# 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 37 • Number 1 • Spring 2011

Table of Contents Page

# *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 **Northern Illinois University** and the **University of Alabama-Birmingham**.

Subscription and SIG membership information can be obtained from:
**Cynthia Campbell, Managing Editor**
**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*.

# Multiple Linear Regression Viewpoints

*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 (SIG/MLR: GLM). 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* (6th ed., 2010). Manuscripts should be prepared in Word, be doubled-spaced, use 12 font, contain a 100 word abstract, have author(s) identifying information appear on the title page only, and consist of no more than 30 pages in length (including equations, footnotes, quotes, and references). Mathematical and Greek symbols should be clear and concise. Tables, figures, and diagrams must be photo copy ready for publication. All manuscripts should be submitted electronically to the editor.

Once received by the editor, manuscripts will be anonymously peer-reviewed by two editorial board members. The review process will take approximately 2 to 3 months. A letter acknowledging receipt of the manuscript will be sent to the first author, and upon review completion, a letter indicating the peer-review decision will be sent to the first author. Potential authors are encouraged to contact the editor to discuss ideas for contributions or determine if their manuscript is suitable for publication in *MLRV*.

**EDITOR**

David Walker
Northern Illinois University
Department of Educational Technology,
Research and Assessment
204A Gabel Hall
DeKalb, IL 60115
Phone: (815) 753-7886
Fax: (815) 753-9388
Email: dawalker@niu.edu

**ASSOCIATE EDITOR**

T. Mark Beasley
University of Alabama-Birmingham
Department of Biostatistics
School of Public Health
Ryals Public Health Bldg.
Birmingham, AL 35294
Phone: (205) 975-4957
Fax: (205) 975-2540
Email: MBeasley@ms.soph.uab.edu

**ORDER INFORMATION**

Cynthia Campbell, Managing Editor
Northern Illinois University
Department of Educational Technology, Research and Assessment
DeKalb, IL 60115
Phone: (815) 753-8471
Fax: (815) 753-9388
Email: ccampbell@niu.edu

# Improving the Accuracy of Parameter Estimation of Proportional Hazards Regression with Kernel Resampling

**Haiyan Bai**
University of Central Florida

The accuracy of parameter estimation of proportional hazards regression (PHR) has been a concern. To improve the accuracy of the estimation, the bootstrap has been used; unfortunately, prior research revealed inconsistent findings. The current study applies a new resampling method, the kernel resampling technique (KRT), to PHR. Two empirical datasets were employed to cross-validate and compare the accuracy and stability of the estimation results through multiple replications from KRT with those from the naïve bootstrap as well as the maximum likelihood method. The study results revealed that KRT outperformed the bootstrap and maximum likelihood method in estimating parameters of PHR. The application of KRT to PHR improved the accuracy of the parameter estimation.

Proportional hazards regression (PHR) (or Cox model) is a method for investigating the effect of several variables upon the time-specified outcome for an event to occur. PHR is most commonly applied in time-to-event studies (Cox, 1972). It assumes that the effects of the predictor variables upon survival are constant over time and are additive in one scale. If the assumptions are met, the PHR model can provide better estimates of survival probabilities and cumulative hazard than those provided by the Kaplan-Meier function; a log-rank test method for comparing survival curves in two or more groups (Cox). The PHR model has been used widely in medical studies and increasingly employed in a variety of disciplines under various rubrics, for example, "event-history analysis" in sociology (Allison, 1984), or "teacher survivals" and "student retention" in Education (cf. Adams, 1996; Adams & Dial, 1993; Plank, DeLuca, & Estacion, 2008). However, the accuracy of the estimation of the PHR model parameters has been a concern because estimating density functions or hazard rate functions is complicated (Burr, 1994). To improve the estimation accuracy from PHR models, the bootstrap method was implemented. Unfortunately, the effectiveness of this method is questionable due to the inconsistent findings of the performance of the bootstrap in the PHR model in prior research (Burr; Hjort, 1985; Singh, 1981).

Studies on the PHR model using the bootstrap are classified into two types: one for PHR model selections and the other for parameter estimation. The following is a brief review of these studies. Chen and George (1985) conducted a primary study using the bootstrap to investigate the variable selection in PHR, but they neither considered the prognostic implications for individuals nor discussed the accuracy of the parameter estimation. Extending Chen and George's study, Sauerbrei and Schumacher (1992) proposed a bootstrap-model selection procedure, but this study still focused on the model selection without considering the use of the bootstrap procedures directly in the parameter estimation. Altman and Andersen (1989) explored the confidence interval estimation of hazard ratios while conducting a bootstrap investigation of the stability of the PHR model, and the results revealed that the bootstrap intervals were graphically wider than those obtained from the original model. Hjort (1985) discussed using the bootstrap in the PHR model and found that the bootstrap procedure was first-order equivalent to the standard procedure. This was consistent with later research findings (e.g., Burr, 1994).

Burr (1994) presented a comprehensive study focusing on the methodological discussion about using the bootstrap procedures in PHR parameter estimation. This study compared bootstrap confidence intervals for the following three types of parameters in PHR: the regression parameters, the survival function at fixed time points, and the median survival time at fixed values of a covariate. The study revealed that the bootstrap-t intervals consistently outperformed both bootstrap percentile and hybrid interval estimations. The results also showed that the bootstrap did not improve the quality of regression parameter estimation on the asymptotic method, but it did improve the estimation of the survival function. Burr provided useful information to employ the bootstrap for parameter estimation in PHR; however, as Burr ( p. 1301) stated, "We would like to be able to recommend a single method appropriate for all parameters, but currently this is not possible." Therefore, further research in this area is desirable. The current study aims at exploring the potential improvement of parameter estimation in PHR using kernel resampling procedures.

## PHR Model

PHR for hazard rate was first introduced by Cox (1972) and it is often expressed as:

$$\lambda(t; \mathbf{X}) = \lambda_0(t) \exp(\mathbf{X}\beta),$$

where $\lambda(t; \mathbf{X})$ is the hazard (risk of event) at time $t$ with respect to covariate matrix $\boldsymbol{X}$. The parameter $\beta$ is a log relative risk and $\exp(\beta)$ is a relative risk of response. PHR is sometimes called relative risk regression, Cox regression, or Cox model. $\lambda_0(t)$ represents a reference point that depends on time, which is the "baseline" hazard (when covariates $\mathbf{X}$ are zero) just as $\beta_0$ denotes an arbitrary reference point in other types of regression analysis. PHR is a useful tool for studying patient survival time in medical studies, historical event in social science, company bankruptcy in economic investigations, and students' departure and teachers' survival in educational research.

## The Bootstrap and KRT

As a modern statistical technique, the bootstrap has been used in many procedures to improve the validity of studies through estimating more accurate standard errors (Efron & Tibshirani, 1993). The basic concept of the bootstrap is to construct empirical distributions of parameter estimates to assess the standard errors or confidence intervals to obtain improved statistical estimates. The bootstrap empirical distribution is usually constructed from bootstrap resamples, which are obtained through resampling from the original data with replacement. Existing studies have revealed the usefulness of the bootstrap in PHR (Gonzalez, Pena, & Delicado, 2010)

Kernel resampling technique (KRT) is an alternative resampling method which extends the bootstrap by sampling with random errors from Gaussian Kernels using a fixed bandwidth (Bai & Pan, 2009). KRT is a product of integrating the distribution theory into the smoothing technique. By design, KRT is fundamentally different from the bootstrap and its variant, the smoothed bootstrap, which requires researchers to find the optimal bandwidth to smooth the bootstrap distribution. KRT uses the Gaussian kernel technique to capture the covariance structure of multivariate data (Silverman, 1986; Simonoff, 1996).

The multivariate Gaussian kernel is defined as

$$K(\mathbf{x}) \sim N_d(\mathbf{X}_i, \mathbf{H}^2),$$

where $d$ is the number of variables, $\mathbf{X}_i$ ($i = 1, \ldots, n$) are multivariate data or a vector from a $d$-dimensional space $R_d$, $n$ is the number of cases, and $\mathbf{H}$ is the bandwidth matrix that can be chosen as an optimal one to minimize the *mean integrated square error* (MISE) (Silverman, 1986; Simonoff, 1996):

$$\mathbf{H}_{\mathrm{o}} = \left(\frac{4}{d+2}\right)^{1/(d+4)} \mathbf{S}^{1/2} n^{-1/(d+4)}$$

KRT has been successfully used in multiple regression models for increasing the accuracy of parameter estimation (Bai & Pan, 2009).

## Purpose of the Study

Considering the usefulness of the bootstrap in PHR (Gonzalez et al., 2010), the current study was proposed to use the KRT, an alternative to the bootstrap, to improve the accuracy of parameter estimation of PHR. The application of KRT to a multiple regression model has successfully provided more accurate parameter estimation than both naïve bootstrap and smoothed bootstrap (Bai, 2008; Bai & Pan, 2009); therefore, the purpose of the current study is to examine the performance of the application of KRT to PHR. Empirical data from education were employed for the methodological comparison through resampling at multiple numbers of replications to study the accuracy and stability of the estimation, while the medical data set was used to cross-validate the results. The findings from the applications of KRT to the PHR model using both data sets are compared with those from the bootstrap and the classical maximum likelihood (ML) method to determine which method is the most effective.

## Studies with Empirical Data

### *Study 1: Educational Data*

The study data were collected from an urban public school district in the Southeast region of the United States after obtaining Institutional Review Board approval at the author's university. This data set documented the departure records of 8462 students who departed from public schools between 2006 and 2010 including regular graduations. There were conceivably seven ways of departure in the data set that are listed in Table 1. In the current study, the PHR model was used to study the hazard rate for students' departure from public schools versus regular graduations.

For this study, two variables, student age and accumulated GPA, were used in the PHR as the covariates for the purpose of methodological evaluation of the performance of KRT application to PHR. The two variables were utilized as covariates (i.e., predictors in the PHR) because of their influential impacts on high school student departure based on the extant literature. Hauser, Simmons, and Pager (2000) stated that the likelihood of student departure increased with age in general; therefore, high school students tend to have a higher dropout rate than elementary and middle school students. The association between academic performance and dropout rates has been well studied (cf. Fagan & Pabon, 1990; Krohn, Thornberry, Collins-Hall, & Lizotte, 1995; Rumberger, 1987). Student academic performance is a major predictor of graduation rates and departure rates (Battin-Pearson et al., 2000). Prior studies examined and identified many influential factors or predictors for high school student departure including a variety of demographic, individual, family, and school characteristics (Neild, Stoner-Eby, & Furstenberg, 2008). However, for the focus of the current study on methodological discussions, only two major factors were included, student age and cumulative GPA, in the model to compare the accuracy of the statistics from different statistical procedures with no intention of providing any statistical inferences from the empirical example. The variables used in the model:

**Table 1.** The Numbers of Public High School Students' Departure and Regular Graduation

| Type of Departure | Students | Male/Female |
|---|---|---|
| Non-Public | 277 | 156/121 |
| Nowhere | 157 | 87/70 |
| Home School | 255 | 128/127 |
| Adult Program | 1770 | 1007/763 |
| Another District | 1548 | 847/701 |
| Out of State | 1257 | 638/619 |
| Regular Diploma | 3398 | 1662/1736 |

- Departure and Graduation: Move to non-public schools, go nowhere, home school, adult program, move to other in-state public schools, or move to other states versus obtain a regular diploma.
- Age: Student age was recorded at the time of departure.
- Cumulative GPA: The student GPA measure was the accumulated GPA since the semester a student entered the public high school.
- Survival Months: Months of staying in the public schools.

### *PHR on Student Departure Data*

A PHR model for the current study was defined as:

$$\text{Log}[\lambda(t;\, \mathbf{X})] = \log[\lambda_0(t)] + \beta\mathbf{X},$$

where **X** represents the predictors, *age* and *Weighted Cumulative GPA,* and $\beta$ is *the logarithm of the ratio* of the hazard rate for students belonging to departure versus regular graduation in the hazard function.

The PHR model was fitted with *age* and *Weighted Cumulative GPA* to estimate the hazard ratio. No evidence was found that students' departure in general depends on age (while adjusting only for *Weighted Cumulative GPA*) with $\chi^2$ = 0.13 ($p$ = 0.98) (see Table 2); therefore, age was eliminated in the final model.

**Table 2.** Estimates for Predictors

| Variable | *df* | β | *SE* | $\chi^2$ | $Pr > \chi^2$ | HazardRatio | 95% CI | |
|---|---|---|---|---|---|---|---|---|
| AGE | 1 | 0.02 | 0.06 | 0.13 | 0.72 | 0.980 | 0.88 | 1.09 |
| GPA | 1 | 0.58 | 0.06 | 83.61 | <0.001 | 0.562 | 0.50 | 0.63 |

**Table 3**. *Comparisons of Estimates, CIs, and Bias of PH Model with Asymptotic, Bootstrap, and KRT.*

| | Estimates | Replicates | Estimate | *SE* | CI(2.5%) | CI(97.5%) | Bias |
|---|---|---|---|---|---|---|---|
| Hazard Ratio | ML | | 0.5660 | 0.0333 | 0.5030 | 0.6360 | |
| | Bootstrap | 200 | 0.5694 | 0.0404 | 0.4983 | 0.6541 | 0.0034 |
| | | 500 | 0.5685 | 0.0421 | 0.4966 | 0.6604 | 0.0025 |
| | | 1000 | 0.5678 | 0.0421 | 0.4925 | 0.6593 | 0.0018 |
| | KRT | 200 | 0.5700 | 0.0167 | 0.5384 | 0.6048 | 0.0040 |
| | | 500 | 0.5678 | 0.0155 | 0.5353 | 0.5993 | 0.0018 |
| | | 1000 | 0.5672 | 0.0130 | 0.5426 | 0.5918 | 0.0012 |
| GPA | ML | | -0.5693 | 0.0598 | -0.6889 | -0.4498 | |
| | Bootstrap | 200 | -0.5656 | 0.0703 | -0.6966 | -0.4245 | 0.0038 |
| | | 500 | -0.5675 | 0.0736 | -0.6999 | -0.4149 | 0.0018 |
| | | 1000 | -0.5687 | 0.0737 | -0.7082 | -0.4166 | 0.0006 |
| | KRT | 200 | -0.5663 | 0.0289 | -0.6210 | -0.5073 | 0.0030 |
| | | 500 | -0.5686 | 0.0254 | -0.6209 | -0.5165 | 0.0007 |
| | | 1000 | -0.5680 | 0.0235 | -0.6127 | -0.5218 | 0.0014 |

### *Results of Study 1*

In order to conduct the methodological study, *Weighted Cumulative GPA* was selected to estimate the hazard ratio to examine the performance of KRT in PHR. Both KRT and the bootstrap procedures were used to obtain parameter estimates of *Weighted Cumulative GPA* and *the estimate of hazard ratio* for comparing the results. Two hundred, 500, and 1000 replications of both the bootstrap and KRT were conducted based on the original student departure data using the SAS macro (SAS Institute Inc., 2008) for parameter estimation and hazard ratio estimation of the PHR model.

From Table 3 we can see that the KRT estimates were comparable to the estimates for both *hazard ratio* and *β* for *Weighted Cumulative GPA* from the bootstrap and ML estimates; however, the standard errors from the KRT estimates for the hazard ratio and *β for Weighted Cumulative GPA* were systematically smaller than those from the bootstrap procedure and the Maximum Likelihood estimates across various numbers of replications with less biases in most cases. The confidence intervals (percentiles) for the estimates using the KRT procedure were narrower than those from both the bootstrap procedure and the Maximum Likelihood method.

### *Study 2: Cross-Validating Data*

To cross-validate the results of Study 1 for further evaluation on the performance of the application of KRT to PHR, a study was conducted using a large national medical data set, *Localized colon carcinoma 1975–1994*, as the original input data collected by the Institute for Statistical and Epidemiological Cancer. *Localized colon carcinoma 1975–1994* contains individual-level data of 6,274 patients diagnosed with localized tumors among 15,564 patients diagnosed with colon carcinoma in Finland 1975-1994 with follow-up to the end of 1995.

For the purpose of the methodological research focusing on comparison of the accuracy of the PHR model parameter estimations, the model selection is not discussed in the current study. With regard to the focus of the current study, the hazard ratio of mortality from colon cancer versus mortality due to other reasons was studied using the PHR model (i.e., mortality among the 6,274 patients diagnosed with localized tumors).

**Table 4**. Localized Stage

| Status | Patient *N* |
|---|---|
| 0: Alive | 2979 |
| 1: Dead: colon cancer | 1734 |
| 2: Dead: other | 1557 |
| 3: Lost to follow-up | 4 |

### *Study Variables*:

In the current study, four variables of interest were used:

- Gender: Gender is defined as male or female.
- Year of Diagnosis: The year diagnosed as having localized tumors.
- Survival Months: Months survived since the time of diagnosed localized tumors.
- Status: Vital status at last date of contact.

