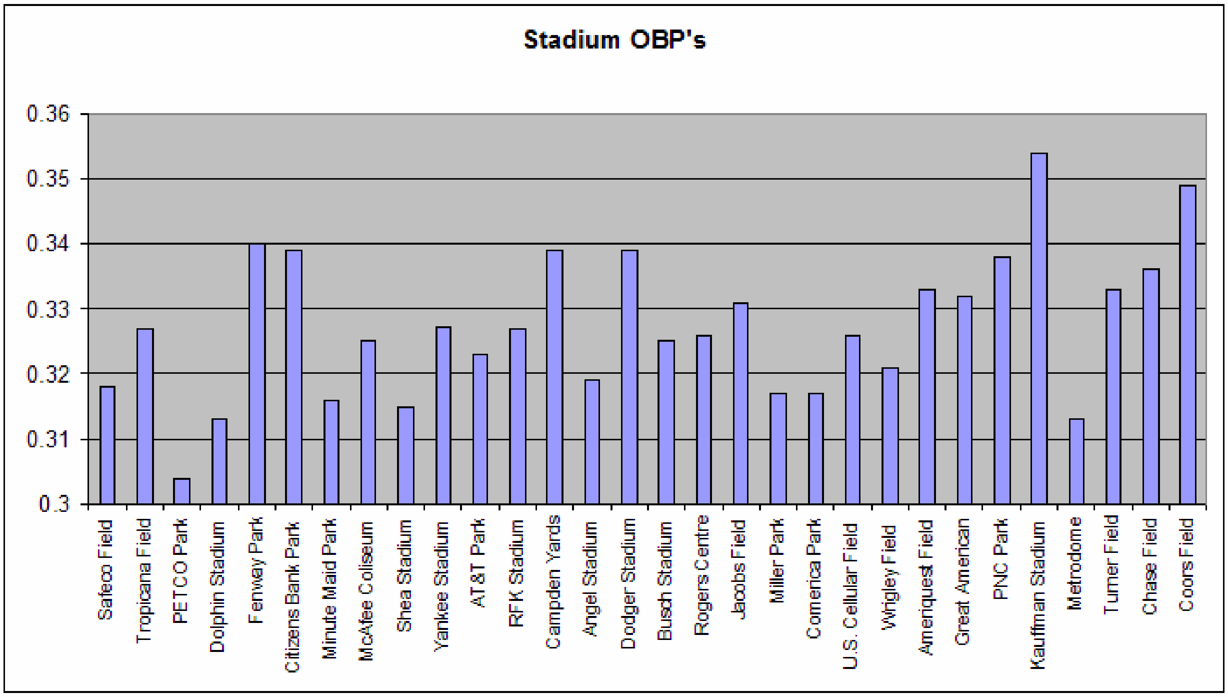


Figure 2. OBP for each stadium.

With less drag, the ball will also get to the outfielders faster in Denver than at Fenway Park in Boston. Players for the Colorado Rockies have noted that in Denver's outfield, 'Fly balls come at you faster and sail farther than you might expect.' Indeed, a hard-hit 'gapper' between the outfielders will reach the 300-foot mark about two-tenths of a second faster in Denver than at sea level, cutting down the pursuit range of an outfielder by five or six feet—not inconsiderable in this game of inches. Even the range of a shortstop covering a line drive or one-hopper will be cut by about a foot in Denver."

When considering the pitching statistics of Rockies starting pitcher Jeff Francis, Adair's theory seems to hold true. In 2006, Francis posted an ERA of 4.30 at home and a 4.05 away. In addition, Francis had a 0.339 on-base percentage against, OBP, at home and a 0.324 on the road. Francis has played his entire MLB career with the Rockies starting in 2004. Thus far, Francis has posted a career ERA of 7.66

