# Statistical Analysis

## Contents

## 1 Data

The data provided consisted of 111 observations which were each comprised of 23 variables. The variables `ATTEND`, `BEHAVE2`, `LETREC2C`, `LETREC2L`, and `NOREC2` were discarded because they contained information which would not be available at the time the parent was attempting to determine whether a prospective student is ready for kindergarten.

To reduce the effects of sparseness in the modeling process, the remaining variables were grouped into three areas: home situation, academic indicators, and testing effects. A short description of each of these groups follows.

### 1.1 Home Situation

The variable `BRTHORDR` is a discrete variable giving the birth order of the child. `TOTCHILD` is a another discrete variable indicating the total number of children in the household.

`GENDER` is a categorical variable which was coded as 0 for *females* and 1 for *males*. `RACE` is another categorical variable which was coded as 1 for *black*, 2 for *white*, 3 for *hispanic*, 4 for *Pacific islander*, 5 for *Asian*, 6 for *Filipino*, and 7 for *other*.

The variable `BILING` is a categorical variable representing whether the child is eligible for bilingual instruction (0=*no*, 1=*yes*), and was viewed as representing whether more than one language was commonly used in the home.

The final variable in this group is `AGE1290` which was computed from `BDAYMO` and `BDAYYR`. It represents the age of the child, in months, in December of 1990.

### 1.2 Academic Indicators

The variables LETREC1C and LETREC1L are discrete variables representing the number of letters (capital and lower case respectively) recognized by the child. Similarly, NOREC1 is a discrete variable representing the number of numbers recognized.

ROUND2 is a categorical variable (0=*no* and 1=*yes*) which indicates whether the student will be in kindergarten for a second time—*i.e.* the student was previously retained.

Finally, BEHAVE1 is an ordinal variable (1=*good*, 2=?, 3=?, and 4=*bad*) representing the teacher's perception of the students behavior at entry. While this value was not assessed prior to the child's entry, it could easily be assessed earlier.

### 1.3 Testing Effects

Information on both teacher and school were recorded to control for these effects. The variable TEACHER is a categorical variable which indicates which of the four teachers the child was assigned to. SCHOOL is a categorical variable used to indicate which school the child attended. Due to the method of sampling—only one teacher was used at Ramona—teacher three was confounded with school two.

### 1.4 Retention

The most important variable in the group was RETAINED. This dichotomous categorical variable (0=*no* and 1=*yes*) indicated whether the teacher felt the student should be retained. Actual retention could not be used as the outcome because of legal issues.

## 2 Simple Descriptives

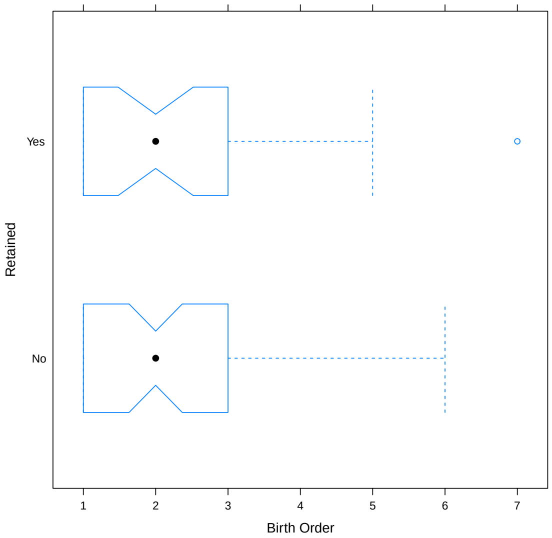
Simple descriptive statistics and plots were generated to help in defining potential significant effects. It is important to remember, particularly when viewing the boxplots, that the variable which is being modeled is RETAINED.

### 2.1 Interval-Ratio

The information in the interval-ratio variables is best viewed through boxplots.

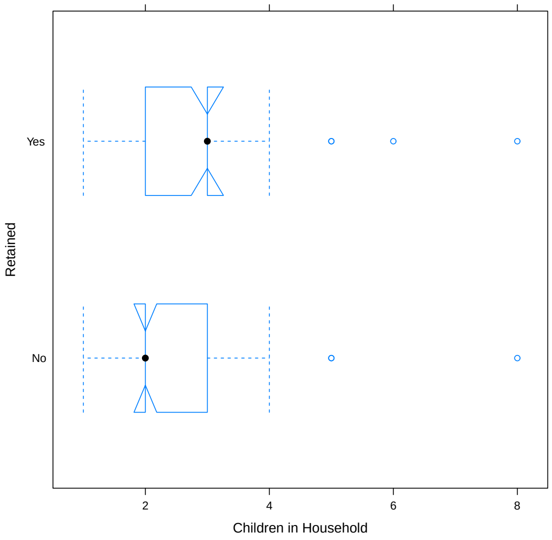
#### 2.1.1 BRTHORDR

The boxplot of BRTHORDR shows none of the significance that the logistic model attributes to it.