**Table 5**. Estimates for Predictors

| Variable | *df* | *β* | *SE* | $\chi^2$ | $Pr > \chi^2$ | Hazard Ratio | 95% CI | |
|---|---|---|---|---|---|---|---|---|
| Gender | 1 | -0.002 | 0.049 | 0.020 | 0.966 | 0.998 | 0.907 | 1.098 |
| Year85--94 | 1 | -0.232 | 0.049 | 22.258 | <0.001 | 0.793 | 0.720 | 0.873 |

### *PHR Model on Localized Colon Carcinoma Data*

A PHR model for the current study was defined as:

$$\text{Log}[\lambda(t; \mathbf{X})] = \log[\lambda_0(t)] + \beta\mathbf{X}$$

where **X** represents the predictors, *gender* and *Year of Diagnosis,* and *β* is *the logarithm of the ratio* of the hazard rate for patients belonging to the *mortality from colon cancer* group versus the *mortality group because of other reasons* in the hazard function. The PHR model was fitted with *gender* and *Year of Diagnosis* as predictors just for the purpose of the methodological discussion focus of this study. No evidence was found that mortality depends on gender while adjusting for year of diagnosis with $\chi^2 = .02$ (p = .966) (see Table 5). Therefore, *Year of Diagnosis* was selected to estimate the parameters and the hazard ratio for examining the performance of KRT in the Cox model with respect to the preliminary model fitting information. The KRT, the bootstrap, and the Maximum Likelihood method were used to obtain parameter estimates of *Year of Diagnosis* and *the estimate of hazard ratio* for comparing the results for examining the performance of KRT in the PHR model.

### *Cross-Validating Results from Study 2*

Table 6 presents the parameter and hazard ratio estimation from the PHR model with 200, 500, and 1000 replications of both the bootstrap and KRT and the results from the Maximum Likelihood applied to the original Localized Colon Carcinoma data. From Table 6 we can see that the KRT estimates were comparable to the estimates for both *hazard ratio* and *β for Year of Diagnosis* from the bootstrap and asymptotic estimates. With this in mind, it is evident that the standard errors from the KRT resamples were systematically smaller. The estimation biases were consistently less in most cases than those from both the bootstrap procedure and the conventional maximum likelihood method across of 200, 500, and 1000 replications. The 95% confidence intervals (percentiles) for the estimates using the KRT procedure were narrower than those from both the bootstrap procedure and the conventional maximum likelihood estimates. Methodologically, the evaluation results from the cross-validating sample were consistent with the results from Study 1 from the educational data; therefore, the findings of the KRT application to PHR model were cross-validated and proved to be replicable.

**Table 6**. Comparisons of Estimates, CIs, and Bias of Cox Model with the Conventional Asymptotic, Bootstrap, and KRT Methods

| | Estimates | Replicates | Estimate | *SE* | CI (2.5%) | CI (97.5%) | Bias |
|---|---|---|---|---|---|---|---|
| Hazard Ratio | ML | | 0.7930 | 0.0383 | 0.7200 | 0.8730 | |
| | Bootstrap | 200 | 0.7947 | 0.0401 | 0.7246 | 0.8748 | 0.0017 |
| | | 500 | 0.7936 | 0.0389 | 0.7194 | 0.8704 | -0.0012 |
| | | 1000 | 0.7932 | 0.0386 | 0.7201 | 0.8728 | -0.0004 |
| | KRT | 200 | 0.7938 | 0.0165 | 0.7651 | 0.8278 | 0.0007 |
| | | 500 | 0.7930 | 0.0154 | 0.7653 | 0.8240 | -0.0008 |
| | | 1000 | 0.7945 | 0.0153 | 0.7688 | 0.8243 | 0.0015 |
| Year of Diagnosis | ML | | -0.2310 | 0.0503 | -0.3222 | -0.1338 | |
| | Bootstrap | 200 | -0.2321 | 0.0492 | -0.3311 | 0.1334 | -0.0011 |
| | | 500 | -0.2324 | 0.0490 | -0.3293 | -0.1388 | -0.0003 |
| | | 1000 | -0.2329 | 0.0487 | -0.3284 | -0.1360 | -0.0005 |
| | KRT | 200 | -0.2296 | 0.0205 | -0.2683 | -0.1920 | 0.0033 |
| | | 500 | -0.2322 | 0.0180 | -0.2653 | -0.1961 | -0.0027 |
| | | 1000 | -0.2329 | 0.0202 | -0.2718 | -0.1919 | -0.0007 |

## Discussion and Further Study

Using data from different research areas, the findings from two studies provide strong evidence that the KRT outperformed both the bootstrap and the Maximum Likelihood method in the PHR parameter estimation. The application of KRT in PHR provided more accurate confidence interval estimation with narrower bands, smaller standard errors with less or comparable biases, and equivalent accurate point estimates. The KRT procedure produced stable estimation results across various replications. KRT application to PHR provides a solution for “a single method appropriate for all parameters” (Burr, 1994, p. 1301). This study produced preliminary results of the KRT application in PHR models for parameter estimation. The findings suggest that applications of KRT to PHR models improve the accuracy of parameter estimation for more valid statistical inference in survival research.

Future studies are desired to compare the results of other types of confidence interval estimation. In the current study, only empirical datasets were used to study the performance of the application of the KRT in a PHR model. Even though the cross-validating study provided strong evidence of the current study findings, a simulation study is expected to provide more information and further confirmation of the study results in terms of the stability of the findings under other conditions. Future studies should engage in (1) comparison of the results of other types of confidence interval estimation and (2) simulation studies with different data conditions (e.g., sample sizes or distributions) to explore the stability of the application results.

## Significance of the Study

In education, teachers’ survival, students’ dropout, and on-time graduation are all important factors influencing the quality of education. Understanding these factors is crucial for educators and educational administrators to work on effective solutions. PHR is an appropriate and effective statistical analytical tool for studies in such areas, and applications of KRT to PHR will improve the accuracy of parameter estimation to provide more valid statistical inference in educational research.

## Acknowledgement

Sincere thanks to Drs. Paul Dickman and Timo Hakulinen for providing the medical data.

## References

Adams, G. J. (1996). Using a PHR regression model to examine voluntary teacher turnover. *Journal of Experimental Education*, *643,* 267-285.

Adams, G. J., & Dial, M. (1993). Teacher survival a PHR regression model. *Education and Urban Society, 26*, 90-99.

Allison, P. D. (1984). *Event history analysis: Regression for longitudinal event data*. Beverly Hills, CA: Sage Publications.

Altman, D. G., & Andersen, P. K. (1989). Bootstrap investigation of the stability of a PHR regression model. *Statistics in Medicine, 8*, 771-783.

Bai, H. (2008). Kernel resampling to improve the performance of multiple regression with small samples. *Journal of Applied Statistical Science, 17*, 573-582.

Bai, H., & Pan, W. (2009). An application of a new multivariate resampling method to multiple regression. *Multiple Linear Regression Viewpoints*, *35*, 1-5.

Battin-Pearson, S., Newcomb, M.D., Abbott, R. D., Hill K. G., Catalano R. F., & Hawkins J. D. (2000). Predictors of early high school drop-out: a test of five theories. *Journal of Educational Psychology. 92*, 568–582.

Burr, D. A. (1994). A comparison of certain bootstrap confidence intervals in the PHR model. *Journal of the American Statistical Association*, *89*, 1290-1302.

Chen, S. H., & George, S. L. (1985). The bootstrap and identification of prognostic factors via PHR’s proportional hazards regression model. *Statistics in Medicine, 4*, 39–46,

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society Series B, 34,* 187–220.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall, Inc.

Fagan, J., & Pabon, E. (1990). Contributions of delinquency and substance use to school dropout among inner-city youth. *Youth & Society, 21,* 306-354.

Gonzalez, J. R., Pena, A. E., & Delicado, P. (2010). Confidence intervals for median survival time with recurrent event data. *Computational Statistics & Data Analysis*, 54, 78-89

Hauser, R. M., Simmons, S., & Pager, D. (2000). *High school dropout, race-ethnicity and social background from the 1970s to the 1990s*. Retrieved from http://www.ssc.wisc.edu/ ~hauser /Trends2001_03.pdf.

Hjort, H. (1985). *Boostrapping PHR's regression model*. Technical Report. Stanford University.

Krohn, M. D., Thornberry, T. P., Collins-Hall, L., & Lizotte, A. J. (1995). School dropout, delinquent behavior, and drug use: An examination of the causes and consequences of dropping out of school. In H. B. Kaplan (Ed.), *Drugs, crime, and other deviant adaptations: Longitudinal studies* (pp. 163-183). New York: Plenum Press.

Neild, R. C., Stoner -Eby, S., & Furstenberg, F. (2008). Connecting entrance and departure: The transition to ninth grade and high school dropout. *Education and Urban, 40*, 543-569.

Plank, S., DeLuca, S., & Estacion, A. (2008). High school dropout and the role of career and technical education: A survival analysis of surviving high school. *Sociology of Education*, *81*, 345-370.

Rumberger, R. W. (1987). High school dropouts: A review of issues and evidence. *Review of Educational Research*, *57*, 101-121.

SAS Institute Inc. (2008). *SAS/STAT user's guide, Version 9.2.* Cary, NC: SAS Institute Inc.

Sauerbrei, W., & Schumacher M. (1992). A bootstrap resampling procedure for model building: Application to the PHR regression model. *Statistics in Medicine*, *11*, 2093–2109.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman & Hall.

Simonoff, J. S. (1996). *Smoothing methods in statistics*. New York: Springer.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics, 9*, 1187-1195.

Send correspondence to: Haiyan Bai
University of Central Florida
Email: haiyan.bai@ucf.edu

# Application of CART, Neural Networks, and Generalized Additive Models: A Case Study

**W. Holmes Finch** **Mei Chang** **Andrew S. Davis**
Ball State University

Statistical prediction of an outcome variable using multiple independent variables is a common practice in the social and behavioral sciences. For example, neuropsychologists are sometimes called upon to provide predictions of pre-injury cognitive functioning for individuals who have suffered a traumatic brain injury. Typically these predictions are made using standard multiple linear regression models with several demographic variables (e.g., gender, ethnicity, education level) as predictors. Prior research has found conflicting evidence regarding the ability of such models to provide accurate predictions of outcome variables such as full-scale intelligence (FSIQ) test scores. The current study had two goals: 1) to demonstrate the utility of a set of alternative prediction methods that have been applied extensively in the natural sciences and business but which have not been frequently explored in the social sciences and 2) to develop models that can be used to predict premorbid cognitive functioning in preschool children. Prediction of Stanford Binet 5 FSIQ scores for preschool aged children is used to compare the performance of a multiple regression model with several of these alternative methods. Results demonstrate that classification and regression trees (CART) provided more accurate prediction of FSIQ scores than the more traditional regression approach. Implications of these results are discussed.

Clinical neuropsychologists are frequently required to determine if an individual has experienced changes in intellectual functioning resulting from a neurological insult such as a traumatic brain injury (TBI) or stroke. Accurate diagnosis and the determination of functional decline relies largely on a clinician's ability to compare current test performance to an estimate of premorbid (i.e., prior to injury) performance. It is common in clinical practice for neuropsychologists to use a discrepancy between a predicted and an obtained test score to assist in the determination of whether organic impairment or a progressive disease is present. Thus, an accurate estimation of premorbid intelligence is necessary to prevent errors such as under or overestimation of a patient's level of cognitive decline (Griffin, Mindt, Rankin, Ritchie, & Scott, 2002) and the availability of techniques demonstrating good validity and reliability for predicting premorbid intellectual functioning is a central concern of clinicians. When premorbid ability levels can be reasonably estimated, a diagnosis can be made with confidence and cognitive rehabilitation programs can be properly designed, monitored, and modified (Reynolds, 1997).

## Traditional Methods of Prediction

A variety of approaches have been proposed and developed for the estimation of premorbid ability estimation (PAE), including (a) historical achievement-based and standardized group assessment data (e.g., Baade & Schoenberg, 2004; Schinka & Vanderploeg, 2000); (b) "hold/don't hold tests" estimates (Blair & Spreen, 1989; Lezak, Howieson, Loring, Hannay, & Fischer, 2004); (c) best current performance estimates (Lezak, 1995); (d) demographic-based regression formulas (e.g., Barona, Reynolds, & Chastain, 1984); (e) combinations of demographic and actual performance data (e.g., Schoenberg, Lange, & Saklofske, 2007a; Schoenberg, Lange, & Saklofske, 2007b; Schoenberg, Lange, Saklofske, Suarez, & Brickell, 2008; Schoenberg, Scott, Duff, & Adams, 2002; Vanderploeg, Schinka, & Axelrod, 1996); and (f) current word reading ability tests (e.g., Blair & Spreen; Wechsler, 2003). However, each approach has been shown to have some limitations in application.

### *Using Multiple Linear Regression to Predict Premorbid IQ*

An alternative to these more ad hoc approaches to predicting premorbid IQ involves the use of multiple linear regression (MLR) to estimate IQ. Researchers in the field have developed models based on demographic variables in conjunction with performance on a task such as word reading or some comparable measure (Sellers, Burns, & Guyrke, 2002; Vanderploeg, Schinka, Baum, Tremont, & Mittenberg, 1998; Yeates & Taylor, 1997), while in other cases, only demographic variables were used. Crawford, Millar, and Milne (2001) found that for adults, the correlation between actual and predicted IQ, based on the demographic variables of education, socio-economic status and age, was 0.76, which was higher than that obtained through clinical judgment. A study focusing on predicting IQ for adolescents included variables such as gender, ethnicity, region of the U.S. in which the subject lived, age, and parental education level (Schoenberg et al., 2007a) and was found to provide predictions of FSIQ. Powell, Brossart and Reynolds (2003) compared the performance of two regression models of the demographic information estimation formula index (DI) (Barona et al., 1984) and the Oklahoma Premorbid Intelligence Estimate (OPIE) (Krull, Sherer, & Adams, 1995) for estimating premorbid cognitive functioning in adults. Both models are based on linear equations that predict cognitive functioning using demographic variables (age, gender, race, education, occupation, urban/rural residence and current performance) on Wechsler Adult Intelligence Scale-Revised (WAIS-R; Wechsler, 1983) Vocabulary and Picture Completion subtests. Their results demonstrated that the DI approach provided more accurate estimates of cognitive decline but were not as accurate when predicting FSIQ for individuals who did not suffer any brain injury. One issue with either of these approaches is that researchers must have access to all of the variables that serve as inputs to the standardized equations. Another issue is that regression-based estimates of premorbid IQ have been shown susceptible to error, particularly in outer ranges of intellectual function (Veiel & Koopman, 2001). In addition, the MLR model assumes a linear relationship exists between the outcome of interest and the predictors, unless the researcher explicitly includes a non-linear term. However, in many instances, it may be unclear whether a non-linear term should be included, and more importantly what type would be most appropriate.

Much of the focus in the prediction of premorbid IQ has been on adults with relatively little research devoted to predicting cognitive functioning in school-aged and younger children (Schoenberg et al., 2007a). However, some research has been conducted on the use of prediction equations based on only demographic variables with school-aged children (Pungello, Iruka, Dotterer, Koonce-Mills, & Reznick, 2009; Schoenberg et al., 2008; Schoenberg, Lange, Brickell, & Saklofske, 2007; Roberts, Bornstein, Slater, & Barrett, 1999; Sellers, Burns, & Guyrke, 1996). These studies included variables such as parental level of education, ethnicity, gender, age, region of the U.S. in which the child resided and parental occupation, and generally found that they could achieve an $R^2$ generally around 0.4 when predicting Full Scale IQ (FSIQ). Despite the publication of several studies, the problem of accurately predicting premorbid IQ, particularly in young children, has not been completely solved (Schoenberg et al., 2007b). Furthermore, prior work with adult populations has not definitively demonstrated linear models to be the universally most effective tool for predicting premorbid IQ scores, as was discussed previously. Therefore, alternative methods for prediction should be investigated in order to find the optimal tool(s) for the important task of obtaining reliable estimates of premorbid intellectual functioning for young children who have undergone a neurological insult and suffered from cognitive impairment.

It should be noted that while the focus of this research was on predicting IQ scores for young children using demographic variables, other recent examples using prediction in the social science literature include predicting school counselor evaluations of student performance (Granello, 2010), college student dropout (Nistor & Neubauer, 2010), impact of character education on social competence (Cheung & Lee, 2010) and parent training effectiveness (Lavigne, LeBailly, & Gouze, 2010) to name but a few. In the vast majority of this research some variant of linear regression was used to obtain predictions. However, because it is limited to linear or relatively simple non-linear forms, regression may not always be the optimal choice for this type of research (Berk, 2008).

The goals of this study were to describe some alternative methods of prediction that could be employed in the context of obtaining estimates of premorbid IQ in preschool-aged children. These methods, including Classification and Regression Trees, Neural Networks and Generalized Additive Models, have all been shown to be effective tools for prediction in simulation research, particularly in the presence of non-linear relationships between outcome and predictor variables (Chang, Finch, & Davis, 2011; Finch & Holden, 2010). Given their positive record of performance, and their relative novelty in the social science literature, it was hoped that a manuscript demonstrating how each could be used to solve a real world prediction problem would add to the quantitative methods literature. In addition, each of these alternative models presents the user with great flexibility in terms of model settings and the like, which can have tremendous impact on the final results of the analysis. Thus, in addition to demonstrating how these tools can be used in practice, a second goal of this manuscript is to discuss these various model settings and provide some guidance for their implementation. It should be noted that this paper is not intended to be a comprehensive review of these methods, but rather an introduction that should provide the interested researcher with the basic tools to conduct analyses with these modeling techniques. A number of more comprehensive works are referenced below for those who want to delve deeper (which we encourage wholeheartedly).