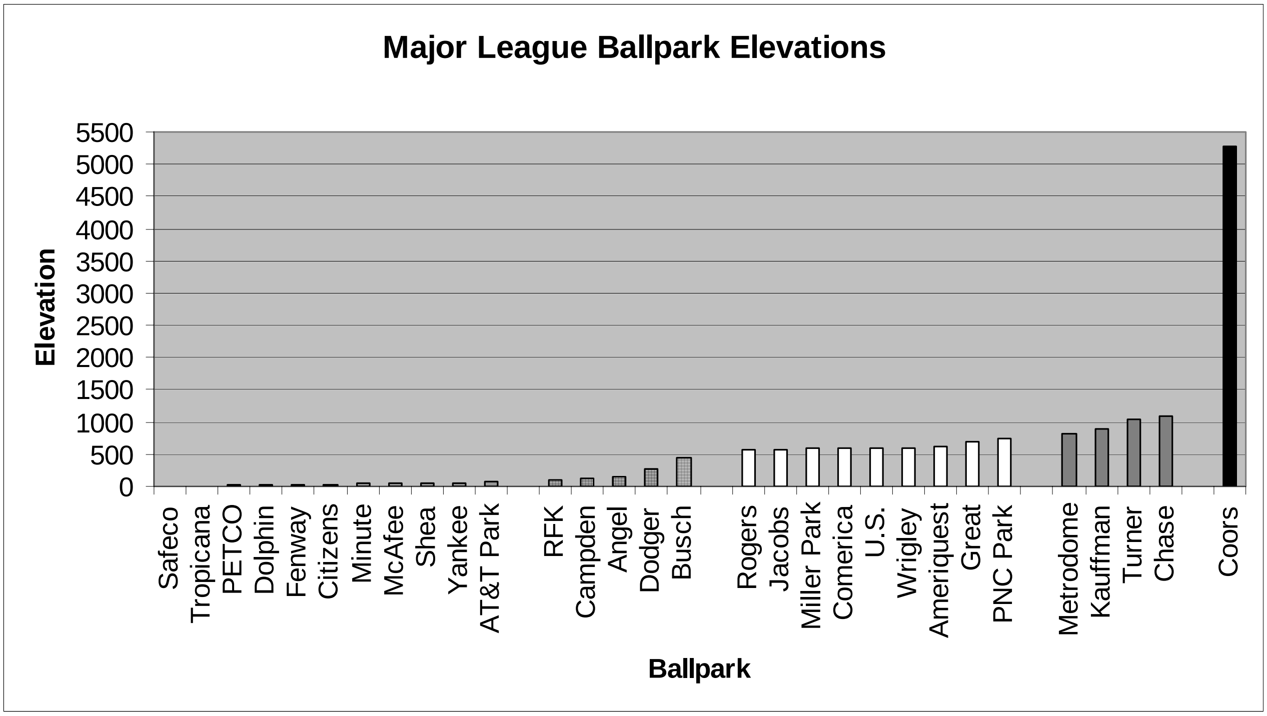


Figure 3. Elevations of major league ballparks.

at Coors Field and a 3.86 ERA on the road. The phenomenon extends beyond just Coors Field. In general, stadiums with high elevation have high pitching statistics. Figures 1 and 2 display 2006 ERA and OBP respectively for each stadium. It should be noted that Coors Field has nearly the highest ERA and OBP when compared with the other stadiums.

Data

Data was taken from sportsnet.ca, a leading Canadian sports company, who obtain their stats from STATS LLC. STATS LLC gathers stats for the Major League Baseball Association. Data was collected such that each pitcher's statistics were tallied for each stadium he pitched at for the 2006 season. The pitching statistics that were collected initially are shown in the appendix 1. Elevations for each ballpark were found using the US Geological survey website, www.usgs.gov, and are shown in Figure 3.

