A Prediction Model for BackPack Programs

D. N. Black, L. Liu, S. Kim, and L. Davis

Abstract—**To help solve the problem of child food insecurity, school backpack programs supply schoolchildren with food to take home on weekends and holiday breaks when school cafeterias are unavailable. It is important to assess and identify the true needs of the children in schools in order to avoid any potential negative effects. This study utilizes linear regression analysis on the data from a backpack program and the data from the schools it serves. The study reveals that the percentage of low income is a significant factor. Through various feature selection methods, a prediction model is obtained, which is then employed to create a backpack needs ranking system for schools in the county not currently being serviced by the backpack program.**

Index Terms—**Backpack Programs, Child Food Insecurity, Data Analysis, Linear Regression Analysis, Prediction Modeling.**

I. Introduction

Food insecurity, defined by the lack of access to healthy food options, is a major world problem that deserves serious attention and analysis. It follows that child food insecurity is a matter of even greater importance than food insecurity of the general population. This has been confirmed by research studies [1] which reveal that food insufficiency is a persistent problem in the United States. In 2017, an estimated one in eight Americans were food insecure, equating to 40 million Americans including more than 12 million children [2]. In fact, more than 46 million people still turn to the Feeding America network each year for extra support.

Child food insecurity is associated with a range of negative developmental consequences, including behavior problems, poor health [3], poor school performance, absenteeism at school, altered daily activities [4], less healthy diets, and inadequate intake of micronutrients such as calcium, iron, and zinc [4]. Food insecurity among African American and Hispanic American children is common [3], as prevalence rates among these minorities constantly exceed the national average.

This project was supported by NSF National Research Traineeship Project Improving Strategies for Hunger Relief and Food Security using Computational Data Science (Award No. DGE-1735258).

D. N. Black, Department of Mathematics, North Carolina Agricultural & Technical State University, Greensboro, NC 27411 USA (e-mail: blackd@aggies.ncat.edu). Presenter

L. Liu, Department of Mathematics, North Carolina Agricultural & Technical State University, Greensboro, NC 27411 USA (e-mail: lliu@ncat.edu). Contact author

S. Kim, Department of Mathematics, North Carolina Agricultural & Technical State University, Greensboro, NC 27411 USA (e-mail: skim@ncat.edu).

L. Davis, Department of Industrial and Systems Engineering, North Carolina Agricultural & Technical State University, Greensboro, NC 27411 USA (e-mail: lbdavis@ncat.edu).

The first documented school-based food backpack program (BPP) began in Arkansas in 1994. A school nurse observed the trend of many children arriving at school on Mondays tired and hungry, which impeded their abilities to learn [5]. The general BPP model is simple to understand. Teachers, school staff, and sometimes parents express concern for schoolchildren suspected of experiencing constant hunger. At the end of each week, these children, with the written permission of their parents, are given backpacks or bags filled with easy-to-prepare, shelf-stable foods to combat their weekend hunger. Children take the food home, happily eat it, and return the empty backpack to school on Monday for a refill at the end of the week [6].

However, not much information is known about recipients of BPP services, their experiences, or their personal impacts as a result of their BPP participation. In considering all these aspects, a recent study [6] assesses the BPP model as it currently exists, concluding that BPPs fit poorly with the needs of most food-insecure children in America, as the current structure of its model enhances risks of detrimental effects related to worry, shame, and family functioning disarray.

Based on the existing literature on food insecurity in children and BPPs, we propose to develop a prediction model that will estimate the BPP need of schools not currently receiving services. We choose the possible factors for our prediction model as percentages of economically disadvantaged students, percentages of students severely absent from school, rural / urban school locations, percentages of African-American and Hispanic students, Title I school statuses, and percentages of low-income students.

II. Methodology

A. Data Collection

Data sets for this study were collected from a backpack program in NC, the district offices of the Guilford County Schools (GCS) public school system, and the State of North Carolina Department of Public Instruction.