## Alternative Methods of Prediction

The following is discussion of a number of alternative approaches for prediction that may prove useful when a relationship between variables is not strictly linear. Given some of the problems discussed earlier in the traditional approach of using MLR for predicting premorbid IQ and because very little work has been done examining the prediction of IQ in very young children, these alternative methods may prove to be interesting tools for this task. After a description of these approaches, the results of a study demonstrating how to use these techniques for the task of predicting IQ will be presented.

### *Classification and Regression Trees (CART)*

CART (Breiman, Friedman, Olshen & Stone, 1984) arrives at predicted values for an outcome variable, *Y*, given a set of predictors by iteratively dividing individual members of the sample into ever more homogeneous groups, or nodes, based on values of the predictor variables. It can be thought of as a nonparametric approach because there are no assumptions regarding the underlying population from which the sample is drawn nor the form of the model linking the outcome and predictor variables. CART begins by placing all subjects into one node, or group, and then searches the set of predictors to find the value of one of those by which it can divide the observations into two new nodes, whose values on *Y* are as homogeneous as possible. For each of these new nodes, the predictors are once again searched for the optimal split by which the subjects can be further divided into ever more homogeneous nodes, where homogeneity is always based on the similarity of values of *Y*. This division of the data continues until a predetermined stopping point is reached, when further splits do not appreciably reduce the heterogeneity of the resulting nodes. At this point, the tree is complete and values of *Y* for new individuals can be obtained using the decision tree developed with this original training sample. The data for the new subject are fed into the tree, following the branches from node to node based on the values of the predictor variables until the individual is placed in one of the final, or terminal nodes. The predicted value for *Y* for each individual is then the mean for the training sample in this terminal node.

CART has a tendency to overfit the training data when developing the initial prediction tree (Berk, 2008), meaning that the final model may be too closely associated with the training sample to generalize well to other samples from the same population. In addition, trees produced by CART can sometimes contain terminal nodes with few individuals or terminal nodes that are very heterogeneous, which is characteristic of tree instability and an inability to generalize to the broader population (Hothorn, Hornik, & Zeileis, 2006). One commonly used method for ameliorating overfitting is the practice of pruning trees. This process, which is demonstrated in the results section below, involves the removal of terminal nodes that appear to provide little predictive power, and when included in the final model might lead to overfitting of the training sample. Pruning is not an automated process, and requires the direct involvement of the researcher, typically through an examination of results for multiple pruned trees, as is discussed below. In order to ascertain how much pruning is necessary, the researcher typically refers to a plot of the number of nodes by total model deviance. Total deviance for a tree corresponds to the sum of the sum of squared residuals within the terminal nodes (i.e., the sum of squared differences between the predicted and actual IQ scores in this example). The larger the deviance value, the greater the heterogeneity of scores within the terminal nodes, and the worse the CART solution. As terminal nodes are removed from the tree, the deviance will increase because more heterogeneous individuals are grouped together in the new, larger terminal node. The decision regarding the number of terminal nodes to retain in the pruned tree is based upon balancing this increased heterogeneity with a desire to have a more parsimonious tree, and one that generalizes better to other samples.

### *Neural Networks (NNET)*

Another prediction method examined in this study is Neural Networks (NNET) (e.g., Marshall & English, 2000; see Garson, 1998 for a more technical description of the method). NNETs create a prediction model for *Y* by using a search algorithm that examines a large number of subsets of the predictors, as well as interactions among them. Interactions and powers of the predictors (referred to as hidden layers) are computed in conjunction with weights that are akin to regression slopes. Main effects and hidden layers to be included in the final model are selected by the algorithm so as to minimize the least squares criterion used in standard linear regression (i.e., minimizing the sum of squared differences between the observed and predicted values). The hidden layers are generally much more complex than the two and three way interactions common in regression, involving several predictors and higher order versions of the predictors in a single interaction (Schumacher, Robner, & Vach, 1996). In addition, they are not specified *a priori* by the researcher, but instead are identified by the NNET algorithm based on their contribution to reducing the sum of squared residuals. In order to reduce the likelihood of finding locally optimal results that will not generalize beyond the training sample, random changes to the subset of predictors and interactions, not based on model fit, are also made. This method of obtaining optimal model fit is known as back-propagation, where the difference between actual and predicted outputs is used to find optimal weights for main effects and hidden layers. It is one of the most commonly used approaches in NNET applications (Garson, 1998).

A primary strength of NNET models is that they can identify complex interactions among the predictor variables in the hidden layer that other approaches may ignore (Marshall & English, 2000). For example, whereas in regression it is common to express the interaction of two predictors as their product, or to square or cube a single variable if the relationship with the response is believed not to be linear, a NNET will create hidden layers as weighted products of perhaps several variables, thus allowing the model to be influenced by the predictors to varying degrees. The result is that fairly obscure relationships between the outcome and predictors will be automatically identified without the researcher having to explicitly include them in the model.

Conversely, this ability to identify extremely specific models to fit the data presents a potential problem in that NNETs can substantially overfit the training data used to estimate the model (Schumacher et al., 1996). In order to combat this problem, most NNET models apply what is called weight decay, which penalizes (i.e., reduces) the largest weights found in the original NNET analysis, in effect assuming that very large weights are at least partially driven by random variation unique to the training data. The researcher typically sets the value of the decay parameter, λ, with larger values shrinking the weights for non-linear terms to a greater degree, and thus reducing their impact on the final model, hopefully ameliorating problems of overfitting. Generally speaking, the value of λ for a given problem is selected by examining the ability of the model to correctly predict the outcome variable for a cross-validation sample (Hastie, Tibshirani, & Friedman, 2001). In other words, the decay parameter value associated with the most accurate prediction in the cross-validation sample is the one that is selected.

Another choice that the researcher must make when using NNETs is the number of hidden layers that will be allowed in the final model. The larger the number of hidden layers that are permitted in the model, the more complex the model could become by incorporating higher order non-linear terms. Including more hidden layers has both positive and negative aspects. On the one hand, such models are better able to identify complex relationships among the variables, but on the other, they may lead to overfitting of the training sample. Thus, the researcher is advised to try multiple settings for the number of hidden layers and decide on the optimal setting for this parameter based on the accuracy of predictions for a cross-validation sample (Garson, 1998).

Researchers using NNETs also have control over the range of random starting values for the hidden layer weights that will be used in the model. The algorithm selects initial weight values randomly within a predefined range. The weights are then updated based upon the minimization of the least squares criterion. When the weights are near 0, hidden layers are deemphasized and the model becomes essentially linear in form. As these weights increase, the hidden layers play a greater role in determining predicted values for the outcome variable. In the initial model setup, the randomly selected starting values are typically drawn from a fairly restricted range near 0; e.g. -0.5 to 0.5 in the case of R. However, the researcher can change the range of these starting values and attempt to find the optimal setting based upon prediction accuracy for a cross-validation sample, if (s)he believes that hidden layers will play a more (or less) important role in the final model. It is important to note that if starting values for the weights are too large, the final model performance may be compromised due to overfitting (Hastie et al., 2001). Examples of manipulation of each of these settings are provided in the results section.

### *Generalized Additive Models (GAM)*

GAMs are a class of very flexible models that allow for the linking of *Y* with one or more predictor variables, using a wide variety of smoothing functions common in statistics. Each function is fit using a smoothing technique such as a thin plate spline (default in R), cubic spline or a P spline, with the goal of minimizing the penalized sum of squares criterion (Simonoff, 1996). For an excellent discussion of smoothing and splines, the reader is encouraged to refer to Keele (2008), and the aforementioned Simonoff . The penalized sum of squares (PSS) is based on the standard sum of squared residuals with a penalty applied for model complexity (i.e., the number of main effects and interactions included). The GAM algorithm works in an iterative fashion, beginning with the setting of the model intercept to the mean of *Y*. Subsequently, the smoothing function of choice is applied to each of the independent variables in turn, selecting the smoothed predictor that minimizes the PSS. This iterative process continues until the smoothing functions stabilize (i.e., the PSS cannot be appreciably reduced further), at which point final model parameter estimates are obtained. The optimal model is typically selected so as to minimize the Generalized Cross Validation (GCV) score, which is based on an approximation of a jackknifed cross validation check of the training data. Essentially, the GCV score is a measure of prediction accuracy based on a sum of squares value when jackknifing (leave one out) is used. However, it is constructed such that actual jackknifing, which can be quite laborious for large datasets, is not necessary. Smaller values of GCV are associated with more accurate and generalizable models. In addition to the GCV, selection of optimal GAMs is also aided by the popular Akaike Information Criterion (AIC), for which smaller values indicate better model fit.

As was the case for CART and NNET, overfitting of the data can also be a problem with GAMs. In order to avoid overfitting, the researcher using GAM can change the smoothing parameter, γ, which appears in the equation for the GCV score. By default γ is set to 1, where larger values correspond to identifying a smoother model as optimal for the data. Kim and Gu (2004) found that γ of approximately 1.4 was effective at correcting the overfitting problem while not compromising model fit. The researcher also has control over the actual smoothing spline to be used in developing the GAM, a selection of which was mentioned above. Indeed, different smoothing splines could be used with different predictor variables, or combinations of these variables. Finally, the researcher has the option of selecting what is essentially the complexity of the smoothing function through the dimension of the function used by the smoothing algorithm. Larger values of this parameter, *k*, allow for more degrees of freedom in the smoother, which corresponds to a potentially more complex smoothing function. Typically, this value is not set extremely high in order to avoid the possibility of overfitting. It should also be noted that in practice, the value of *k* frequently has a minor impact on the final performance of the model in terms of prediction accuracy (Wood, 2006).

## Current Study

As mentioned above, of particular interest in the current study is the investigation of how manipulating tuning parameters impacts each of the alternative methods (i.e., CART, NNET, and GAM). In much prior work, these methods have been studied using either default or generally recommended settings for these parameters. However, in actual practice analysis results can change with different values for these tuning parameters. Given that the general recommendation for these methods is to, in fact, try several values for these settings in order to find the optimal model (Hastie et al., 2001), the current study seeks to add to the literature regarding the most effective use of each approach by demonstrating how these settings can be manipulated in a common software package (R). It should be noted, however, that this work is not intended to represent a complete training in the use of these alternative prediction methods. Indeed, for each of them complete books are available to walk the researcher through planning, conducting and interpreting analyses. Rather, this study is intended to introduce interested readers to the basic sequence of using these methods for prediction, and to encourage further investigation of those methods that appear to be most appropriate for a given research scenario. There are many fine texts available for each approach, several of which are included in the references to this manuscript, and we encourage the interested reader to peruse these for a more complete discussion of the fine details of conducting each analysis. We are hopeful, however, that this paper will serve as a strong starting point from which a researcher interested in using one or more of these methods can begin their analysis with some confidence.

## Methodology

### *Participants and Procedures*

Participants for this study included 200 (*n* = 103 females; *n* = 97 males) preschool children. The sample was obtained from preschool facilities near a mid-sized city in the Midwest Demographic information for the total sample appears in Table 1. Only children who did not receive special education or related services, and whose parental consent was obtained, were included as participants. Once a signed parental permission form was obtained, the children were administered the Stanford-Binet Intelligence Scales – Fifth Edition (SB5; Roid, 2003) under standardized conditions by trained examiners. In addition, selected demographic data were also collected for all study participants.

***Instrumentation***

The SB5 (Roid, 2003) is an individually administered assessment of IQ appropriate for people between the ages of 2 and 85 years. It is theoretically grounded in the Cattell-Horn-Cattell (CHC) theory and intends to represent 5 CHC factors, including Fluid Intelligence (G*f*), Crystallized Knowledge (G*c*), Quantitative Knowledge (G*q*), Visual Processing (G*v*), and Short-Term Memory (G*sm*). The entire SB5 (5 verbal and 5 nonverbal subtests) was administered to the participants, and generated a Full Scale IQ (FSIQ). In relation to this study, the SB5 FSIQ score was used to indicate the children's comprehensive cognitive abilities. The SB5 was selected for use in this study because it is strongly grounded in CHC theory, has been normed for children as young as those used in this study, and has been shown to be a valid and reliable tool for such assessments.

**Table 1**. Descriptive Statistics for Total, Training and Cross-Validation Samples

| Variable | Total Sample | Training | Cross-Validation |
|---|---|---|---|
| Gender | | | |
| Male | 97 (48.5%) | 73 (48.7%) | 24 (48%) |
| Female | 103 (51.5%) | 77 (51.3%) | 26 (52%) |
| Ethnicity | | | |
| Caucasian | 124 (62%) | 93 (62%) | 31 (62%) |
| African-American | 49 (24.5%) | 38 (25.3%) | 11 (22%) |
| Hispanic/Latino | 2 (1%) | 2 (1.3%) | 0 (0%) |
| Bi-racial | 20 (10%) | 15 (10%) | 5 (10%) |
| Other | 3 (1.5%) | 0 (0%) | 3 (6%) |
| No report | 2 (1%) | 2 (1.3%) | 0 (0%) |
| Father's education | | | |
| Less than High school | 30 (15%) | 24 (16%) | 6 (12%) |
| High school/GED | 78 (39%) | 56 (37.3%) | 22 (44%) |
| 1-3 years of college | 44 (22%) | 35 (23.3%) | 9 (18%) |
| 4+ years of college | 29 (14.5%) | 21 (14%) | 8 (16%) |
| No report | 19 (9.5%) | 14 (9.4%) | 5 (10%) |
| Mother's education | | | |
| Less than High school | 16 (8%) | 13 (8.7%) | 3 (6%) |
| High school/GED | 48 (24%) | 35 (23.3%) | 13 (26%) |
| 1-3 years of college | 89 (49.5%) | 65 (43.3%) | 24 (48%) |
| 4+ years of college | 39 (19.5%) | 31 (20.7%) | 8 (16%) |
| No report | 8 (4%) | 6 (4%) | 2 (4%) |
| Age (months) Mean | 58.86 (5.38) | 59.73 (5.50) | 60.28 (5.04) |
| FSIQ (SD) | 98.10 (11.81) | 98.29 (11.17) | 97.54 (13.67) |

***Prediction Models***

The outcome variable of interest was the FSIQ from the SB5, while the predictors included years of education each for mother and father, and the child's age. These predictors were selected because they are typically available for any subject for who predicted IQ is required, and will not be impacted by a CNS injury. They have also been used in prior IQ prediction studies (Sellers et al., 1996). The models used to predict FSIQ with these demographic variables included MLR as well as CART, NNET, and GAM. All analyses were carried out using the R software package (R Development Core Team, 2007). These prediction methods were selected because they have been demonstrated in prior research to be effective tools in predicting continuous outcome variables (Chang et al., 2011; Finch &Holden, 2010).

In order to assess the predictive accuracy of the models, the original sample of 200 subjects was randomly divided into training (*N*=150) and cross-validation samples (*N*=50). For each method, the training sample was used to estimate a predictive model, which was in turn applied to the cross-validation sample to obtain predicted values for FSIQ. Prediction accuracy for the cross-validated sample was assessed through the bias of the predicted IQ: $\text{Bias} = \theta_{\text{Actual}} - \theta_{\text{Predicted}}$ and the Root Mean Square Error (RMSE) of the predictions for the cross-validation sample:

$$\text{RMSE} = \sqrt{\frac{\sum\left(\theta_{\text{Actual}} - \theta_{\text{Predicted}}\right)^2}{n}}\ .$$

Bias serves as a measure of the estimation accuracy, while RMSE reflects both accuracy and precision of the predicted values. In general, results with lower bias and lower RMSE can be viewed as better fitting.

## Results

Following is a description of FSIQ prediction results for the cross-validation sample using CART, NNET, and GAM, along with ordinary least squares (OLS) regression, which will serve as the baseline for comparison with the alternative methods. The R commands necessary to run these analyses appear in italics in the text. Table 2 contains the bias and RMSE results for each model.