Methods

An ANOVA procedure was used to determine if elevation had a significant effect on pitching statistics. An unbalanced, repeated measures design with nested factors was used. A treatment was considered each combination of ballpark and elevation. The subjects in this experiment were the pitchers. This is considered an unbalanced design because none of the pitchers play at every ballpark. The response variables were ERA and OBP. In this study, ERA and OBP were weighted by innings pitched. The model used is shown below in equation 1:

$$X_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \gamma_k + \varepsilon_{ijk} \; ; \qquad (1)$$

where, X_{ijk} = ERA or OBP of a pitcher; α_i = effect of elevation; $\beta_{j(i)}$ = effect of ballpark; γ_k = effect of player; μ = overall mean ERA; and ε_{ijk} = random error.

The effects for elevation, ballpark and player were treated as fixed effects. The data used in this study were not a random sample of ballparks, elevation or players from a larger population, but rather a collection of pitching statistics from the entire major league for the 2006 season.

In order to test whether elevation had a significant effect on pitchers both ERA and on-base percentage against were used. ERA is one of the oldest and most popular statistics gathered on pitchers. ERA is calculated by equation (2).

$$\text{ERA} = \frac{\textit{Number of Earned Runs} * 9}{\textit{Number of Innings Pitched}} \qquad (2)$$