For the purposes of this study, it was decided that the numbers of food bags distributed to each school was the most relevant for analysis, thus used in this study as the response variable. Public data collected from the GCS district website [7] [8] included School Information Dashboard data such as race/ethnicity percentages, chronic absence rates, and lists of Title I schools defined as follows:

> Title I, Part A (Title I) of the Every Student Succeeds Act provides financial assistance to school districts and schools with high numbers or high percentages of children from low-income families to help ensure that all

children meet challenging state academic standards. [8] Also, the State of NC Department of Public Instruction provided data with percentages of economically-disadvantaged [9] and low-income students [10].

B. Statistical Modeling

This study utilized the predictive modeling method of multiple linear regression, an extension of the simple linear regression model that could be used for several predictors simultaneously [11]. For p predictors, the full multiple linear regression model is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p = \hat{\beta}_0 + \left(\sum_{i=1}^{p} \hat{\beta}_i X_i \right) \#(1)$$

for estimated response $\hat{Y}$, coefficient estimates $\hat{\beta}_0, \ldots, \hat{\beta}_p$, and predictors $X_1, \ldots, X_p$.

Data set columns most relevant to study outcomes were combined in one data file and used as predictor variables. These predictors, measured for each Guilford County school, were *PctEDS* (percentage of economically-disadvantaged students), *PctSevAbs* (percentage of severely absent students), *RurUrb* (0 for rural schools, 1 for urban schools), *PctMinority* (sum of percentages of Hispanic and Black students), *TitleI_NoYes* (0 for schools not receiving Title I funding, 1 for schools receiving Title I funding), and *PctLowInc* (percentage of low-income students). The response, *Score*, represented the need of each school for BPP services. *Score* values were included for schools already receiving BPP services based on the total number of food bags provided to those schools for the school year. Full data sets from 2016-2017 and 2017-2018 school years were separated into training data and prediction data. Training data sets included schools already receiving BPP services, and prediction data sets included schools not yet receiving those services.

With these predictors, RStudio [12] linear regression function *lm* was used to estimate intercept parameter β_0 and slope parameters $\beta_1, \ldots, \beta_6$ for each of the six predictor variables. These estimates were used to construct full model equations. Diagnostic checks were performed on each full model with Residual vs Fitted plots, Normal Quantile-Quantile plots, Scale-Location plots, and Residual vs Leverage plots to confirm that all classical assumptions of linear regression for sample size *n* were satisfied.

Before variable selection was performed, separate checks for collinearity were necessary. Collinearity occurred when at least two predictor variables were very closely associated [11]. Comparisons were made with the variance inflation factor (VIF). VIF values greater than 5 or 10 indicated levels of collinearity that could lead to trouble in linear regression. For the purposes of this study, predictor variables with VIF > 10 were deleted from the full model for that data set and would not be considered in any variable selection methods.

Backward selection and forward selection processes considered main effect subset models using traditional p-value criteria. Backward, forward, and stepwise selection would also be performed using corrected Akaike Information Criterion (AIC_c) for small samples [13]. Interaction effects were considered during stepwise selection. After forward, backward, and stepwise variable selection methods (using p-values and AIC_c) narrowed these potential subset models down to a few finalists, then model comparison methods were used to select the best of the best. The random forest method, adjusted R^2 comparisons, and *F*-statistic p-value comparisons were used to measure the quality of fit of subset models, while discouraging overfitting – the use of unnecessarily complex models to explain random error in the data.

Each finalist model was used on the prediction data set to generate tables of predicted scores, which were then sorted by score in descending order. Distributions of these predicted scores were plotted on histograms for side-by-side comparisons of the finalist models. With several well-defined methods and criteria in place, the best models for each training data set were selected. Data visualization generated for these best models included training data scatterplots of best-model predictors, prediction data scatterplots of predicted scores by school name, and prediction data histograms of best-model predictors.

III. Results and Discussions

GCS data were imported in RStudio [12]. Each data set was separated into training and prediction data as described in Chapter II. Training data for 2016-2017 and 2017-2018 school years were of sizes 20 and 21, respectively. Thus, sample size to predictor ratios $\frac{n}{p}$ for all training data sets were small, which justified the use of small-sample variable selection criteria and methods described in the previous chapter.