**Table 2**. RMSE and Bias Values for Cross-Validation Sample: OLS, CART, GAM and NNET Models

| Model | RMSE | Bias |
|---|---|---|
| OLS | 15.0567 | -6.13 |
| CART, 13 nodes | 15.2883 | -6.24 |
| CART, 10 nodes | 15.46269 | -6.43 |
| CART, 8 nodes | 15.19276 | -6.51 |
| NNET, 2 hidden layers | 15.05665 | -6.13 |
| NNET, 5 hidden layers | 15.05665 | -6.13 |
| NNET, 10 hidden layers | 16.16655 | -6.11 |
| NNET, 20 hidden layers | 14.71749 | -5.2 |
| NNET, 2 hidden layers, decay=0.5 | 16.43546 | -6.08 |
| NNET, 5 hidden layers, decay=0.5 | 15.48161 | -6.35 |
| NNET, 10 hidden layers, decay=0.5 | 15.08878 | -5.18 |
| NNET, 20 hidden layers, decay=0.5 | 15.31007 | -7.06 |
| NNET, 2 hidden layers, decay=0.75 | 16.16945 | -6.34 |
| NNET, 5 hidden layers, decay=0.75 | 15.27232 | -4.96 |
| NNET, 10 hidden layers, decay=0.75 | 15.67306 | -5.88 |
| NNET, 20 hidden layers, decay=0.75 | 15.67511 | -6.67 |
| NNET, 2 hidden layers, range=-1 to 1 | 15.05665 | -6.13 |
| NNET, 5 hidden layers, range=-1 to 1 | 15.05665 | -6.13 |
| NNET, 10 hidden layers, range=-1 to 1 | 15.27986 | -6.59 |
| NNET, 20 hidden layers, range=-1 to 1 | 15.05665 | -6.13 |
| GAM, thin plate | 15.06799 | -5.49 |
| GAM, cubic | 15.87559 | -5.95 |
| GAM, thin plate, g=1.4 | 15.06799 | -5.49 |
| GAM, cubic, g=1.4 | 15.56384 | -6.53 |

### *CART*

CART found a tree with 13 terminal nodes, and underestimated FSIQ in the model was 15.29. In terms of conducting analysis in R, the *library* command loads the tree library (which we would have previously installed in our version of R), and the *iq.cart<-tree(IQ~age+fathered+mothered)* command creates the prediction tree and saves it in the R object *iq.cart*. The descriptive output, produced by the *summary* command, appears below. This output shows us the deviance value for the tree (Residual mean deviance), where larger values indicate a greater difference in observed and predicted FSIQ for the training sample.

```
library(tree)
iq.cart<-tree(IQ~age+fathered+mothered)
summary(iq.cart)
```

```
Regression tree:
tree(formula = IQ ~ age + fathered + mothered)
Number of terminal nodes:  13
Residual mean deviance:  83.15 = 11390 / 137
Distribution of residuals:

Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-34.330  -6.375   1.255   0.000   6.176  20.860
```

In order to determine how much pruning should be done, we used the following set of commands to create the graph in Figure 1, showing the relationship between the deviance and the number of terminal nodes.

```
iq.cart.prune<-prune.tree(fsiqtrain.cart)
plot(iq.cart.prune)
```

Moving right to left on the x-axis, we can see that the first big increase in deviance occurs between 10 and 9 terminal nodes. Thus, we may elect to fit a tree with only 10 terminal nodes rather than the original 13, using the following commands in R. Note that the subcommand *best=10* requests that the 10 terminal node tree with the smallest deviance be selected. The deviance for this tree was 85.08, which is not much larger than the 83.15 for the original 13 node tree, suggesting that losing the three weakest terminal nodes did not substantially damage model fit for the training sample. In addition, the mean bias and RMSE values in Table 2 for the 10 node tree were very similar to those for the full tree.

```
 iq.cart.prune10<-prune.tree(iq.cart,best=10)
 summary(iq.cart.prune10)

 Regression tree:
 snip.tree(tree = iq.cart, nodes = c(4, 22, 15))
 Number of terminal nodes:  10
 Residual mean deviance:  85.08 = 11910/140
 Distribution of residuals:

 Min. 1st Qu.  Median    Mean 3rd Qu.   Max.
-34.330  -6.000   1.160   0.000   5.784  24.150
```

Likewise, we produced a tree with 8 terminal nodes, which also had very similar bias and RMSE values to the other two trees.

```
 iq.cart.prune8<-prune.tree(iq.cart,best=8)
 summary(iq.cart.prune8)

 Regression tree:
 snip.tree(tree = iq.cart, nodes = c(4, 15, 11))
 Number of terminal nodes:  8
 Residual mean deviance:  88.59 = 12580/142
 Distribution of residuals:
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-34.330  -6.281   1.465   0.000   5.899  20.380
```

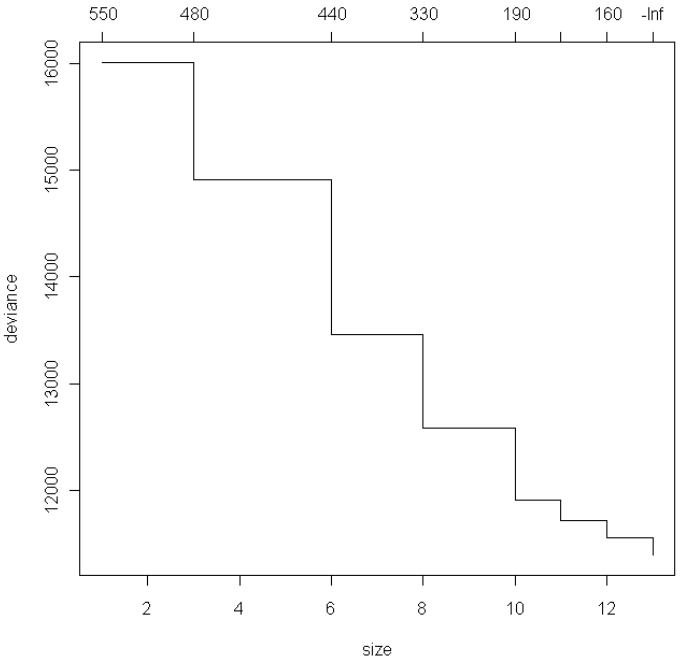

***Figure 1***.Total tree deviance by number of terminal nodes

Taken together, these results would suggest that for the purposes of predicting FSIQ for the cross-validation sample, the 8 terminal node tree was just as effective as the full 13 node tree. Furthermore, all of the tree models produced generally comparable results to those for the other alternative models included in this study.

***NNET***

Of the three alternative modeling techniques examined here, NNET has the largest number of potential settings that a researcher can change. In this study, we estimated a large number of NNET models, results of which are presented in Table 2 with regard to prediction accuracy for the cross-validation sample. In addition, compared to the other two approaches featured here, the output from a NNET analysis is not particularly informative regarding either model fit or the actual nature of the model itself. Indeed, the most useful information regarding the fit of a NNET model comes from its ability to accurately predict the outcome variable for the cross-validation sample. As an initial example, the following are the commands for running a basic NNET with 2 hidden layers, and default weight decay of 0 and range of random weight starting values from -0.5 to 0.5.

```
 library(nnet)
iq.nnet2<-
nnet(IQ~age+fathered+mothered,size=2,
linout=T,skip=T)

# weights:  14
initial  value 3088737.481193
iter  10 value 28190.551478
iter  20 value 21395.190892
iter  30 value 21354.815727
iter  40 value 21298.496237
iter  50 value 21296.878978
iter  50 value 21296.878961
iter  50 value 21296.878961
final  value 21296.878961
converged
```

The library to be used in this case is nnet, which was previously installed in our version of R. The *linout=T* command is necessary when the outcome variable is continuous, as is the case for FSIQ. The default setting for NNET in R is a categorical outcome, leading to a model corresponding to logistic regression. The *skip=T* subcommand allows for the inclusion of main effects as well as hidden layers in the final model. If this command is set to *F*, there will be no weights directly linking each of the main effects to the outcome variable. We can view the weights by typing the command to the right.
In this table, b is the intercept, i1, i2, and i3 are the three independent variables, h1 and h2 are the hidden layers, and o is the outcome variable. Thus, we can see that the weight for variable i1, age, to the first hidden layer is -0.52, while the weight of father's education (i2) to this hidden layer is 0.55, and so on. Finally, looking at the last line, we can see that the variable with the largest direct weight to the outcome variable is mother's education (i3), with a value of 1.69. At the same time, neither age nor father's education had a strong direct relationship with FSIQ. In addition, hidden layer 1 (h1), which was dominated primarily by an interaction of age and father's education, had a somewhat stronger impact on FSIQ than did hidden layer 2. It should be noted that from this output, we do not have an indication of which of these weights could be statistically significant in the classic hypothesis testing sense, nor do we know how well this particular model fits the training data. We can, however, examine its fit to the cross-validation data in Table 2, and see that it performs similarly to the CART models described earlier. Finally, we can set the number of hidden layers with the *size* subcommand, as below for a NNET with 10 hidden layers.

```
 summary(iq.nnet2)
a 3-2-1 network with 14 weights
options were - skip-layer connections  linear output units
 b->h1 i1->h1 i2->h1 i3->h1
  0.55  -0.52   0.55  -0.26
 b->h2 i1->h2 i2->h2 i3->h2
 -0.69  -0.57  -0.61  -0.29
 b->o h1->o h2->o i1->o i2->o i3->o
93.04  0.37 -0.21 -0.01  0.12  1.69
```

```
iq.nnet10<-nnet(IQ~age+fathered+mothered,size=10,linout=T,skip=T)
```

In order to change the value of the weight decay (*λ*) parameter to 0.5, we would use the following command in R. Remember that larger values of *λ* tend to shrink the size of the weights for the hidden layers. To select the optimal *λ* value various values would be tried and their relative impact determined through an examination of the accuracy of results for the cross-validation sample (see Table 2).

```
iq.nnet2.decay<-nnet(IQ~age+fathered+mothered,size=2,decay=.5,linout=T,skip=T)
```

In addition to manipulating the number of hidden nodes and the weight decay parameter, the researcher also has the option of changing the range of random starting values for the weights. By default, R draws these weights randomly from between -0.5 and 0.5. However, as discussed above, the range of starting values can be changed in order to reflect *a priori* beliefs regarding the importance of the hidden layers. A larger range of starting values for the weights allows for the possibility that the hidden layers are more important than if the range of starting values is tightly clustered near 0. As an example, the following R commands include 2 hidden layers, a *λ*value of 0 and the range of starting values between -1 and 1.

```
iq.nnet2.range1<-nnet(IQ~age+fathered+mothered,size=2,linout=T,skip=T,rang=1)

# weights:  14
initial  value 2588587.290328
iter  10 value 15631.624853
final  value 15597.617140
converged

summary(iq.nnet2.range1)

a 3-2-1 network with 14 weights
options were - skip-layer connections  linear output units
 b->h1  i1->h1  i2->h1  i3->h1
 -0.36    0.40    0.68    0.81
 b->h2  i1->h2  i2->h2  i3->h2
 -2.38 -142.77   -6.38  -11.60
  b->o   h1->o   h2->o   i1->o   i2->o   i3->o
 47.02   46.02  -86.92   -0.01    0.12    1.69
```

We can see that the weights for the hidden layers in this model are generally larger those of the 2 hidden node with starting value range from -0.5 to 0.5 above. Bias and RMSE results for a number of NNET models with the cross-validation data appear in Table 2. It appears that the NNET model with 20 hidden layers provided the best fit to the cross-validation sample, across all of the models examined in this study.

### *GAM*

In order to build a GAM with the thin plate smoothing spline (the default in R), we would use the following command sequence. GAM is included in the *mgcv* library of R functions that we would have previously installed in our version of R. The actual *gam* command used here sets the dimensions of the basis function, *k*, equal to 6 for both mother's and father's education. The reason for this is that the number of observed values for these variables was only 6, and there cannot be more dimensions to the basis function than there are values of the variable. The default value in R is 10. We will note the GCV score of 111.6503 and compare it with the GCV scores for alternative GAMs below.

```
library(mgcv)
iq.gam.tp<-gam(IQ~s(age, bs="tp")+s(fathered, k=6, bs="tp")
+s(mothered, k=6, bs="tp"),family=gaussian)
iq.gam.tp
```

```
        Family: gaussian
        Link function: identity

        Formula:
        IQ ~ s(age, bs = "tp") + s(fathered, k = 6, bs = "tp")
        + s(mothered,     k = 6, bs = "tp")

        Estimated degrees of freedom:
        1 1 1  total = 6

        GCV score: 111.6503
```

In order to use an alternative smoother, such as the cubic spline, we would change the previous commands as follows, replacing *tp* with *cs*.

```
iq.gam.cs<-gam(IQ~s(age, bs="cs")+s(fathered, k=6, bs="cs")+s(mothered,
k=6, bs="cs"),family=gaussian)
iq.gam.cs
```

```
        Family: gaussian
        Link function: identity

        Formula:
        IQ ~ s(age, bs = "cs") + s(fathered, k = 6, bs = "cs") + s(mothered,
           k = 6, bs = "cs")

        Estimated degrees of freedom:
        4.4562e+00 3.0283e+00 6.7655e-10  total = 10.48453

        GCV score: 105.3774
```

Note that the GCV score for the cubic spline GAM is somewhat lower than that of the thin plate spline model, indicating that it provides a better fit to the data.

We can change the degree of smoothing itself by setting $\gamma$=1.4, for example here with the cubic spline smoother. In this instance, the GCV actually increased from the cubic spline model with the default $\gamma$=1, suggesting that this latter model does not provide as good a fit to the data.

```
iq.gam.cs.gamma14<-gam(IQ~s(age, bs="cs")+s(fathered, k=6, bs="cs")
     +s(mothered, k=6, bs="cs"),family=gaussian,gamma=1.4)
iq.gam.cs.gamma14
```

```
Family: gaussian
Link function: identity

Formula:
IQ ~ s(age, bs = "cs") + s(fathered, k = 6, bs = "cs")
   + s(mothered,     k = 6, bs = "cs")

Estimated degrees of freedom:
4.8614e-03 6.6852e-03 7.6595e-06  total = 3.011554

GCV score: 107.4748
```

```
AIC(iq.gam.tp)
[1] 1129.699
AIC(iq.gam.cs)
[1] 1125.545
AIC(iq.gam.cs.gamma14)
[1] 1126.74
```

In addition to the GCV, it is also possible to compare the fit of GAMs using the AIC, which can be obtained with the R commands to the right. Remember that using this criterion, the optimal model is the one with the smallest AIC value, which in this case is the cubic spline model with $\gamma$ =1, which was also the best fitting model based on the GCV score.

We estimated models for GAMs with cubic and thin plate splines and for both $\gamma$=1 and $\gamma$=1.4. Results for these in terms of prediction of the cross-validation sample appear in Table 2. The GAMs with a thin plate spline and $\gamma$=1 or $\gamma$=1.4 provided the lowest bias and RMSE values of the GAMs, despite the fact that based on the GCV and AIC values, the cubic splines appeared to be slightly better. In addition, the GAMs had slightly better RMSE and bias values than both OLS and the CART models, and performed comparably to the NNETs, though as noted above the NNET model with 20 hidden layers provided the best fit for the cross-validation sample. While not at all dramatic, the small discrepancy in terms of which model appears to be best fitting for the training and cross-validation samples does suggest the need for extra care with regard to the problem of overfitting when using these complex modeling techniques. In short, it appears that the cubic spline models may have overfit the training data somewhat, when compared with the thin plate splines.

## Discussion

Prediction is an important aspect of statistical practice in psychology and the other social sciences, which had traditionally been done using standard MLR. For example, prior studies in the area of premorbid IQ prediction have generally been based on MLR models with adolescent and older populations. However, it has been argued that the regression based approach may not always be optimal (Veiel & Koopman, 2001), nor has it been shown that such predictions can be accurately made for preschool age children. Problems with relying too completely on strictly linear model forms are not limited to the prediction of premorbid IQ. In recent years, a number of alternative prediction modeling methods have become more widely available in popular software packages such as R. While offering the promise of greater prediction accuracy, however, these more complex models also present the researcher with a sometimes bewildering array of tuning parameters that must be set in order for them to perform optimally. Thus, a primary goal of this study was to demonstrate how one might use these modern methods of prediction in practice with a real prediction problem. The methods featured here were selected because they have been shown to be effective in both simulation and applied research, as noted above.

To briefly summarize the results of this study, it appears that in terms of the outcome variables included here, bias and RMSE all of the methods provided generally comparable predictions of FSIQ, with the NNET model with 20 hidden nodes being somewhat more accurate than the others. This approach demonstrated both the least bias and the lowest RMSE value. MLR, CART and GAM performed very similarly to one another. Given that prior Monte Carlo simulation work has shown that linear models such as MLR perform poorly for prediction when there are a number of interactions among predictor variables in the population (Garson, 1998), we may be able to infer from the relative success of MLR in this case that the relationships among these predictors and FSIQ are largely linear.

In conclusion, we hope that this manuscript contributes to research practice by demonstrating three proven and effective methods of prediction that are available in situations where it is known or believed that the relationships between predictor and outcome variables is not linear in nature. Furthermore, we have attempted to demonstrate how one can optimize these models through the judicious use of tuning parameters and/or pruning, in the case of CART. In the final analysis, the selection of optimal models and settings should be based on their accuracy with respect to a cross-validation sample. In all three cases, there is a distinct risk of overfitting the training data, which results in models that are not generalizable to the broader population. Therefore, simply assessing model fit for the training sample will likely leave the researcher with a less than optimal model for practice. However, systematically altering the tuning parameters and examining their impact on prediction for the cross-validation sample can result in selection of a model that provides the most accurate predictions possible for samples from across the population.

## References

Baade, L. E., & Schoenberg, M. R. (2004). A proposed method to estimate premorbid intelligence utilizing group achievement measures from school records. *Archives of Clinical Neuropsychology*, *19*, 227-243.

Barona, A., Reynolds, C.R., & Chastain, R. (1984). A demographically based index of premorbid intelligence for the WAIS-R. *Journal of Consulting and Clinical Psychology, 52,* 885-887.

Berk, R. A. (2008). *Statistical learning from a regression perspective*. New York: Springer.

Blair, J. R., & Spreen, O. (1989). Predicting premorbid IQ: A revision of the National Adult Reading Test. *The Clinical Neuropsychologist, 3,* 129-136.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C.J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Chang, M., Finch, W. H., & Davis. A. S. (2011, April). *The prediction of intelligence in preschool children using alternative models to regression*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Cheung, C-K, & Lee, T-Y. (2010). Improving social competence through character education. *Evaluation and Program Planning, 33,* 255-263.

Crawford, J. R., Millar, J., & Milne, A. B. (2001). Estimating premorbid IQ from demographic variables: A comparison of regression equation vs. clinical judgment. *British Journal of Clinical Psychology, 40,* 97-105.