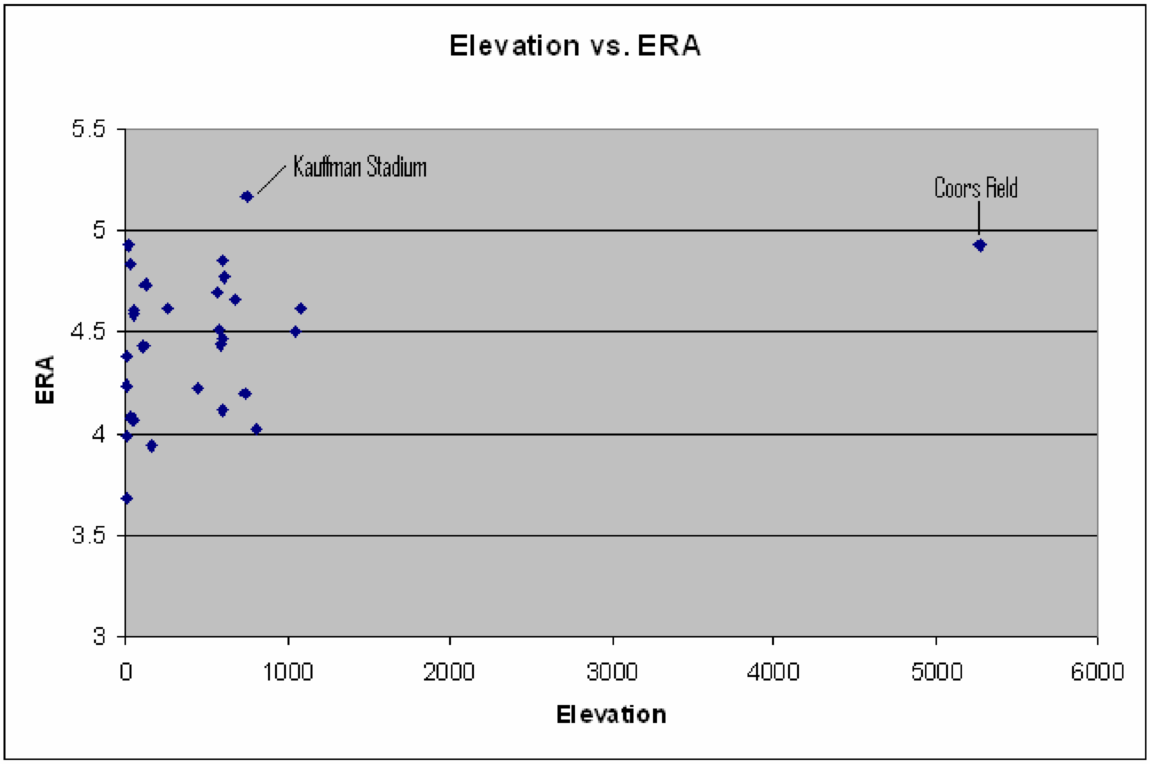


Figure 4. Elevation versus ERA

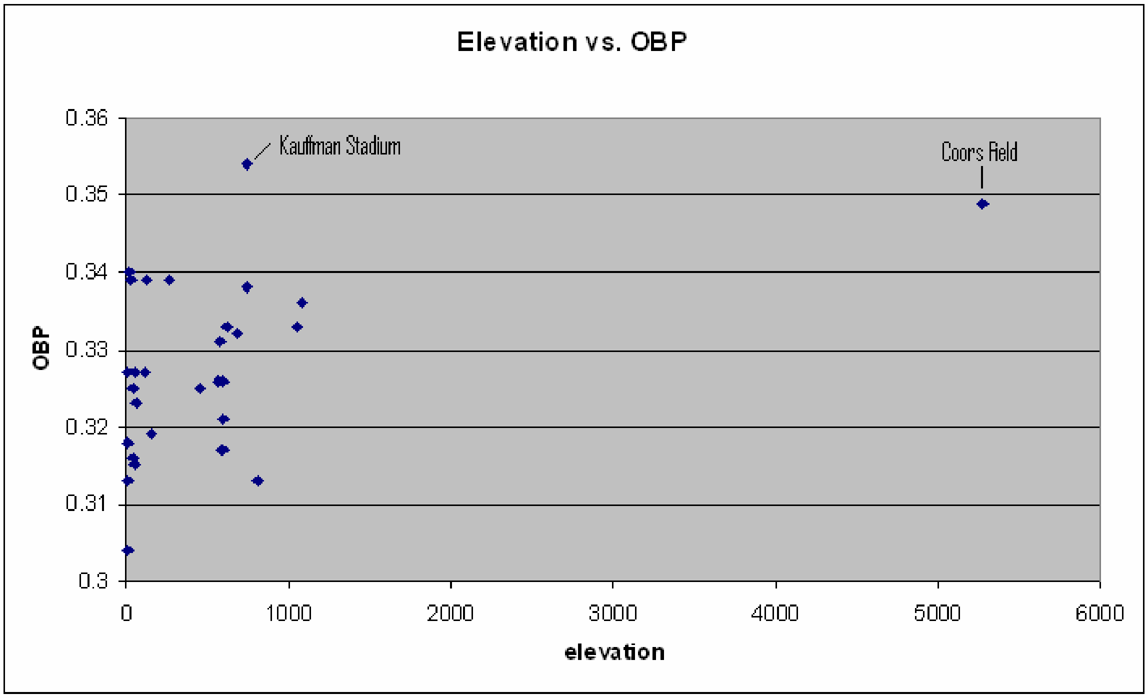


Figure 5. Elevation versus OBP.

Earned runs do not include errors by the catcher or position players, as the pitcher did not control this.

In addition to ERA, OBP was used to capture hits given up by pitchers. While batting average against is a widely used statistic to capture hits given up, it does not account for walks, players hit by a pitch, and sacrifice flies. OBP takes into account these additional statistics and is calculated by equation (3).

$$\text{OBP} = \frac{Hits + Walks + Batters\ Hit\ By\ Pitch + Sacrifice\ Flies}{Number\ of\ Opponent\ At - Bats} \qquad (3)$$

Table 1. Levels of Factor Elevation

Level	Elevation Range	Number of Ballparks
1	under 100 ft	11
2	between 100ft and 500ft	5
3	between 500ft and 800ft	9
4	between 800ft and 1,000ft	4
5	over 1.100 ft	1

Using elevation as a numerical variable presents analytical difficulties, as Coors Field is 5,277 feet above sea level and the other 29 ballparks are all below 1,100 feet. Figures 4 and 5 show ERA versus elevation and OBP versus elevation respectively, for each major league stadium. As illustrated in both figures, Coors Field is an obvious outlier with respect to elevation. Due to the leverage point created by Coors Field, it would be detrimental to fit any type of regression model to the data using elevation as a continuous variable. Therefore elevation was categorized into five levels, so that a reasonable number of teams would be distributed into each level of elevation. Coors Field, in Denver, Colorado was categorized into a level of its own for reasons discussed previously. In addition, ballparks with elevations less than 100 feet were considered their own level due to the large amount of ballparks that fit this into that category. The elevation range and number of stadiums in each level is shown in Table 1. The elevation of each major league city is displayed in Figure 3 with a space between each level for the factor elevation.