A. 2016-2017 School Year Analysis

Initial diagnostics on full model *Score* ~ *PctEDS* + *PctSevAbs* + *RurUrb* + *PctMinority* + *TitleI_NoYes* + *PctLowInc* revealed high VIF values for *PctMinority* (11.722406) and *PctLowInc* (10.767096). Since the highest value belonged to *PctMinority*, this variable was removed from the full model. This yielded new VIF values less than 10 (*PctEDS* = 1.672238, *PctSevAbs* = 2.130364, *RurUrb* = 1.070120, *TitleI_NoYes* = 3.686646, *PctLowInc* = 6.083727) for the five remaining predictors.

After performing manual backward selection, backward selection with AIC_c, manual forward selection, forward selection with AIC_c, and stepwise selection with AIC_c, one model emerged as the unanimous best. This model was *Score ~ PctLowInc*. Thus, random forest, adjusted R^2, and *F*-statistic p-value comparisons were not required for 2016-2017 analysis.

Diagnostic plots were generated for this model to check for the validity of linear regression assumptions, as shown in Fig. 1. The Residual vs Fitted plot showed randomness in residuals centered around a line with some decrease and increase, but a generally horizontal pattern. This was not the most ideal result desired, but the line was not distorted extremely enough to give reason for concern. The Normal Q-Q plot showed the greatest deviation from normality in the first few quartiles, but a vast majority of the standardized residual quartiles closely aligned with the quartiles of $N(0, \sigma^2)$. Scale-Location plot analysis (not shown) was similar to that of Residual vs Fitted;

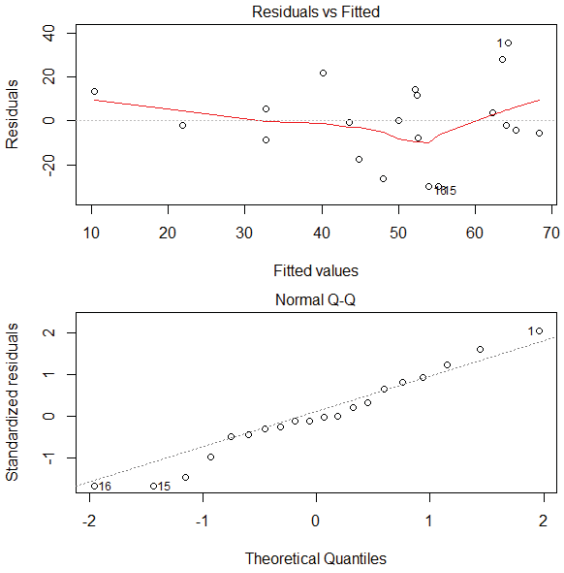

Fig. 1. Diagnostic plots for the best model for data 2016-2017.

randomness among plot points, while not the most ideal, were evident. Finally, Residuals vs Leverage (not shown) revealed all points within the Cook's Distance threshold (less than 0.5); therefore, no influential points existed within the 2016-2017 training data set for this model.

A table of predicted scores was generated using best model *Score ~ PctLowInc*, and the histogram of the distribution of those predicted scores was plotted in Fig. 2. This histogram showed an approximately symmetrical distribution of scores predicted from the test data set with a median score in the 30-40 range. Linear regression output allowed us to generate the equation used to derive the predicted scores from this histogram. Thus, predictions of scores for 2016-2017 school year test data were best approximated with the formula

$$Score = -0.4495 + 86.9292 * (PctLowInc). \#(2)$$

B. 2017-2018 School Year Analysis

Full model *Score ~ PctEDS + PctSevAbs + RurUrb + PctMinority + TitleI_NoYes + PctLowInc* initial diagnostics had high VIF values for *PctMinority* (9.858125), *TitleI_NoYes* (9.049403) and *PctLowInc* (9.366301). To combat potential collinearity issues, the predictor with the largest VIF greater than 10 would be removed from the full model. The VIF values of the predictors of this full model did not exceed 10. However, the VIF for *PctMinority* was very close to 10, so it was removed. VIF was recalculated for the new full model, and all remaining predictors had satisfactory VIF results (*PctEDS* = 6.397304, *PctSevAbs* = 2.358749, *RurUrb* = 1.068436, *TitleI_NoYes* = 8.395424, *PctLowInc* = 4.598785).