Finch, W. H., & Holden, J. E. (2010). Prediction accuracy: A Monte Carlo comparison of several methods in the continuous variable case. *Multiple Linear Regression Viewpoints, 36*, 13-28.

Garson, G. D. (1998). *Neural networks: An introductory guide for social scientists*. London: SAGE Publications.

Granello, D. H. (2010). Cognitive complexity among practicing counselors: How thinking changes with experience. *Journal of Counseling & Development, 88,* 92-100.

Griffin, S. L., Mindt, M. R., Rankin, E. J., Ritchie A. J., & Scott, J. G. (2002). Estimating premorbid intelligence: Comparison of traditional and contemporary methods across the intelligence continuum. *Archives of Clinical Neuropsychology, 17*, 497–507.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference and prediction*. New York: Springer.

Hothorn, T., Hornick, K., Zeilleis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of computational and graphical statistics, 15,* 651-674.

Keele, L. (2008). *Semiparametric regression in the social sciences*. Hoboken, NJ: John Wiley & Sons.

Kim, Y. J. & Gu, C. (2004). Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *Journal of the Royal Statistical Society, Series B, 66,* 337-356.

Krull, K. R., Sherer, M., & Adams, R. L. (1995). A comparison of indices of premorbid intelligence in clinical populations. *Applied Neruopsychology, 2,* 35-38.

Lavigne, J. V., LeBailly, S. A., & Gouze, K. R. (2010). Predictors and correlates of completing behavioral parent training for the treatment of oppositional defiant disorder in pediatric primary care. *Behavior Therapy, 41,* 198-211.

Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford UniversityPress.

Lezak, M. D., Howieson, D. B., Loring, D. W., Hannay, H. J., & Fischer, J. S. (2004). *Neuropsychological assessment* (4th ed.). New York: Oxford University Press.

Marshall, D. B., & English, D. J. (2000). Neural Network modeling of risk assessment in child protective services. *Psychological Methods*, *5,* 102-124.

Nistor, N., & Neubauer, K. (2010). From participation to dropout: Quantitative participation patterns in online university courses. *Computers & Education, 55,* 663-672.

Powell, B. D., Brossart, D. F., & Reynolds, C. R. (2003). Evaluation of the accuracy of two regression-based methods for estimating premorbid IQ. *Archives of clinical neuropsychology, 18,* 277-292.

Pungello, E. P., Iruka, I. U., Dotterer, A. M., Koonce-Mills, R., & Reznick, J. S. (2009). The effects of socioeconomic status, race, and parenting on language development in early childhood. *Developmental Psychology, 45,* 544-557.

R Development Core Team. (2007). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Reynolds, C. R. (1997). Postscripts on premorbid ability estimation: Conceptual addenda and a few words on alternative and conditional approaches. *Archives of Clinical Neuropsychology, 12,* 769-778.

Roberts, E., Bornstein, M. H., Slater, A. M., & Barrett, J. (1999). Early cognitive development and parental education. *Infant and Child Development, 8*, 49-62.

Roid, G. H. (2003). *Stanford-Binet intelligence scales (5th ed.)*: Technical Manual. Itasca, IL: Riverside Publishing.

Schinka, J. A., & Vanderploeg, R. D. (2000). Estimating premorbid level of functioning. In R. D. Vanderploeg (Ed.), *Clinician's guide to neuropsychological assessment* (2nd ed.) Mahwah, NJ: Lawrence Erlbaum Associates.

Schoenberg, M. R., Lange R. T., Brickell T. A., & Saklofske D. H. (2007). Estimating premorbid general cognitive functioning for the American WISC-IV: Demographic and current performance approaches. *Journal of Child Neurology, 22*, 379-388.

Schoenberg M. R., Lange R. T., & Saklofske D. H. (2007a). Estimating premorbid FSIQ scores for the Canadian WISC-IV: Demographic and combined estimation procedures. *Journal of Clinical and Experimental Neuropsychology, 29,* 867-878.

Schoenberg, M. R., Lange, R. T., & Saklofske, D. H. (2007b). A proposed method to estimate full scale intelligence quotient (FSIQ) for the Canadian Wechsler Intelligence Scale for Children—Fourth Edition (WISC-IV) using demographic and combined estimation procedures. *Journal of Clinical and Experimental Neuropsychology*, *29,* 867-878.

Schoenberg, M. R., Lange, R. T., Saklofske, D. H., & Suarez, M. (2008). Validation of the child premorbid intelligence estimate method to predict premorbid Wechsler Intelligence Scale for Children—Fourth Edition Fall Scale IQ among children with brain injury. *Psychological Assessment, 20,* 377-384.

Schoenberg, M. R., Scott, J. G., Duff, K., & Adams, R. L. (2002). Estimation of WAIS-III intelligence from combined performance and demographic variables: Development of the OPIE-3. *The Clinical Neuropsychologist, 16*, 426-438.

Schumacher, M., Robner, R. & Vach, W. (1996). Neural networks and logistic regression: Part I. *Computational Statistics and Data Analysis*, *21*, 661-682.

Sellers, A. H., Burns, W. J., & Guyrke, L. (2002). Differences in young children's IQs on the Wechsler Preschool and Primary Scale of Intelligence-Revised as a function of stratification variables. *Applied Neuropsychology, 9,* 65-73.

Sellers, A. H., Burns, W. J., & Guyrke, L. (1996). Prediction of premorbid intellectual functioning of young children using demographic information. *Applied Neuropsychology, 9,* 65-73.

Simonoff, J. S. (1996). *Smoothing methods in statistics.* New York: Springer.

Vanderploeg, R. D., Schinka, J. A., & Axelrod, B. N. (1996). Estimation of WAIS-R premorbid intelligence: current ability and demographic data used in a best-performance fashion. *Psychological Assessment, 8,* 404-411.

Vanderploeg, R. D., Schinka, J. A., Baum, K. M., Tremont, G., & Mittenberg, W. (1998). WISC-III premorbid prediction strategies: Demographic and best performance approaches. *Psychological Assessment, 10(3),* 277-284.

Veiel, H. O. F., & Koopman, R. F. (2001). The bias in regression-based indices of premorbid IQ. *Psychological Assessment, 13,* 356-368.

Wechsler, D. (1983). *Wechsler Adult Intelligence Scale-Revised (WAIS-R)*. Technical Manual. New York: Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children—Fourth Edition*. San Antonio, TX: Psychological Corporation.

Wood, S.N. (2006). *Generalized Additive Models: An introduction with R.* Boca Raton, FL: Chapman & Hall.

Yeates, K. O., & Taylor, H. G. (1997). Predicting premorbid neuropsychological functioning following pediatric traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology, 19,* 825-.

Send correspondence to: W. Holmes Finch
Ball State University
Email: whfinch@bsu.edu

# Unbalanced Sampling Effect on the Power at Level-1 in the Random Coefficient Model

**Bonnie J. Steele**
Colorado Mountain College

**Daniel J. Mundfrom**
Eastern Kentucky University

**Jamis Perrett**
Texas A & M University

Researchers often disregard the potentially negative effects of unbalanced sampling on power estimates when using multilevel models. The purpose of this study was to investigate the effects that unbalanced sampling had on the estimated level-one power in multilevel random coefficient models. Twelve combinations of three effect sizes (0.5, 0.8, and 1.0) and four intraclass correlations (0.2, 0.1, 0.05, and 0.01) were investigated with each of three sampling ratios (0.25:0.75, 0.20:0.80, and 0.15:0.85) and three sample sizes (200, 500, and 800) to compare the effects that the different sampling ratios had on the level-1 power in the random coefficient model. Results indicated that as sampling ratios changed from 0.25:0.75, to incrementally a larger unbalanced sampling ratio of 0.15:0.85, the estimated power was lower in almost every case. This effect was more pronounced for the smaller sample sizes. Fourteen cases displayed differences larger than 5% in aggregate power estimates.

Hierarchical Linear Modeling (HLM) is a derivative or extension of the standard regression model adapted to address the problem of multilevel data, which allows the researcher to confront restrictions previously imposed by single-level analyses (Heck & Thomas, 2000). As a widely utilized technique, HLM has a rich literature containing recommendations regarding the appropriate balanced sample sizes necessary to ensure adequate power in simultaneous variation testing for both within-groups and between-group(s) comparisons (Kreft & De Leeuw, 1998; Raudenbush & Bryk, 2002; Raudenbush & Liu, 2000). For example, the use of HLM with balanced sample sizes is cited in mental health research (Bond, Miller, Krumweid, & Ward, 1988), education (Finn & Achilles, 1990; Mosteller, 1995), and medicine (Haddow, 1991). However, the literature pertaining to unbalanced sample size recommendations in HLM is meager (Raudenbush & Liu). This study builds on the work from previous balanced research perspectives by providing insight into the effect that unbalanced sampling has on the power estimates at level-1 in the random coefficient model with three dissimilar conditions of effect size and four intraclass correlations.

Kraemer and Thiemann (1987) broadly summarized and discussed the effects of sampling in a number of single level models. They found that small differences in sample sizes across groups for single-level analyses may not lower the estimated power of a test, but larger differences become problematic, indicating that unbalanced sampling tends to lower the model's expected power. Larger differences are those with proportional sampling differences that can incrementally differ by as much as 75% or as little as 25%. Such unbalanced sampling occurs as the rule rather than the exception in many educational settings where, for example, several classes are sampled to obtain a sample size of 200. Suppose that, in a particular school, each of 5 sampled classes have 10 students and each of 10 other sampled classes have 15 students. With sampling that is unbalanced to this extent, it would not be unreasonable to expect the same decreasing effect on power in multilevel models as is seen in single-level models.

Raudenbush and Liu (2000) provided a comprehensive summary of expected power estimate calculations based upon parameter estimates, balanced sample sizes, and overall resource expense. Unbalanced sampling, on the other hand, was only minimally addressed as a focus of suggested further research. Likewise, Reise and Duan (2003) suggested that the unbalanced nature of educational data produced design flaws in need of further research to investigate its effects on model efficiency.

## Method

Building on the work of Kraemer and Thiemann (1987) and Raudenbush and Liu (2000), three different overall sample sizes (as suggested by Raudenbush and Bryk, 2002) with three unbalanced sampling ratios (as suggested by Kraemer and Thiemann) were investigated in this study, with the focus on the differences in the amount of unbalanced sampling being used to determine if there was a recognizable effect on model power. Other model conditions that were varied included effect size and intraclass correlations (ICC) as possible contributors to decreases in power with multilevel models. The overall resource expense ratio was held to 1.

***Sampling Schemes***

Three levels of proportionally unbalanced data (75% to 25%, 80% to 20%, and 85% to 15%) were calculated for three different sample sizes of 200, 500, and 800 with each of 12 combinations of effect size and intraclass correlation. Design conditions were limited by the restriction that the number of classes ($C_1$) times the number of students in each of those classes ($S_1$) at the first sampling proportion (e.g., 25%) plus the number of classes ($C_2$) times the number of students in each of those classes ($S_2$) at the second sampling proportion (e.g., 75%) must equal the desired sample size (*N*), i.e., $(C_1S_1) + (C_2S_2) = N$, where in the 25%-75% sampling ratio, $(C_2S_2)$ must be 3 times larger than $(C_1S_1)$. This algebraic equation was used to generate 312 possible sampling ratios that would reflect possible classroom scenarios, where in each case the $C_1$, $S_1$, $C_2$, and $S_2$ values were integers. To be specific, 84 possible sampling combinations were used that corresponded to a total sample size of 200, another 84 possible sampling combinations were used with a total sample size of 500, and 144 possible sampling combinations were used with a total sample size of 800. Whereas Schumacker and Lomax (1996) (as cited in Heck & Thomas, 2000) provided a rule of thumb suggestion that a minimum of 100-150 subjects be included in a study, Heck and Thomas considered anything with $N < 400$ to be a small sample. Sample sizes, sampling schemes, and sampling ratio differences used in this study are displayed in Table 1 where a sample size of $N = 200$ is used as a representative of a small sample, $N = 500$ to represent a moderate sample size, and $N = 800$ to represent a large sample.

**Table 1**. Possible Sampling Combinations Equal to Unbalanced Samples of 200, 500, and 800.

| Ratio | Classes Trt 1 | Subjects Trt 1 | Classes Trt 2 | Subjects Trt 2 |
|---|---|---|---|---|
| | *N* = 200 | | | |
| .25-.75 | 5 | 10 | 10 | 15 |
| | 5 | 10 | 5 | 30 |
| .20-.80 | 2 | 20 | 10 | 16 |
| | 4 | 10 | 10 | 16 |
| | 2 | 20 | 5 | 32 |
| | 4 | 10 | 5 | 32 |
| .15-.85 | 2 | 15 | 10 | 17 |
| | *N* = 500 | | | |
| .25-.75 | 5 | 25 | 15 | 25 |
| .20-.80 | 4 | 25 | 20 | 20 |
| | 10 | 10 | 20 | 20 |
| | 4 | 25 | 16 | 25 |
| | 10 | 10 | 16 | 25 |
| .15-.85 | 3 | 25 | 17 | 25 |
| | 5 | 15 | 17 | 25 |
| | *N* = 800 | | | |
| .25-.75 | 8 | 25 | 24 | 25 |
| | 10 | 20 | 24 | 25 |
| | 8 | 25 | 30 | 20 |
| | 10 | 20 | 30 | 20 |
| .20-.80 | 10 | 16 | 20 | 32 |
| | 10 | 16 | 40 | 16 |
| | 5 | 32 | 20 | 32 |
| | 5 | 32 | 40 | 16 |
| .15-.85 | 4 | 30 | 17 | 40 |
| | 4 | 30 | 20 | 34 |
| | 5 | 24 | 17 | 40 |
| | 5 | 24 | 20 | 34 |

***Data Simulation***

After the unbalanced sampling schemes with corresponding sample sizes were determined, Step 1 of the simulation began. Ten thousand outcome variables were simulated for each of the 312 different sampling schemes using the SAS PROC IML (see the Appendix). These outcome variables were mechanically constrained to fit within given values for effect size, intraclass correlations, sample sizes, and proportions of unbalanced data. Step 2, performing an HLM analysis on each of the 10,000 iterations of the 312 sampling schemes using SAS PROC MIXED, produced the partitioned level-1 and level-2 power parameters. The 312 possible sampling combinations were grouped and aggregated according to sampling schema to generate 108 estimated level-1 power values.

Upon completion of the simulations, a comparison table was created where the effect of each level of each design characteristic (i.e., sample size, proportion of unbalanced data, effect size, and intraclass correlation) on model power could be investigated on the resultant dependent variable (the calculated estimate of level-1 power) for the simulated unbalanced random coefficient model. Visual comparisons were made.

## Results and Conclusions

Simulated data were created following the recommendations of Raudenbush and Liu (2000) for relative magnitudes of effect size, intraclass correlation, total sample size, and proportions of unbalanced data. The 312 possible sampling ratios and 108 level-1 aggregate power estimates from Step 2 are presented in Tables 2A – 4C. Power estimates when *N* = 200 are in Tables 2A – 2B, with *N* = 500 in Tables 3A – 3C, and for *N* = 800 in Tables 4A – 4C.