As seen in Figures 4 and 5, Kauffman Stadium, home to the Kansas City Royals was the only stadium that posted higher ERA and OBP than Coors Field. When considering the characteristics of the Royals, it may be safe to say that factors other than elevation played an important role in Kauffman Stadium's poor statistical showing. With a record of 62 wins and 100 losses, the Royals were one half of a game away from holding the worst record in the major league baseball in 2006. The Royals also posted a league worst 5.68 ERA at home. The overall Kauffman Stadium average for ERA was 5.17 indicating mainly the Royals struggled in the ERA category at Kauffman stadium. This would imply that the elevation may have less to do with Kauffman stadium's high ERA and more to do with the poor pitching on the part of the Royals.

Baseball is different from other team sports in that each ballpark has its own set of unique characteristics. These characteristics can either hinder or help pitchers. The most obvious of these characteristics is that of the length of outfield fences. However other factors may also play an important role. Stadiums such as McAfee Coliseum in Oakland, California and Dodger Stadium in Los Angeles, California have expansive foul territories which aid pitchers and hinder batters. Faster playing surfaces such as Astroturf® and Fieldturf® can make it much easier for ground balls to get through to the outfield for hits. Weather conditions also vary greatly from ballpark to ballpark, with some teams playing indoors in a climate controlled environment, such as Minnesota Twins, and while others like the White Sox in Chicago battle the wind. Rather than trying to account for all the stadium factors separately, an overall ballpark factor was used. The overall ballpark factor is nested within the factor elevation.

It is also of interest to note that the American League has consistently higher run production over the National League, because the rules dictate that teams in the American League may have a designated hitter to hit in place of their pitcher, while teams in the National League do not have this luxury. A simple t-test comparing the means of both ERA and OBP for each league showed there was not a significant difference. Thus, league was not included as a factor in the model.

The opposing team may be an important factor in the model shown in equation 1, as certain teams tend to have better offensive production than others. Unfortunately, those statistics were not available at the time of this study. Therefore, opposing team offense was not considered as a factor.

Results

Results were found using PROC GLM in SAS. An ANOVA table for ERA and OBP was generated for the model shown in equation 1. The model using ERA as a dependent variable is shown in Table 2 while the model using OBP is shown in Table 3.

The R^2's for the ERA and OBP models were 0.14 and 0.18 respectively, which seems to indicate that both models do not account for a large portion of variation for their respective dependent variables. However, low R^2 values have been common in previous baseball studies. In the analysis performed by Schaffer and Heiny (2006), they stated, "Additional independent variables could be added, but most likely the randomness of baseball never can be accounted for completely. Players go through hot and cold streaks for reasons they do not even understand." The "hot and cold streaks" that Schaffer and Heiny

(2006) refer to are independent of ballpark and elevation, but still contribute variability to the study. Previous studies have made similar conclusions regarding the randomness of baseball. Hofacker (1998) analyzed major league baseball data for the 1982 season. The study attempted to measure a baseball team's offensive ability independent of opponent and ballpark. Hofacker used runs scored as a dependent variable, and opponent, park, league and home versus away as independent variables. The R^2 for this study was 0.267. Hofacker defended his low R^2 by stating, "While it is true that researchers in some fields might scoff at such a low R^2, perhaps the better way to think about the current result is that it offers insight into just how stochastic baseball must be. Such considerations necessarily imply that the analysis presented be considered exploratory." The purpose of this study is not to predict ERA or OBP for a pitcher, but rather to determine if elevation is a significant effect on pitching.