Manual backward selection, backward selection with AIC_c, manual forward selection, forward selection with AIC_c, and stepwise selection with AIC_c were then performed on 2017-2018 training data. Three models emerged from these methods as best model finalists. Final Model 1, *Score ~ PctSevAbs + TitleI_NoYes*, was the best model from backward selection with AIC_c and stepwise selection with AIC_c. Final Model 2, *Score ~ PctLowInc*, was the best model from manual forward selection and forward selection with AIC_c. Also recall that this model was the best model from the previous school year. Final Model 3, *Score ~ PctSevAbs*, was the best model from manual backward selection.

To choose the best model, these three finalists were compared by random forest, adjusted R^2, and *F*-statistic p-value computations. Final Model 1 had the highest adjusted R^2 value, while Final Model 2 was the preferred model from random forest and had the lowest *F*-statistic p-value.

After carefully considering all output, plots, and tables, Final Model 2 was chosen as the best model to approximate the observed 2017-2018 training data. Therefore, the best model equation used for score predictions for 2017-2018 school year test data was

$$Score = 14.84 + 57.16 * (PctLowInc). \#(3)$$

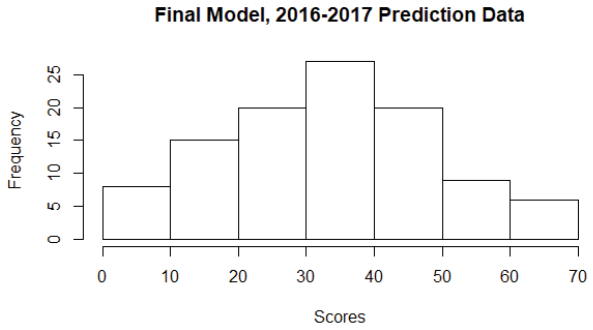

Fig. 2 Histogram of predicted scores, 2016-2017 best model.

C. Ranking of Schools

The prediction output from R clearly ordered the schools not currently receiving BPP services from highest score (greatest need for BPP) to lowest score (least amount of need for BPP). The top ten schools – the schools with the greatest need – were especially observed for future discussion, as shown in Tables I and II.

Comparing elementary school rankings in these tables reveals that among elementary schools, Fairview Elementary has the greatest food insecurity need, followed by Bessemer Elementary, Waldo C Falkener Sr Elementary, and Kirkman

TABLE I
TOP TEN SCHOOLS, PREDICTIONS FROM 2016-2017 DATA

Rank	School	Score
1	Scale School [Alternative]	64.74735856
2	Kirkman Park Elementary [Elementary]	62.26987685
3	Otis L Hairston Sr Middle [Middle]	61.87000261
4	Bessemer Elementary [Elementary]	61.85261677
5	Fairview Elementary [Elementary]	61.5309788
6	Waldo C Falkener Sr Elementary [Elementary]	61.04417537
7	Jackson Middle [Middle]	59.44467841
8	Newcomers School [Alternative]	57.02804712
9	Murphey Traditional Academy [Elementary]	56.75856666
10	Scale in High Point (Pruette Scale) [Alternative]	56.24568448

TABLE II
TOP TEN SCHOOLS, PREDICTIONS FROM 2017-2018 DATA

Rank	School	Score
1	Fairview Elementary [Elementary]	60.87585242
2	Waldo C Falkener Sr Elementary [Elementary]	60.80726623
3	Bessemer Elementary [Elementary]	60.33859393
4	Bluford Elementary [Elementary]	58.82398222
5	Scale School [Alternative]	58.14383582
6	Otis L Hairston Sr Middle [Middle]	57.97237035
7	Kirkman Park Elementary [Elementary]	56.9835861
8	Jackson Middle [Middle]	55.98908633
9	Hunter Elementary [Elementary]	54.74310387
10	Julius I Foust Elementary [Elementary]	54.69166422

Park Elementary, in that order. Clearly, most schools in Tables I and II were elementary schools, but other school levels were also present. Scale School (an alternative school) and middle schools Otis L Hairston Middle and Jackson Middle were on both lists. Among these schools, it is recommended that priority is given to Scale School, followed by Otis L Hairston Middle and then Jackson Middle.