**Table 2A**. Aggregate Power for Sample Size of 200 for Sampling Ratios 0.25:0.75 and 0.20:0.80 by Four ICCs & Three Effect Sizes

| Ratio | ICC | ES | C1 | S1 | C2 | S2 | Power | Average |
|---|---|---|---|---|---|---|---|---|
| 0.25:0.75 | 0.2 | 1.0 | 5 | 10 | 10 | 15 | 0.9639 | 0.9663 |
| | | | 5 | 10 | 5 | 30 | 0.9686 | |
| | | 0.8 | 5 | 10 | 10 | 15 | 0.8977 | 0.8684 |
| | | | 5 | 10 | 5 | 30 | 0.8391 | |
| | | 0.5 | 5 | 10 | 10 | 15 | 0.6643 | 0.6266 |
| | | | 5 | 10 | 5 | 30 | 0.5889 | |
| | 0.1 | 1.0 | 5 | 10 | 10 | 15 | 0.9945 | 0.9942 |
| | | | 5 | 10 | 5 | 30 | 0.9939 | |
| | | 0.8 | 5 | 10 | 10 | 15 | 0.9612 | 0.9602 |
| | | | 5 | 10 | 5 | 30 | 0.9591 | |
| | | 0.5 | 5 | 10 | 10 | 15 | 0.7322 | 0.7370 |
| | | | 5 | 10 | 5 | 30 | 0.7417 | |
| | 0.05 | 1.0 | 5 | 10 | 10 | 15 | 0.9986 | 0.9937 |
| | | | 5 | 10 | 5 | 30 | 0.9887 | |
| | | 0.8 | 5 | 10 | 10 | 15 | 0.9838 | 0.9829 |
| | | | 5 | 10 | 5 | 30 | 0.9819 | |
| | | 0.5 | 5 | 10 | 10 | 15 | 0.7837 | 0.7100 |
| | | | 5 | 10 | 5 | 30 | 0.6363 | |
| | 0.01 | 1.0 | 5 | 10 | 10 | 15 | 0.9999 | 0.9985 |
| | | | 5 | 10 | 5 | 30 | 0.9970 | |
| | | 0.8 | 5 | 10 | 10 | 15 | 0.9612 | 0.9620 |
| | | | 5 | 10 | 5 | 30 | 0.9627 | |
| | | 0.5 | 5 | 10 | 10 | 15 | 0.8483 | 0.7544 |
| | | | 5 | 10 | 5 | 30 | 0.6604 | |
| 0.20:0.80 | 0.2 | 1.0 | 2 | 20 | 10 | 16 | 0.8577 | 0.9066 |
| | | | 4 | 10 | 10 | 16 | 0.9381 | |
| | | | 2 | 20 | 5 | 32 | 0.8823 | |
| | | | 4 | 10 | 5 | 32 | 0.9484 | |
| | | 0.8 | 2 | 20 | 10 | 16 | 0.7725 | 0.8246 |
| | | | 4 | 10 | 10 | 16 | 0.8614 | |
| | | | 2 | 20 | 5 | 32 | 0.7950 | |
| | | | 4 | 10 | 5 | 32 | 0.8693 | |
| | | 0.5 | 2 | 20 | 10 | 16 | 0.5803 | 0.6181 |
| | | | 4 | 10 | 10 | 16 | 0.6239 | |
| | | | 2 | 20 | 5 | 32 | 0.6200 | |
| | | | 4 | 10 | 5 | 32 | 0.6482 | |

**Table 2B**. Aggregate Power for Sample Size of 200, 0.20:0.80 and 0.15:0.85 Sampling Ratios by Four ICCs & Three Effect Sizes

| Ratio | ICC | ES | C1 | S1 | C2 | S2 | Power | Average |
|---|---|---|---|---|---|---|---|---|
| 0.20:0.80 | 0.1 | 1.0 | 2 | 20 | 10 | 16 | 0.9545 | 0.9704 |
| | | | 4 | 10 | 10 | 16 | 0.9826 | |
| | | | 2 | 20 | 5 | 32 | 0.9576 | |
| | | | 4 | 10 | 5 | 32 | 0.9867 | |
| | | 0.8 | 2 | 20 | 10 | 16 | 0.8526 | 0.9007 |
| | | | 4 | 10 | 10 | 16 | 0.9299 | |
| | | | 2 | 20 | 5 | 32 | 0.8767 | |
| | | | 4 | 10 | 5 | 32 | 0.9437 | |
| | | 0.5 | 2 | 20 | 10 | 16 | 0.6001 | 0.6475 |
| | | | 4 | 10 | 10 | 16 | 0.6669 | |
| | | | 2 | 20 | 5 | 32 | 0.6317 | |
| | | | 4 | 10 | 5 | 32 | 0.6912 | |
| | 0.05 | 1.0 | 2 | 20 | 10 | 16 | 0.9821 | 0.9913 |
| | | | 4 | 10 | 10 | 16 | 0.9969 | |
| | | | 2 | 20 | 5 | 32 | 0.9899 | |
| | | | 4 | 10 | 5 | 32 | 0.9962 | |
| | | 0.8 | 2 | 20 | 10 | 16 | 0.9304 | 0.9503 |
| | | | 4 | 10 | 10 | 16 | 0.9668 | |
| | | | 2 | 20 | 5 | 32 | 0.9355 | |
| | | | 4 | 10 | 5 | 32 | 0.9686 | |
| | | 0.5 | 2 | 20 | 10 | 16 | 0.6585 | 0.6974 |
| | | | 4 | 10 | 10 | 16 | 0.7138 | |
| | | | 2 | 20 | 5 | 32 | 0.6813 | |
| | | | 4 | 10 | 5 | 32 | 0.7358 | |
| | 0.01 | 1.0 | 2 | 20 | 10 | 16 | 0.9982 | 0.9991 |
| | | | 4 | 10 | 10 | 16 | 0.9997 | |
| | | | 2 | 20 | 5 | 32 | 0.9991 | |
| | | | 4 | 10 | 5 | 32 | 0.9992 | |
| | | 0.8 | 2 | 20 | 10 | 16 | 0.9824 | 0.9871 |
| | | | 4 | 10 | 10 | 16 | 0.9905 | |
| | | | 2 | 20 | 5 | 32 | 0.9854 | |
| | | | 4 | 10 | 5 | 32 | 0.9902 | |
| | | 0.5 | 2 | 20 | 10 | 16 | 0.7453 | 0.7680 |
| | | | 4 | 10 | 10 | 16 | 0.7784 | |
| | | | 2 | 20 | 5 | 32 | 0.7585 | |
| | | | 4 | 10 | 5 | 32 | 0.7896 | |
| 0.15:0.85 | 0.2 | 1.0 | 2 | 15 | 10 | 17 | 0.8522 | 0.8522 |
| | | 0.8 | 2 | 15 | 10 | 17 | 0.7469 | 0.7469 |
| | | 0.5 | 2 | 15 | 10 | 17 | 0.5471 | 0.5471 |
| | 0.1 | 1.0 | 2 | 15 | 10 | 17 | 0.9357 | 0.9357 |
| | | 0.8 | 2 | 15 | 10 | 17 | 0.8274 | 0.8274 |
| | | 0.5 | 2 | 15 | 10 | 17 | 0.5561 | 0.5561 |
| | 0.05 | 1.0 | 2 | 15 | 10 | 17 | 0.9697 | 0.9697 |
| | | 0.8 | 2 | 15 | 10 | 17 | 0.9063 | 0.9063 |
| | | 0.5 | 2 | 15 | 10 | 17 | 0.6086 | 0.6086 |
| | 0.01 | 1.0 | 2 | 15 | 10 | 17 | 0.9958 | 0.9958 |
| | | 0.8 | 2 | 15 | 10 | 17 | 0.9658 | 0.9658 |
| | | 0.5 | 2 | 15 | 10 | 17 | 0.6737 | 0.6737 |

In general, with *N* = 200, adequate average power was achieved with effect sizes equal to 1.0 and 0.8 for all three sampling ratios and all four ICC values (with 3 exceptions—sampling ratio of 0.20:0.80, ICC = 0.2, ES = 0.8; sampling ratio of 0.15:0.85, ICC = 0.2, ES = 0.8; and sampling ratio of 0.15:0.85, ICC = 0.1, ES = 0.8). None of the scenarios with effect size = 0.5 showed adequate average power.

With *N* = 500, adequate average power was achieved for all three ratios of unbalanced sampling, all four ICC values, and all three effect sizes with four exceptions each with effect size = 0.5: sampling ratio of 0.25:0.75, ICC = 0.2; sampling ratio of 0.20:0.80, ICC = 0.2; sampling ratio of 0.15:0.85, ICC = 0.2; and sampling ratio of 0.15:0.85, ICC = 0.1.

**Table 3A**. Aggregate Power for Sample Size of 500, 0.25:0.75 and 0.20:0.80 Sampling Ratios by Four ICCs & Three Effect Sizes

| Ratio | ICC | ES | C1 | S1 | C2 | S2 | Power | Average |
|---|---|---|---|---|---|---|---|---|
| 0.25:0.75 | 0.2 | 1.0 | 5 | 25 | 15 | 25 | 0.9943 | 0.9943 |
| | | 0.8 | | | | | 0.9598 | 0.9598 |
| | | 0.5 | | | | | 0.7886 | 0.7886 |
| | 0.1 | 1.0 | 5 | 25 | 15 | 25 | 0.9998 | 0.9998 |
| | | 0.8 | | | | | 0.9967 | 0.9967 |
| | | 0.5 | | | | | 0.8733 | 0.8733 |
| | 0.05 | 1.0 | 5 | 25 | 15 | 25 | 1.0 | 1.0 |
| | | 0.8 | | | | | 0.9996 | 0.9996 |
| | | 0.5 | | | | | 0.9522 | 0.9522 |
| | 0.01 | 1.0 | 5 | 25 | 15 | 25 | 1.0 | 1.0 |
| | | 0.8 | | | | | 1.0 | 1.0 |
| | | 0.5 | | | | | 0.9929 | 0.9929 |
| 0.20:0.80 | 0.2 | 1.0 | 4 | 25 | 20 | 20 | 0.9745 | 0.9885 |
| | | | 10 | 10 | 20 | 20 | 0.9997 | |
| | | | 4 | 25 | 16 | 25 | 0.9804 | |
| | | | 10 | 10 | 16 | 25 | 0.9994 | |
| | | 0.8 | 4 | 25 | 20 | 20 | 0.9193 | 0.9569 |
| | | | 10 | 10 | 20 | 20 | 0.9884 | |
| | | | 4 | 25 | 16 | 25 | 0.9338 | |
| | | | 10 | 10 | 16 | 25 | 0.9861 | |
| | | 0.5 | 4 | 25 | 20 | 20 | 0.7351 | 0.8016 |
| | | | 10 | 10 | 20 | 20 | 0.8619 | |
| | | | 4 | 25 | 16 | 25 | 0.7512 | |
| | | | 10 | 10 | 16 | 25 | 0.8581 | |
| | 0.1 | 1.0 | 4 | 25 | 20 | 20 | 0.9985 | 0.9991 |
| | | | 10 | 10 | 20 | 20 | 1.0 | |
| | | | 4 | 25 | 16 | 25 | 0.9979 | |
| | | | 10 | 10 | 16 | 25 | 1.0 | |
| | | 0.8 | 4 | 25 | 20 | 20 | 0.9802 | 0.9910 |
| | | | 10 | 10 | 20 | 20 | 0.9997 | |
| | | | 4 | 25 | 16 | 25 | 0.9854 | |
| | | | 10 | 10 | 16 | 25 | 0.9985 | |
| | | 0.5 | 4 | 25 | 20 | 20 | 0.8234 | 0.8840 |
| | | | 10 | 10 | 20 | 20 | 0.9346 | |
| | | | 4 | 25 | 16 | 25 | 0.8427 | |
| | | | 10 | 10 | 16 | 25 | 0.9351 | |

**Table 3B.** Aggregate Power for *N* = 500, 0.20:0.80 Sampling Ratio by 2 ICCs & 3 Effect Sizes

| Ratio | ICC | ES | C1 | S1 | C2 | S2 | Power | Average |
|---|---|---|---|---|---|---|---|---|
| 0.20:0.80 | 0.05 | 1.0 | 4 | 25 | 20 | 20 | 1.0 | 1.0 |
| | | | 10 | 10 | 20 | 20 | 1.0 | |
| | | | 4 | 25 | 16 | 25 | 0.9999 | |
| | | | 10 | 10 | 16 | 25 | 1.0 | |
| | | 0.8 | 4 | 25 | 20 | 20 | 0.9988 | 0.9991 |
| | | | 10 | 10 | 20 | 20 | 0.9999 | |
| | | | 4 | 25 | 16 | 25 | 0.9980 | |
| | | | 10 | 10 | 16 | 25 | 0.9998 | |
| | | 0.5 | 4 | 25 | 20 | 20 | 0.9051 | 0.9388 |
| | | | 10 | 10 | 20 | 20 | 0.9663 | |
| | | | 4 | 25 | 16 | 25 | 0.9144 | |
| | | | 10 | 10 | 16 | 25 | 0.9694 | |
| | 0.01 | 1.0 | 4 | 25 | 20 | 20 | 1.0 | 1.0 |
| | | | 10 | 10 | 20 | 20 | 1.0 | |
| | | | 4 | 25 | 16 | 25 | 1.0 | |
| | | | 10 | 10 | 16 | 25 | 1.0 | |
| | | 0.8 | 4 | 25 | 20 | 20 | 0.9999 | 1.0 |
| | | | 10 | 10 | 20 | 20 | 1.0 | |
| | | | 4 | 25 | 16 | 25 | 1.0 | |
| | | | 10 | 10 | 16 | 25 | 1.0 | |
| | | 0.5 | 4 | 25 | 20 | 20 | 0.9803 | 0.9859 |
| | | | 10 | 10 | 20 | 20 | 0.9892 | |
| | | | 4 | 25 | 16 | 25 | 0.9835 | |
| | | | 10 | 10 | 16 | 25 | 0.9904 | |

**Table 3C**. Aggregate Power for *N* = 500, 0.15:0.85 Sampling Ratio by 4 ICCs & 3 Effect Sizes

| Ratio | ICC | ES | C1 | S1 | C2 | S2 | Power | Average |
|---|---|---|---|---|---|---|---|---|
| 0.15:0.85 | 0.2 | 1.0 | 3 | 25 | 17 | 25 | 0.9510 | 0.9660 |
| | | | 5 | 15 | 17 | 25 | 0.9810 | |
| | | 0.8 | 3 | 25 | 17 | 25 | 0.8891 | 0.9328 |
| | | | 5 | 15 | 17 | 25 | 0.9764 | |
| | | 0.5 | 3 | 25 | 17 | 25 | 0.6901 | 0.7154 |
| | | | 5 | 15 | 17 | 25 | 0.7406 | |
| | 0.1 | 1.0 | 3 | 25 | 17 | 25 | 0.9920 | 0.9956 |
| | | | 5 | 15 | 17 | 25 | 0.9992 | |
| | | 0.8 | 3 | 25 | 17 | 25 | 0.9565 | 0.9725 |
| | | | 5 | 15 | 17 | 25 | 0.9884 | |
| | | 0.5 | 3 | 25 | 17 | 25 | 0.7529 | 0.7878 |
| | | | 5 | 15 | 17 | 25 | 0.8227 | |
| | 0.05 | 1.0 | 3 | 25 | 17 | 25 | 0.9999 | 1.0 |
| | | | 5 | 15 | 17 | 25 | 1.0 | |
| | | 0.8 | 3 | 25 | 17 | 25 | 0.9930 | 0.9957 |
| | | | 5 | 15 | 17 | 25 | 0.9984 | |
| | | 0.5 | 3 | 25 | 17 | 25 | 0.8474 | 0.8725 |
| | | | 5 | 15 | 17 | 25 | 0.8975 | |
| | 0.01 | 1.0 | 3 | 25 | 17 | 25 | 1.0 | 1.0 |
| | | | 5 | 15 | 17 | 25 | 1.0 | |
| | | 0.8 | 3 | 25 | 17 | 25 | 0.9999 | 1.0 |
| | | | 5 | 15 | 17 | 25 | 1.0 | |
| | | 0.5 | 3 | 25 | 17 | 25 | 0.9470 | 0.9557 |

**Table 4A**. Aggregate Power for *N* = 800, 0.25:0.75 Sampling Ratio by 4 ICCs and 3Effect Sizes.

| Ratio | ICC | ES | C1 | S1 | C2 | S2 | Power | Average |
|---|---|---|---|---|---|---|---|---|
| 0.25:0.75 | 0.2 | 1.0 | 8 | 25 | 24 | 25 | 0.9997 | 0.9994 |
| | | | 10 | 20 | 24 | 25 | 0.9993 | |
| | | | 8 | 25 | 30 | 20 | 0.9986 | |
| | | | 10 | 20 | 30 | 20 | 1.0 | |
| | | 0.8 | 8 | 25 | 24 | 25 | 0.9922 | 0.9932 |
| | | | 10 | 20 | 24 | 25 | 0.9936 | |
| | | | 8 | 25 | 30 | 20 | 0.9898 | |
| | | | 10 | 20 | 30 | 20 | 0.9973 | |
| | | 0.5 | 8 | 25 | 24 | 25 | 0.8927 | 0.9017 |
| | | | 10 | 20 | 24 | 25 | 0.9134 | |
| | | | 8 | 25 | 30 | 20 | 0.8880 | |
| | | | 10 | 20 | 30 | 20 | 0.9125 | |
| | 0.1 | 1.0 | 8 | 25 | 24 | 25 | 1.0 | 1.0 |
| | | | 10 | 20 | 24 | 25 | 1.0 | |
| | | | 8 | 25 | 30 | 20 | 1.0 | |
| | | | 10 | 20 | 30 | 20 | 1.0 | |
| | | 0.8 | 8 | 25 | 24 | 25 | 1.0 | 0.9999 |
| | | | 10 | 20 | 24 | 25 | 1.0 | |
| | | | 8 | 25 | 30 | 20 | 0.9997 | |
| | | | 10 | 20 | 30 | 20 | 1.0 | |
| | | 0.5 | 8 | 25 | 24 | 25 | 0.9679 | 0.9703 |
| | | | 10 | 20 | 24 | 25 | 0.9733 | |
| | | | 8 | 25 | 30 | 20 | 0.9593 | |
| | | | 10 | 20 | 30 | 20 | 0.9807 | |
| | 0.05 | 1.0 | 8 | 25 | 24 | 25 | 1.0 | 1.0 |
| | | | 10 | 20 | 24 | 25 | 1.0 | |
| | | | 8 | 25 | 30 | 20 | 1.0 | |
| | | | 10 | 20 | 30 | 20 | 1.0 | |
| | | 0.8 | 8 | 25 | 24 | 25 | 1.0 | 1.0 |
| | | | 10 | 20 | 24 | 25 | 1.0 | |
| | | | 8 | 25 | 30 | 20 | 1.0 | |
| | | | 10 | 20 | 30 | 20 | 1.0 | |
| | | 0.5 | 8 | 25 | 24 | 25 | 0.9929 | 0.9938 |
| | | | 10 | 20 | 24 | 25 | 0.9959 | |
| | | | 8 | 25 | 30 | 20 | 0.9904 | |
| | | | 10 | 20 | 30 | 20 | 0.9961 | |
| | 0.01 | 1.0 | 8 | 25 | 24 | 25 | 1.0 | 1.0 |
| | | | 10 | 20 | 24 | 25 | 1.0 | |
| | | | 8 | 25 | 30 | 20 | 1.0 | |
| | | | 10 | 20 | 30 | 20 | 1.0 | |
| | | 0.8 | 8 | 25 | 24 | 25 | 1.0 | 1.0 |
| | | | 10 | 20 | 24 | 25 | 1.0 | |
| | | | 8 | 25 | 30 | 20 | 1.0 | |
| | | | 10 | 20 | 30 | 20 | 1.0 | |
| | | 0.5 | 8 | 25 | 24 | 25 | 1.0 | 0.9999 |
| | | | 10 | 20 | 24 | 25 | 0.9999 | |
| | | | 8 | 25 | 30 | 20 | 0.9998 | |
| | | | 10 | 20 | 30 | 20 | 1.0 | |