Table 2. ANOVA Table for ERA

Source	*df*	Sums of Squares	Mean Square	*F*-value	*p*-value
Model	651	79,393.12	121.96	1.39	<.0001
Error	5,625	494,360.76	87.89		
Corrected Total	6,276	573,753.88			

Table 3. ANOVA Table for OBP

Source	*df*	Sums of Squares	Mean Square	*F*-value	*p*-value
Model	651	58.77	.09	1.95	<.0001
Error	5,625	260.59	.05		
Corrected Total	6,276	319.36			

Table 4. Repeated Measures ANOVA Table for ERA

Source	*df*	Sums of Squares	Mean Square	*F*-value	*p*-value
Elevation	4	1027.60	256.90	2.92	0.0199
Park(Elevation)	25	2971.59	118.86	1.35	0.1127
Player	622	74,424.28	119.65	1.36	<.0001
Error	5,625	494,360.76	87.89		

Table 4. Repeated Measures ANOVA Table for OBP

Source	*df*	Sums of Squares	Mean Square	*F*-value	*p*-value
Elevation	4	0.94	0.23	5.09	0.0004
Park(Elevation)	25	2.38	0.09	2.05	0.0015
Player	622	53.81	0.09	1.87	<.0001
Error	5,625	260.59	0.05		

Tables 4 and 5 list each factor with its degrees of freedom, Type III sums of squares, mean squares, *F*-statistics and *p*-values for ERA and OBP, respectively. Elevation has a statistically significant contribution to both ERA and OBP. The *p*-value for ERA was 0.0199 and 0.0004 for OBP. The compound symmetry assumption for this study was violated due to the hot and cold streaks that pitchers experience during the season. Pitches thrown in a game are likely to be highly correlated, while pitches thrown in a different game will be less correlated. Regarding the compound symmetry assumption Neter et al. (1996) stated, "In repeated measures studies, the compound symmetry assumption will be violated, for instance, if repeated responses over time are more highly correlated for observations closer together than for observations further apart in time." Neter et al. (1996, *p*. 1170) suggested using a more conservative critical value because the test becomes more liberal when the compound symmetry assumption has been violated. Even with this more conservative critical value, $F(0.95;\ 1,\ 622) = 3.84$, elevation still has a statistically significant effect on OBP. However, using the more conservative critical value for the ERA model shows elevation is no longer statistically significant effect.

Despite ERA no longer being significant, using the conservative degrees of freedom suggested by Neter et al., a post hoc test was still performed in order see if differences still existed between elevations assuming the compound symmetry assumption had not been violated. A post hoc test was also performed on the OBP model. Glass and Hopkins (1996) suggest using Student Newman-Keuls (SNK) due to its power and high degree of protection for the entire [omibus] null hypothesis. The results from the SNK tests for ERA and OBP are shown in Tables 6 and 7 respectively. Groups with different SNK groupings (A versus B) are statistically different from each other. N represents the number of players who pitched at that particular elevation level. In addition, the means for ERA and OBP are shown for each level of Elevation in Tables 6 and 7.

Table 6. ERA by Elevation.

SNK Grouping		Mean	*N*	Elevation
A		4.9330	216	5
B	A	4.5800	828	4
B	A	4.5296	1870	3
B		4.3941	1037	2
B		4.3043	2326	1

Table 6 shows that Coors Field, in elevation 5, is statistically different from ballparks located in the two lowest elevations, 1 and 2. It should be noted that as elevation increases, intuitively ERA increases too. By looking at the mean ERA per elevation, it would appear there are three relative groups. The two lowest elevations, 1 and 2, have a mean ERA near 4.35, while ERA for elevations 3 and 4 are near 4.55, and the highest level, Coors Field, is 4.93. The mean ERA for the Coors Field elevation is about 8.4% higher than the "middle elevations" and 13.3% higher that the "low elevations."

Table 6. ERA by Elevation.