IV. CONCLUSION AND FUTURE RESEARCH

Data from the school district office website and from the North Carolina State Board of Education Department of Public Instruction website were collected and consolidated. Simple and multiple linear regression methods were used to express trends from this data as mathematical models, which could then be used for prediction analysis. It was determined that *PctLowInc* was the most relevant variable needed to predict scores for 2016-2017 and 2017-2018 data. Once ordered lists of predicted scores were generated, feasible solutions were suggested to the backpack program to help maximize the impact of its community outreach. These solutions included student participation in focus groups, thorough end-of-school-year evaluations, strengthened relationships with school administrators (especially school social workers), and periodic data monitoring.

Additional research is necessary to provide more insight and perspective to this topic. First, this analysis can be extended to investigate a large area such as a whole state including many backpack programs for the schools in state. A deeper research should be done to find out other food assistance programs that are also offered in schools. This is an important factor because food assistance programs provided by other local agencies decrease the need for backpack services at those schools. Also, there may be other nonlinear regression methods that are more appropriate to use for modeling these data sets. Perhaps the best linear regression models could be compared with the best nonlinear regression models for prediction accuracy. Finally, visualization methods could be used to show priority regions. A heat map could take county food deserts into consideration when assigning ranks to schools. Further research with data visualization can give a more holistic view of the child food insecurity situation in the county.

REFERENCES

[1] Alaimo, K., Briefel, R.R., Frongillo, E.A., Jr., and Olson, C.M., "Food insufficiency exists in the United States: Results from the third National Health and Nutrition Examination Survey (NHANES III)," *American Journal of Public Health,* vol. 88, no. 3, pp. 419-426, Mar. 1998.

[2] Coleman-Jensen, A., Rabbitt, M.P., Gregory, C.A., and Singh, A., "Household food security in the United States in 2017," United States Department of Agriculture, Economic Research Service, Washington, DC, USA, Rep. 256, Sep. 2018.

[3] Casey, P.H., Szeto, K.L., Robbins, J.M., Stuff, J.E., Connell, C., Gossett, J.M., and Simpson, P.M., "Child health-related quality of life and household food security," *Archives of Pediatrics and Adolescent Medicine*, vol. 159, no. 1, pp. 51-56, Jan. 2005.

[4] Bernal, J., Frongillo, E.A., Herrera, H.A., and Rivera, J.A., "Food insecurity in children but not in their mothers is associated with altered activities, school absenteeism, and stunting," *The Journal of Nutrition*, vol. 144, no. 10, pp. 1619-1626, Oct. 2014.

[5] Education World, "For hungry kids, backpacks lighten load," Colchester, CT, USA, n.d. [Online]. Available: https://www.educationworld.com/a_admin/admin/admin495.shtml

[6] Fram, M.S. and Frongillo, E.A., Jr., "Backpack programs and the crisis narrative of child hunger: A critical review of the rationale, targeting, and potential benefits and harms of an expanding but untested model of practice," *Advances in Nutrition*, vol. 9, no. 1, pp. 1-8, Jan. 2018.

[7] Guilford County Schools, "Enrollment (3 year) snapshot; Chronic Absences," Greensboro, NC, USA, 2019. [Data file; Online]. Available: https://www.gcsnc.com/Page/44123

[8] Guilford County Schools, "About Title I," Greensboro, NC, USA, 2019. [Online]. Available: https://www.gcsnc.com/domain/5042

[9] North Carolina State Board of Education, Department of Public Instruction, "Free & reduced meals application data," Raleigh, NC, USA, 2018. [Online]. Available: http://www.ncpublicschools.org/fbs/resources/data/#meal-application

[10] North Carolina State Board of Education, Department of Public Instruction, "Title I schools," Raleigh, NC, USA, 2018. [Online]. Available: http://www.ncpublicschools.org/program-monitoring/titleIA/

[11] James, G., Witten, D., Hastie, T., and Tibshirani, R., *An introduction to statistical learning with applications in R*. New York, NY, USA: Springer, 2013.

[12] R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria. [Online]. Available: https://www.R-project.org

[13] Burnham, K.P. and Anderson, D.R., "Second-order information criterion: 1978," in *Model selection and multimodel inference: A practical information-theoretic approach*, 2nd ed., New York, NY, USA: Springer, 2002, ch. 2, sec. 4, pp. 66-67.