**Table 4B**. Aggregate Power for *N* = 800, 0.20:0.80 Sampling Ratio by 4 ICCs and 3 Effect Sizes

| Ratio | ICC | ES | C1 | S1 | C2 | S2 | Power | Average |
|---|---|---|---|---|---|---|---|---|
| 0.20:0.80 | 0.2 | 1.0 | 10 | 16 | 20 | 32 | 0.9990 | 0.9964 |
| | | | 10 | 16 | 40 | 16 | 1.0 | |
| | | | 5 | 32 | 20 | 32 | 0.9948 | |
| | | | 5 | 32 | 40 | 16 | 0.9917 | |
| | | 0.8 | 10 | 16 | 20 | 32 | 0.9945 | 0.9762 |
| | | | 10 | 16 | 40 | 16 | 0.9930 | |
| | | | 5 | 32 | 20 | 32 | 0.9633 | |
| | | | 5 | 32 | 40 | 16 | 0.9541 | |
| | | 0.5 | 10 | 16 | 20 | 32 | 0.9046 | 0.8555 |
| | | | 10 | 16 | 40 | 16 | 0.8914 | |
| | | | 5 | 32 | 20 | 32 | 0.8260 | |
| | | | 5 | 32 | 40 | 16 | 0.8000 | |
| | 0.1 | 1.0 | 10 | 16 | 20 | 32 | 1.0 | 0.9999 |
| | | | 10 | 16 | 40 | 16 | 1.0 | |
| | | | 5 | 32 | 20 | 32 | 0.9997 | |
| | | | 5 | 32 | 40 | 16 | 0.9999 | |
| | | 0.8 | 10 | 16 | 20 | 32 | 0.9996 | 0.9983 |
| | | | 10 | 16 | 40 | 16 | 0.9997 | |
| | | | 5 | 32 | 20 | 32 | 0.9975 | |
| | | | 5 | 32 | 40 | 16 | 0.9962 | |
| | | 0.5 | 10 | 16 | 20 | 32 | 0.9677 | 0.9294 |
| | | | 10 | 16 | 40 | 16 | 0.9650 | |
| | | | 5 | 32 | 20 | 32 | 0.8983 | |
| | | | 5 | 32 | 40 | 16 | 0.8867 | |
| | 0.05 | 1.0 | 10 | 16 | 20 | 32 | 1.0 | 1.0 |
| | | | 10 | 16 | 40 | 16 | 1.0 | |
| | | | 5 | 32 | 20 | 32 | 1.0 | |
| | | | 5 | 32 | 40 | 16 | 1.0 | |
| | | 0.8 | 10 | 16 | 20 | 32 | 1.0 | 1.0 |
| | | | 10 | 16 | 40 | 16 | 1.0 | |
| | | | 5 | 32 | 20 | 32 | 0.9999 | |
| | | | 5 | 32 | 40 | 16 | 0.9999 | |
| | | 0.5 | 10 | 16 | 20 | 32 | 0.9926 | 0.9787 |
| | | | 10 | 16 | 40 | 16 | 0.9904 | |
| | | | 5 | 32 | 20 | 32 | 0.9666 | |
| | | | 5 | 32 | 40 | 16 | 0.9651 | |
| | 0.01 | 1 | 10 | 16 | 20 | 32 | 1.0 | 1.0 |
| | | | 10 | 16 | 40 | 16 | 1.0 | |
| | | | 5 | 32 | 20 | 32 | 1.0 | |
| | | | 5 | 32 | 40 | 16 | 1.0 | |
| | | 0.8 | 10 | 16 | 20 | 32 | 1.0 | 1.0 |
| | | | 10 | 16 | 40 | 16 | 1.0 | |
| | | | 5 | 32 | 20 | 32 | 1.0 | |
| | | | 5 | 32 | 40 | 16 | 1.0 | |
| | | 0.5 | 10 | 16 | 20 | 32 | 0.9996 | 0.9989 |
| | | | 10 | 16 | 40 | 16 | 0.9994 | |
| | | | 5 | 32 | 20 | 32 | 0.9983 | |
| | | | 5 | 32 | 40 | 16 | 0.9984 | |

**Table 4C**. Aggregate Power for *N* = 800, 0.15:0.85 Sampling Ratio by 4 ICCs and 3 Effect Sizes

| Ratio | ICC | ES | C1 | S1 | C2 | S2 | Power | Average |
|---|---|---|---|---|---|---|---|---|
| 0.15:0.85 | 0.2 | 1.0 | 4 | 30 | 17 | 40 | 0.9839 | 0.9886 |
| | | | 4 | 30 | 20 | 34 | 0.9899 | |
| | | | 5 | 24 | 17 | 40 | 0.9903 | |
| | | | 5 | 24 | 20 | 34 | 0.9903 | |
| | | 0.8 | 4 | 30 | 17 | 40 | 0.9474 | 0.9539 |
| | | | 4 | 30 | 20 | 34 | 0.9424 | |
| | | | 5 | 24 | 17 | 40 | 0.9618 | |
| | | | 5 | 24 | 20 | 34 | 0.9640 | |
| | | 0.5 | 4 | 30 | 17 | 40 | 0.7846 | 0.7969 |
| | | | 4 | 30 | 20 | 34 | 0.7814 | |
| | | | 5 | 24 | 17 | 40 | 0.8150 | |
| | | | 5 | 24 | 20 | 34 | 0.8064 | |
| | 0.1 | 1.0 | 4 | 30 | 17 | 40 | 0.9988 | 0.9992 |
| | | | 4 | 30 | 20 | 34 | 0.9987 | |
| | | | 5 | 24 | 17 | 40 | 0.9999 | |
| | | | 5 | 24 | 20 | 34 | 0.9993 | |
| | | 0.8 | 4 | 30 | 17 | 40 | 0.9888 | 0.9919 |
| | | | 4 | 30 | 20 | 34 | 0.9923 | |
| | | | 5 | 24 | 17 | 40 | 0.9920 | |
| | | | 5 | 24 | 20 | 34 | 0.9946 | |
| | | 0.5 | 4 | 30 | 17 | 40 | 0.8621 | 0.8799 |
| | | | 4 | 30 | 20 | 34 | 0.8689 | |
| | | | 5 | 24 | 17 | 40 | 0.8919 | |
| | | | 5 | 24 | 20 | 34 | 0.8966 | |
| | 0.05 | 1.0 | 4 | 30 | 17 | 40 | 1.0 | 1.0 |
| | | | 4 | 30 | 20 | 34 | 1.0 | |
| | | | 5 | 24 | 17 | 40 | 1.0 | |
| | | | 5 | 24 | 20 | 34 | 1.0 | |
| | | 0.8 | 4 | 30 | 17 | 40 | 0.9994 | 0.9996 |
| | | | 4 | 30 | 20 | 34 | 0.9997 | |
| | | | 5 | 24 | 17 | 40 | 0.9996 | |
| | | | 5 | 24 | 20 | 34 | 0.9998 | |
| | | 0.5 | 4 | 30 | 17 | 40 | 0.9370 | 0.9472 |
| | | | 4 | 30 | 20 | 34 | 0.9401 | |
| | | | 5 | 24 | 17 | 40 | 0.9582 | |
| | | | 5 | 24 | 20 | 34 | 0.9536 | |
| | 0.01 | 1.0 | 4 | 30 | 17 | 40 | 1.0 | 1.0 |
| | | | 4 | 30 | 20 | 34 | 1.0 | |
| | | | 5 | 24 | 17 | 40 | 1.0 | |
| | | | 5 | 24 | 20 | 34 | 1.0 | |
| | | 0.8 | 4 | 30 | 17 | 40 | 1.0 | 1.0 |
| | | | 4 | 30 | 20 | 34 | 1.0 | |
| | | | 5 | 24 | 17 | 40 | 1.0 | |
| | | | 5 | 24 | 20 | 34 | 1.0 | |
| | | 0.5 | 4 | 30 | 17 | 40 | 0.9913 | 0.9941 |
| | | | 4 | 30 | 20 | 34 | 0.9936 | |
| | | | 5 | 24 | 17 | 40 | 0.9959 | |
| | | | 5 | 24 | 20 | 34 | 0.9954 | |

With *N* = 800, adequate average power was achieved for all three ratios of unbalanced sampling, all four ICC values, and all three effect sizes with only one exception: sampling ratio of 0.15:0.85, ICC = 0.2 and ES = 0.5. Overall, with effect size set at 1.0 or 0.8, the only scenarios for which adequate average power was not achieved was with the small sample size of *N* = 200. With *N* = 500 or 800, the only scenarios for which adequate average power was not achieved all had effect size = 0.5 and ICC values equal to 0.2 or 0.1.

Aggregate data in Table 5 represent summaries for sample sizes of *N* = 200, 500, and 800 by each of the three sampling ratios, the four ICC values, and the three effect sizes for a total of 108 average power estimates. These results indicate that increasing the width of the sampling ratios has the effect of lowering the estimated power in most cases. For each sample size in each column, the aggregate estimated power decreased as the width of the sampling ratio increased. This effect is more pronounced for the smaller sample size of *N* = 200.

Five of these 108 aggregated power estimates (4.6%) exhibited an exception to the decreasing pattern, where contrary to every other estimate, power showed a slight increase. In every case, this exception occurs in the sampling ratio of 0.20:0.80 three times with *N* = 200 and twice with *N* = 500. These exceptions are shaded in Table 5.

Plausible explanations for these differences come from the work of Kreft and De Leeuw (1998) and Raudenbush and Liu (2000) where each researcher determined that using a larger number of groups have a greater positive effect on estimated power than having more subjects within groups. The five aggregate power estimates that are exceptions come from samples in which the group size is small. It also appears to generally be the case that even when intraclass correlations and effect sizes are high, power estimates are not compromised at the lower sample size. A wider sampling ratio, or greater unbalanced sampling, has the most pronounced effect on power and presents the greatest threat to research results.

**Table 5**. Comparisons of Aggregate Power Estimates for All Variables

Comparison of Power for ICC = 0.2 and Effect Size = 1.0, 0.8, & 0.5

| | ICC = 0.2 & ES = 1.0 | | | ICC = 0.2 & ES = 0.8 | | | ICC = 0.2 & ES = 0.5 | | |
|---|---|---|---|---|---|---|---|---|---|
| Ratio | 200 | 500 | 800 | 200 | 500 | 800 | 200 | 500 | 800 |
| 0.25 : 0.75 | 0.9663 | 0.9943 | 0.9994 | 0.8684 | 0.9598 | 0.9932 | 0.6266 | 0.7886 | 0.9017 |
| 0.20 : 0.80 | 0.9066 | 0.9885 | 0.9964 | 0.8246 | 0.9569 | 0.9762 | 0.6181 | 0.8016 | 0.8555 |
| 0.15 : 0.85 | 0.8522 | 0.9660 | 0.9886 | 0.7469 | 0.9328 | 0.9539 | 0.5471 | 0.7154 | 0.7969 |

Comparison of Power for ICC = 0.1 and Effect Size = 1, 0.8, & 0.5

| | ICC = 0.1 & ES 1.0 | | | ICC = 0.1 & ES = 0.8 | | | ICC = 0.1 & ES = 0.5 | | |
|---|---|---|---|---|---|---|---|---|---|
| Ratio | 200 | 500 | 800 | 200 | 500 | 800 | 200 | 500 | 800 |
| 0.25 : 0.75 | 0.9942 | 0.9998 | 1.0000 | 0.9602 | 0.9967 | 0.9999 | 0.7370 | 0.8733 | 0.9703 |
| 0.20 : 0.80 | 0.9704 | 0.9991 | 0.9999 | 0.9007 | 0.9910 | 0.9983 | 0.6475 | 0.8840 | 0.9294 |
| 0.15 : 0.85 | 0.9357 | 0.9956 | 0.9992 | 0.8274 | 0.9725 | 0.9919 | 0.5561 | 0.7878 | 0.8799 |

Comparison of Power for ICC = 0.05 and Effect Size = 1, 0.8, & 0.5

| | ICC = 0.05 & ES = 1.0 | | | ICC = 0.05 & ES = 0.8 | | | ICC = 0.05 & ES = 0.5 | | |
|---|---|---|---|---|---|---|---|---|---|
| Ratio | 200 | 500 | 800 | 200 | 500 | 800 | 200 | 500 | 800 |
| 0.25 : 0.75 | 0.9937 | 1.0000 | 1.0000 | 0.9829 | 0.9996 | 1.0000 | 0.7100 | 0.9522 | 0.9938 |
| 0.20 : 0.80 | 0.9913 | 1.0000 | 1.0000 | 0.9503 | 0.9991 | 1.0000 | 0.6974 | 0.9388 | 0.9787 |
| 0.15 : 0.85 | 0.9697 | 1.0000 | 1.0000 | 0.9063 | 0.9957 | 0.9996 | 0.6086 | 0.8725 | 0.9472 |

Comparison of Power for ICC = 0.01 and Effect Size = 1, 0.8, & 0.5

| | ICC = 0.01 & ES = 1.0 | | | ICC = 0.01 & ES = 0.8 | | | ICC = 0.01 & ES = 0.5 | | |
|---|---|---|---|---|---|---|---|---|---|
| Ratio | 200 | 500 | 800 | 200 | 500 | 800 | 200 | 500 | 800 |
| 0.25 : 0.75 | 0.9984 | 1.0000 | 1.0000 | 0.9620 | 1.0000 | 1.0000 | 0.7544 | 0.9929 | 0.9999 |
| 0.20 : 0.80 | 0.9990 | 1.0000 | 1.0000 | 0.9871 | 1.0000 | 1.0000 | 0.7680 | 0.9859 | 0.9989 |
| 0.15 : 0.85 | 0.9958 | 1.0000 | 1.0000 | 0.9658 | 1.0000 | 1.0000 | 0.6737 | 0.9557 | 0.9941 |

Aggregate data presented in Table 6 exhibits level-1 power estimates for sample sizes of 200, 500, and 800 with three sampling ratios, four intraclass correlations, and three effect sizes for a total of 108 mean power estimates. Cohen (1988) suggests aggregate power estimates at or above .80 that possess adequate magnitude to ensure research integrity. Being slightly more conservative, this study considered power estimates that were less than .85 to possess inadequate magnitude to ensure research integrity.

Counting the number of estimates that fell below the selected value, the results indicated that increasing the width of the three sampling ratios has the effect of lowering the estimated power in most cases. For each sample size in the last three columns of Table 6, the aggregate estimated power reduced as the breadth of the three sample ratios increased. For example, within the column of sample sizes = 200, 42% of the power estimates were below .85.

**Table 6**. Aggregated Estimated Power

| Levels of Proportionally Unbalanced Data | Intraclass Correlations of Level 2 Units | Effect Size | Aggregate Level-1 Power *N* = 200 | Aggregate Level-1 Power *N* = 500 | Aggregate Level-1 Power *N* = 800 |
|---|---|---|---|---|---|
| 0.25 : 0.75 | 0.2 | 1 | 0.9663 | 0.9943 | 0.9994 |
| | | 0.8 | 0.8684 | 0.9598 | 0.9932 |
| | | 0.5 | 0.6266 | 0.7886 | 0.9017 |
| | 0.1 | 1 | 0.9942 | 0.9998 | 1.0 |
| | | 0.8 | 0.9602 | 0.9967 | 0.9999 |
| | | 0.5 | 0.7370 | 0.8733 | 0.9703 |
| | 0.05 | 1 | 0.9937 | 1.0 | 1.0 |
| | | 0.8 | 0.9829 | 0.9996 | 1.0 |
| | | 0.5 | 0.7100 | 0.9522 | 0.9938 |
| | 0.01 | 1 | 0.9985 | 1.0 | 1.0 |
| | | 0.8 | 0.9620 | 1.0 | 1.0 |
| | | 0.5 | 0.7544 | 0.9929 | 0.9999 |
| 0.20 : 0.80 | 0.2 | 1 | 0.9066 | 0.9885 | 0.9964 |
| | | 0.8 | 0.8246 | 0.9569 | 0.9762 |
| | | 0.5 | 0.6181 | 0.8016 | 0.8555 |
| | 0.1 | 1 | 0.9704 | 0.9991 | 0.9999 |
| | | 0.8 | 0.9007 | 0.9910 | 0.9983 |
| | | 0.5 | 0.6475 | 0.8840 | 0.9294 |
| | 0.05 | 1 | 0.9913 | 1.0 | 1.0 |
| | | 0.8 | 0.9503 | 0.9991 | 1.0 |
| | | 0.5 | 0.6974 | 0.9388 | 0.9787 |
| | 0.01 | 1 | 0.9991 | 1.0 | 1.0 |
| | | 0.8 | 0.9871 | 1.0 | 1.0 |
| | | 0.5 | 0.7680 | 0.9859 | 0.9989 |
| 0.15 : 0.85 | 0.2 | 1 | 0.8522 | 0.9660 | 0.9886 |
| | | 0.8 | 0.7469 | 0.9328 | 0.9539 |
| | | 0.5 | 0.5471 | 0.7154 | 0.7969 |
| | 0.1 | 1 | 0.9357 | 0.9956 | 0.9992 |
| | | 0.8 | 0.8274 | 0.9725 | 0.9919 |
| | | 0.5 | 0.5561 | 0.7878 | 0.8799 |
| | 0.05 | 1 | 0.9697 | 1.0 | 1.0 |
| | | 0.8 | 0.9063 | 0.9957 | 0.9996 |
| | | 0.5 | 0.6086 | 0.8725 | 0.9472 |
| | 0.01 | 1 | 0.9958 | 1.0 | 1.0 |
| | | 0.8 | 0.9658 | 1.0 | 1.0 |
| | | 0.5 | 0.6737 | 0.9557 | 0.9941 |

**Table 7**. Number & Percentage of Power Estimates < 0.85 for Unbalanced Sample Schemes

| Levels | *N* = 200 | *N* = 500 | *N* = 800 |
|---|---|---|---|
| 0.25 : 0.75 | 4 (33%) | 1(8%) | 0 (0%) |
| 0.20 : 0.80 | 5 (42%) | 1(8%) | 0 (0%) |
| 0.15 : 0.85 | 6 (50%) | 2 (17%) | 1 (8%) |

Additionally, comparing samples of 200, where the unbalanced levels are measured at 0.25:0.75, 33% of the power estimates fell below .85. At 0.20:0.80, 42% of the power estimates fell below .85 and at 0.15:0.85, 50% of the power estimates fell below .85. This effect is more pronounced for the smaller sample size of 200. The majority of the scenarios with sample sizes equal to 500 and 800 produced smaller comparative power differences (see Table 7). In the 108 possible power estimates, five (4.6%) exceptions to the decreasing pattern are seen where contrary to every other estimate, the power increases slightly.