SNK Grouping	Mean	*N*	Elevation
A	0.342	216	5
B	0.329	828	4
B	0.325	1870	3
B	0.323	1037	2
B	0.319	2326	1

The post hoc test for OBP showed that there is a statistically significant difference between the elevation of Coors Field and the other four levels of elevation. The OBP means for each level of elevation exhibit the same trait as ERA; as elevation level increases, so does OBP. Contrary to the findings of ERA, the post hoc test for OBP did not show three distinct groups, but rather two distinct groups, Coors Field versus the other four levels. The mean OBP for Coors Field is 4% higher than elevation 4, and 7.2% higher than elevation 1.

It should be noted that due to Coors Field being an numerical outlier in elevation, and the only ballpark in elevation category 5, the ballpark effect and elevation effect are confounded. However, Coors Field is one of the largest ballparks in the Major Leagues. In fact, Coors Field has the second longest left field dimension, and third longest center and right field dimensions in Major League Baseball. Despite these large dimensions, the z-scores for Coors Field within each set of 30 left field, right field and center field dimensions were 1.70, 1.32 and 1.80 respectively; indicating Coors Field is not an outlier with respect to ballpark dimensions. Additionally, Coors Field has a relatively small foul territory, but is still similar to most other ballparks. The abnormally large foul territories of Dodger Stadium and Network Coliseum are the exception rather than the rule.

It would appear that Denver's outlier status with regards to elevation is the only factor that makes Coors Field significantly different from the other ballparks. It is therefore reasonable to conclude the high elevation is the primary cause for the high ERA and OBP exhibited at Coors Field.

Conclusions

The model used in this paper has demonstrated that elevation significantly impacts OBP, independent of ballpark and player. The model also showed that elevation marginally impacts ERA, but is consistent with the theory that ERA increases as the elevation increases. At Coors Field, ERA is approximately 8.4% higher than the middle elevations between 500 and 1,100 feet, and 13.3% higher than the low elevations less than 500 feet. Differences in OBP showed Coors Field in "a league of its own", with the other four elevations grouping close together at lower values.

Given that young Rockies prospect Jeff Francis has played his entire career at the with the Rockies, it may interest a team owner, manager, sports writer or fan to know how well he might do if he were traded to different team. If Francis was traded to a middle elevation team his home ERA of 4.30 in 2006 would be adjusted down to 3.94 and OBP would have to be adjusted from 0.339 to 0.323. This would give Francis an overall ERA of about 4.00 and an OBP of about 0.324 if he were to be traded to a middle elevation team. If he were to be traded to a low elevation team, his expected ERA and OBP would drop even further. His home ERA would now drop to 3.73, and his OBP would drop to 0.317. This would bring his overall ERA to a respectable 3.89 and his OBP to 0.321.

This study determined that the effect of elevation and ballpark are confounded in Denver. However an examination of the ERA and OBP effects of each ballpark versus elevation level and the dimensions of Coors Field with respect to the other ballparks rule out the ballpark effect explaining the high OBP and ERA experienced at Coors Field. In fact, elevation appears to be the only viable explanation.

The results seem to indicate that there may be other variables and factors (weather, opposing team batting average, etc) that may influence ERA and OBP. While additional independent variables may be added to the model to account for more error, most likely the model would only be improved marginally. The randomness of baseball can probably never be fully accounted for. Players go through hot and cold streaks throughout the season that even they cannot explain. Certainly, even some players may have an off season, or some may be partial to pitching at certain times during the day.

Future studies may want to examine the effect of the humidor used by the Rockies, and other teams, to see if it has an effect on the game. It may also be interesting to follow several Rockies players over the course of several years, in a longitudinal study, to see if their pitching or hitting statistics change over the course of their transition from the Rockies to a different team and vice-versa. Variables such as temperature, left versus right handed, and years in the league could also be on importance variables in future studies.

References

Adair, R.K. (2002). *The Physics of Baseball*, (3rd Ed.). New York: Harper-Collins.

Coors Field. Retrieved April 27, 2007 from http://www.baseball-reference.com/bullpen/Coors_Field

Glass, G., & Hopkins, K (1996). *Statistical Methods in Education and Psychology*, (3rd Ed.). Allyn & Bacon.