## Implications

The cost of utilizing larger sampling techniques to ensure model adequacy may not meet the challenges of today's dwindling budgets. "Doing more with less" would be the preferred method despite the mixed messages inferred from previous research. For example, Bassari (1988) estimates detection of cross-level effects with sufficient power needed at least 30 groups with 30 participants per group or a total sample of 900. Kreft and De Leeuw (1998) (as cited in Heck & Thomas, 2000) found groups as low as 20 were sufficient to determine cross-level effects (i.e., with a total sample size of 600). The results of the present study can help to update educational researchers concerning the recommended sample sizes needed to achieve adequate power when utilizing unbalanced sampling in multilevel models.

## References

Bassiri, D. (1988). *Large and small sample properties of maximum likelihood estimates for the hierarchical linear model.* Unpublished doctoral dissertation, Michigan State University.

Bond, G., Miller, L., Drumweid, R., & Ward, R. (1988). Assertive cases management in three CMHs: A controlled study. *Hospital and Community Psychiatry, 9,* 411-418.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal, 27,* 557-577.

Haddow, J. (1991). Cotinine-assisted intervention in pregnancy to reduce smoking and low birthweight delivery. *British Journal of Obstetrics and Gynecology, 98,* 859-865.

Heck, R. H., & Thomas, S. L. (2000). *An introduction to multilevel modeling techniques*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research.* Newbury Park, CA: Sage Publications, Inc.

Kreft, I., & De Leeuw, J., (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage Publications, Inc.

Mosteller, F. (1995). Optimal design in psychological research. *Psychological Methods, 2,* 3-19.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models*. Thousand Oaks, CA: Sage Publications, Inc.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5,* 199-213.

Reise, S. P., & Duan, N. (2003). Design issues in multilevel studies. In S. Reise & N. Duan (Eds.), *Multilevel modeling methodological advances, issues, and applications* (pp. 285 – 298). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Schumacker, R E., & Lomax, R. G. (1996). *A beginner's guide to structural equation modeling.* Mahwah, NJ: Lawrence Erlbaum Associates.

Yeates, K. O., & Taylor, H. G. (1997). Predicting premorbid neuropsychological functioning following pediatric traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology, 19,* 825-.

Send correspondence to: Bonnie J. Steele
Colorado Mountain College
Email: bsteele@coloradomtn.edu

# APPENDIX

## SAS Program for Simulating Outcome Variables and Estimating Power

The following is the SAS code for running the simulation that produces 10,000 outcome variables for each designated case then takes these outcome variables through SAS PROC MIXED procedure to estimate power for each set of sampling ratios. Note: Intraclass correlations and effect sizes must be defined as fractions, not decimals or the program will cease to run.

```
/** Generate unbalanced two-level data **/
%let icc=1/100; *intraclass correlation coefficient;
%let g1=2; *number of classes in treatment group 1;
%let g2=10; *number of classes in treatment group 2;
%let n1=20; *number of subjects/class in treatment group 1;
%let n2=16; *number of subjects/class in treatment group 2;
%let ti=2; *number of treatments.  DO NOT CHANGE THIS VALUE.;
%let es=1; *effect size;
%let se=1; *standard deviation of individuals (level 2);
%let iter=10000; *this is the number of times you want the simulation to
iterate;

*Note: standard deviation of classes is determined computationally by
 the standard deviation of individuals as well as the effect size.;

title;
data tests;
probf=1;
delete;
run;

/** Generate Data **/
%macro datagen;
   ods select none;
   proc iml;
   icc=&icc; g1=&g1; n1=&n1; g2=&g2; n2=&n2; ti=&ti; se=&se; es=&es;
   mu=j(ti,1,1);
   mu[ti]=mu[ti]+es*se;
   se=1;
   sd=sqrt((icc/(1-icc))*se*se);
   y={0 0 0 0};
   CREATE datagen From y [colname={trt,class,student,y}];
   j=1;
      do k=1 to g1;
         z=normal(0);
         do i=1 to n1;
            w=rannor(0);
            y[1]=j;y[2]=k;y[3]=i;
            y[4]=mu[j]+sd*z+se*w;
            APPEND FROM y;
         end;
      end;
   j=2;
      do k=1 to g2;
         z=normal(0);
         do i=1 to n2;
            w=normal(0);
            y[1]=j;y[2]=k;y[3]=i;
            y[4]=mu[j]+sd*z+se*w;
            APPEND FROM y;
         end;
      end;

   close datagen;
   quit;
   proc mixed data=datagen;
      class class;
      model y=trt;
      random class;
   ods output tests3=tests3;
   run;quit;

   data tests;
   set tests tests3;
   run;

   ods select all;
%mend datagen;
%macro iterate;
   options nonotes nodate nonumber;ods results off;
   %do i=1 %to &iter;
   %datagen;
   %end;
   options notes;ods results on;

   %if &es+0=0 %then %do;
      title 'This is the simulated value of alpha.';
   %end;
   %if &es+0^=0 %then %do;
      title 'This is the simulated value of power.';
   %end;

data prop;
   set tests;
   rejects=probf<.05;
run;

proc means data=prop mean;
   var rejects;
run;
title;
%mend iterate;
%iterate;
```

# Comparing Cross Validated Classification Accuracies for Alternate Predictor Variable Weighting Algorithms

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

The present research contrasts the effectiveness of four predictor variable weighting algorithms with respect to cross-validated accuracies in classification problems. Ordinary Least Squares Regression (OLS), Ridge Regression (RR), Principle Components (PC), and Logistic Regression (LR), are the techniques that were contrasted on 24 real data sets in terms of optimizing cross-validated classification accuracies. LR was best in only 1 data set, PC was best overall in 16%, RR was best in 8%, and OLS was best in 8% of the data sets.

This investigation contrasts four weighting algorithms for classifying subjects into a priori groups based upon classification accuracy. Ordinary Least Squares Regression (OLS), also the same as classification using a linear predictive discriminant analysis or in the case of two groups, Fisher's (LDF); Ridge Regression (RR); Principle Components (PC); and Logistic Regression (LR), are the techniques that were compared with respect to their cross-validated classification accuracies in real data.

In a regression context, Darlington (1978) posited that cross validation accuracy is a function of $R^2$, N, VC, where $R^2$ represents the squared multiple correlation, N is the sample size, and VC is defined as the validity concentration. In Darlington's formulation, validity concentration was used to describe a data condition in which the principal components of the predictors with large eigenvalues also have large correlations with the criterion. Thus, validity concentration requires some degree of collinearity. Darlington suggested that the most useful statistical techniques for practical prediction problems, as in personnel selection, may be ridge regression and Stein-type regression. These combine the sensitivity of multiple regression with the resistance to sampling error of other techniques—notably rational (clinical) weights and weights determined by simple correlations. Darlington stated that the new techniques are not recommended for theoretical modeling work because they yield biased estimates of the true least squares weights, typically have higher expected squared errors for estimating some weights, and do not allow the use of ordinary confidence bands or significance tests. Nevertheless, he recommended the use of ridge regression as best for most classification problems.

Morris (1982) re-examined the performance of ridge regression from a different methodological perspective using the same data structures on which Darlington (1978) demonstrated the technique's superiority. Contrary to Darlington's suggestions, Morris (1982) found that ridge regression was never the most accurate prediction technique, although least squares weights, as well as all of the other non-least-squares techniques, were most accurate in some data configurations.

Further, Morris (1983) examined Darlington's (1978) suggestion to utilize a "shrunken inter-correlation matrix" as the input to an ordinary stepwise regression program to accomplish a stepwise ridge regression solution. The algorithm that Darlington suggested calculates the portions of predictable criterion variance attributable to ridge weighted variable subsets incorrectly, causing inappropriate predictor variable subsets to be selected. An alternate stepwise ridge regression procedure is suggested by Morris (1983).

Through simulation, Morris (1982) showed that as $R^2$ decreases, N decreases, and the VC increases, Ridge Regression becomes better than Ordinary Least Squares but, as well, Reduced Rank, Equal Weighting, and other techniques become better than Ridge. In several studies, Morris (1982) and Morris and Huberty (1987) found that the performance of Ridge Regression was inferior to that of Ordinary Least Squares, Principal Components, Reduced Rank, and Equal Weighting in all but a few data structures.

In fact, there is some evidence that cross-validated $R^2$ becomes better with increased VC, even better than the $R^2$ of OLS at low VC. Because Validity Concentration requires collinearity, the interest might be in examining whether collinearity can, under some circumstances, be helpful to prediction. The present research seeks to contrast Logistic Regression, as a popular classification technique frequently proffered in the literature, with the prior three methods examined in Morris and Huberty (1987).

## Method

A similar comparison (Morris & Huberty, 1987) examined only the OLS, RR, and PC methods. Logistic Regression will expand that coverage. Twenty four real data sets with varying degrees of group separation were analyzed using these four methods to ascertain differences in classification accuracies. All predictor weighting algorithms were cross validated using the Leave-One-Out technique. This algorithm is executed by alternately predicting each subject's group membership from the equation generated from the predictor and criterion scores of all other subjects. The resulting hit-rate over all subjects serves as a criterion for cross-validation accuracy.

## Results

The present research seeks to expand upon prior work investigating the effects of three weighting algorithms on classification accuracies: OLS, Ridge, and PC. In those simulation studies, all methods performed better with increasing sample size, larger population multiple correlations, and large degrees of group separation. OLS performed better with smaller levels of validity concentration. As VC increased, the performance of Ridge Regression was superior, and, at very high levels of VC, Principal Components Regression was superior. It is salient to note that in larger samples, this trend was delayed (Morris & Huberty, 1987). Overall, non-OLS methods performed best, or with increased accuracy, in small samples. It should also be noted, however, that even at high levels of VC, and with significant differences in classification accuracies, the differences were often small (Morris & Huberty).

The finding with real data mirrored the simulation results, but with the focus on contrasting results for specific data sets; not a general contrast of methods. Table 1 reflects the results of the contrast in cross-validated hit rates for 24 real data sets with varying degrees of group separation, numbers of subjects, variables, and data co-variance matrices. As can be seen, LR, the additional method being contrasted, is not present in the first 3 data sets. Overall, LR is best in only 1 data set (i.e., # 15 Block 3 & 4). It is tied with other methods in 6 (29%) of the data sets. It is second best in 7 (29%) of the data sets, third in 2 (13%) of the data sets, and worst in 5 (21%) of the data sets. In two of these data sets (i.e., #6 Bisbey 1 & 2 and # 10 Rulon 1 & 3), LR performed the worst of all four of the methods. All methods performed equally well at a 79% hit rate in the #9 Demographics #2 data set. PC was best overall in 4 (16%) of the data sets and tied for best in 3 (13%) of the data sets. RR was best overall in 2 data sets (8%) and tied for best in 8 (33%) of the data sets. Finally, OLS was best overall in 2 (8%) and tied for best in 8 (33%) of the data sets.

## Discussion

To summarize, the present research contrasted four predictor weighting algorithms: Ordinary Least Squares Regression, Ridge Regression, Principle Components, and Logistic Regression. The purpose of the study was to enhance researchers' methodological toolbox with the most accurate methods for selecting predictor variable weights in a cross-validated context. Subsequently, the weights chosen should yield greater classification accuracy for specific real data sets under investigation.

## References

Darlington, R. B. (1978). Reduced variance regression. *Psychological Bulletin*, *85*, 1238-1255.

Morris, J. D. (1982). Ridge regression and some alternative weighting techniques: A comment on Darlington. *Psychological Bulletin*, *91*, 203-210.

Morris, J. D. (1983). Stepwise ridge regression: A computational clarification. *Psychological Bulletin*, *94*, 363-366.

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

Send correspondence to: Mary G. Lieberman<br>
Florida Atlantic University<br>
Email: mlieberm@fau.edu

**Table 1**. Prediction Methods' PRESS Performance: Proportion of Hits

| # | Data Set Source | | | | Method | | | |
|---|---|---|---|---|---|---|---|---|
| | | D | k | N/p | OLS | Ridge | PC | LR |
| 1 | Fisher 1 & 3 | 13.97 | 0.001 | 100/4 | 1.00 | 1.00 | 1.00 | |
| 2 | Fisher 1 & 2 | 10.16 | 0.002 | 100/4 | 1.00 | 1.00 | 1.00 | |
| 3 | Bisbey 1 & 3 | 5.12 | 0.033 | 72/13 | 0.97 | 0.97 | 0.92 | |
| 4 | Fisher 2 & 3 | 3.77 | 0.012 | 100/4 | 0.97 | 0.97 | 0.81 | 0.97 |
| 5 | Rulon 1 & 3 | 2,93 | 0.010 | 152/3 | 0.93 | 0.93 | 0.91 | 0.91 |
| 6 | Bisbey 1 & 2 | 2.89 | 0.071 | 116/13 | 0.89 | 0.89 | 0.87 | 0.88 |
| 7 | Bisbey 2 & 3 | 2.41 | 0.099 | 118/13 | 0.89 | 0.87 | 0.87 | 0.89 |
| 8 | Talent 3 & 5 | 1.97 | 0.164 | 127/14 | 0.79 | 0.79 | 0.79 | 0.80 |
| 9 | Demographics #2 | 1.88 | 0.034 | 279/8 | 0.79 | 0.79 | 0.79 | 0.79 |
| 10 | Rulon 2 & 3 | 1.87 | 0.023 | 159/3 | 0.83 | 0.84 | 0.84 | 0.82 |
| 11 | Rulon 1 & 2 | 1.74 | 0.022 | 179/3 | 0.81 | 0.81 | 0.80 | 0.80 |
| 12 | Talent 1 & 5 | 1.72 | 0.116 | 177/14 | 0.75 | 0.75 | 0.72 | 0.73 |
| 13 | Demographics #3 | 1.36 | 0.064 | 279/8 | 0.73 | 0.72 | 0.66 | 0.74 |
| 14 | Talent 1 & 3 | 0.89 | 0.839 | 116/14 | 0.62 | 0.70 | 0.70 | 0.62 |
| 15 | Block 3 & 4 | 0.85 | 0.307 | 76/4 | 0.67 | 0.67 | 0.58 | 0.69 |
| 16 | Block 1 & 2 | 0.84 | 0.308 | 77/4 | 0.66 | 0.67 | 0.69 | 0.66 |
| 17 | Block 1 & 4 | 0.81 | 0.325 | 78/4 | 0.58 | 0.59 | 0.55 | 0.58 |
| 18 | Block 1 & 3 | 0.74 | 0.387 | 78/4 | 0.62 | 0.60 | 0.58 | 0.60 |
| 19 | Warncke 1 & 3 | 0.69 | 0.950 | 105/10 | 0.61 | 0.58 | 0.59 | 0.61 |
| 20 | Block 2 & 3 | 0.64 | 0.550 | 75/4 | 0.55 | 0.55 | 0.59 | 0.55 |
| 21 | Block 2 & 4 | 0.52 | 0.814 | 75/4 | 0.59 | 0.59 | 0.64 | 0.59 |
| 22 | Demographics #1 | 0.50 | 0.477 | 279/8 | 0.59 | 0.58 | 0.57 | 0.58 |
| 23 | Warncke 1 & 2 | 0.48 | 1.749 | 112/10 | 0.47 | 0.54 | 0.58 | 0.48 |
| 24 | Warncke 2 & 3 | 0.45 | 2.635 | 87/10 | 0.41 | 0.46 | 0.43 | 0.42 |