Kaufman, K. (2002, April). Colorado's Rocky Road. Retrieved March 20, 2007 from http://www.salon.com/news/sports/col/kaufman/2002/04/09/rockies/index.html

Leshansky, J. (2005, February). 2005 Season Preview: Colorado Rockies. Retrieved April 26, 2007 from http://www.athomeplate.com/rockiespre05.shtml

Neter, J., Kutner, M., Nachtsheim, C., & Wasserman, W., (1996). *Applied Linear Statistical Models*, (4th Ed.). McGraw-Hill Companies.

Schaffer, J., & Heiny, E. (2006). The Effects of Elevation on Slugging Percentage in Major League Baseball. *Chance*, *19*(*1*), 28-34..

Send correspondence to: Jay Schaffer
University of Northern Colorado
Email: Jay.Schaffer@unco.edu

Appendix 1

Statistic	**Definition**
G	Games Played
GS	Games Started
ERA	Earned Run Average
W	Wins
L	Losses
SV	Saves
IP	Innings Pitched
H	Hits
R	Runs
ER	Earned Runs
BB	Bases on Balls (Walks)
CG	Complete Games
SVO	Save Opportunities
HR	Home Runs
2B	Doubles Given up
3B	Triples Given up
OPAVG	Opponent Batting Average
OBP	On-Base Percentage Against
SLG	Slugging Percentage

Multiple Linear Regression Viewpoints
Information for Contributors

Multiple Linear Regression Viewpoints (*MLRV*) is a publication sponsored by the American Educational Research Association's Special Interest Group on Multiple Linear Regression: The General Linear Model (MLR/GLM SIG). It is published twice a year to facilitate communication among professionals who focus their research on the theory, application, or teaching of multiple linear regression models and/or the general linear model. Manuscripts submitted to *MLRV* should conform to the language, style, and format of the Publication Manual of the American Psychological Association (5th ed., 2001). Three copies (two blind) of a doubled spaced manuscript (including equations, footnotes, quotes, and references) of approximately 25 pages in length, a 100 word abstract, and an IBM formatted diskette with the manuscript formatted in WordPerfect or Word should be submitted to one of the editors listed below.

Mathematical and Greek symbols should be clear and concise. All figures and diagrams must be photocopy-ready for publication. Manuscripts will be anonymously peer reviewed by two editorial board members. Author identifying information should appear on the title page of only one submitted manuscript. The review process will take approximately 3 months. A letter acknowledging receipt of the manuscript will be sent to the first author, and upon review, a letter indicating the peer review decision will be sent to the first author. Potential authors are encouraged to contact the editors to discuss ideas for contributions or determine if their manuscript is suitable for publication in *MLRV*.

EDITOR INFORMATION

Randall E. Schumacker, Editor *MLRV*
College of Education
P.O. Box 870231
Carmichael Hall 316
The University of Alabama
Tuscaloosa, AL 35487-0231
(205) 348-6062
rschumacker@ua.edu

T. Mark Beasley, Associate Editor
Department of Biostatistics
School of Public Health
309E Ryals Public Health Bldg.
University of Alabama, Birmingham
Birmingham, AL 35294
(205) 975-4957
mbeasley@uab.edu

ORDER INFORMATION
Cynthia Campbell, MLR/GLM SIG Executive Secretary
Department of Educational Technology, Research & Assessment
Northern Illinois University
DeKalb, IL 60115-2854
ccampbell@niu.edu

Check out our website at: http://mlrv.ua.edu/

POSTMASTER: Send address changes to:
Cynthia Campbell, MLR/GLM SIG Executive Secretary
Department of Educational Technology, Research & Assessment
Northern Illinois University
DeKalb, IL   60115-2854

Multiple Linear Regression Viewpoints (ISSN 0195-7171) is published by the
AERA Special Interest Group on Multiple Linear Regression: General Linear Model
through the **University of Alabama at Birmingham** and the **Dallas Independent School District**.