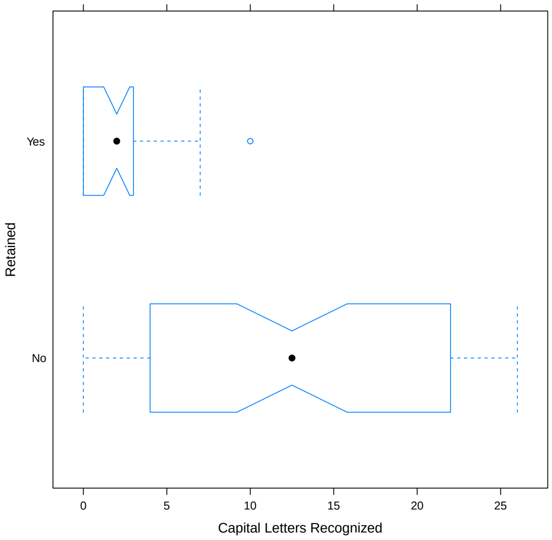
### 2.1.2 TOTCHILD

The boxplot of TOTCHILD indicates that there may be some information about retention in the variable.



### 2.1.3 LETREC1C

LETREC1C appears to be very indicative of later retention.



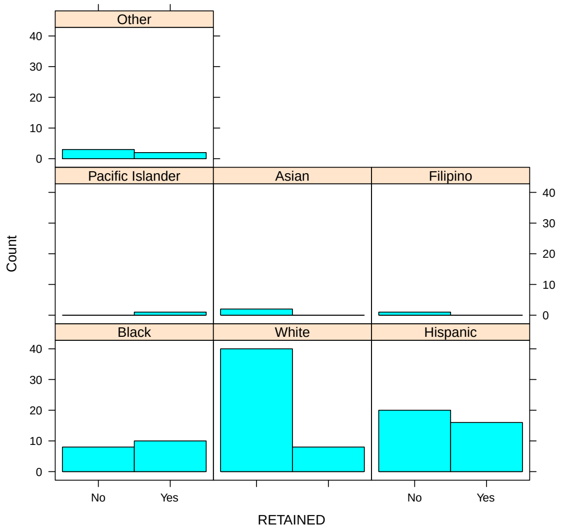
## 2.2 Categorical

Bar graphs and tables are useful for looking at trends in categorical variables.

### 2.2.1 RACE

The information in RACE is apparent when looking at RACE=2 but the others are less clear.

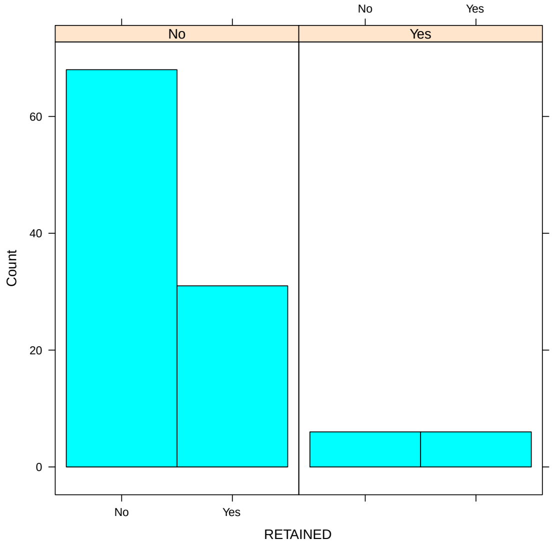
| | RACE | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|---|
| RETAINED | 0 | 8 | 40 | 20 | 0 | 2 | 1 | 3 | 74 |
| | | 10.8 | 54.1 | 27.0 | 0.0 | 2.7 | 1.4 | 4.1 | 100.0 |
| | | 44.4 | 83.3 | 55.6 | 0.0 | 100.0 | 100.0 | 60.0 | 66.7 |
| | 1 | 10 | 8 | 16 | 1 | 0 | 0 | 2 | 33 |
| | | 27.0 | 21.6 | 43.2 | 2.7 | 0.0 | 0.0 | 5.4 | 100.0 |
| | | 55.6 | 16.7 | 44.4 | 100.0 | 0.0 | 0.0 | 40.0 | 33.3 |
| | | 18 | 36 | 1 | 2 | 1 | 5 | 111 | |
| | | 16.2 | 43.2 | 32.4 | 0.9 | 1.8 | 4.5 | 100.0 | |



### 2.2.2 ROUND2

Again, there appears to be some information in ROUND2.

| | ROUND2 | 0 | 1 | |
|---|---|---|---|---|
| RETAINED | 0 | 68 | 6 | 74 |
| | | 91.9 | 8.1 | 100.0 |
| | | 68.7 | 50.0 | 66.7 |
| | 1 | 31 | 6 | 37 |
| | | 83.8 | 16.2 | 100.0 |
| | | 31.3 | 50.0 | 81.3 |
| | | 99 | 12 | 111 |
| | | 89.2 | 10.8 | 100.0 |



# 3 Area Models

As noted above, the three areas were modeled separately to reduce the effects of sparseness in the logistic regression model.[1] Each group of variables was subjected to both stepwise and backward variable selection techniques. Variables which survived the selection process were then tested for interactions.

## 3.1 Teaching Effects

None of the teaching effects variables survived the selection process. This indicates that there is no evidence that the retention rate varied between schools or teachers.

## 3.2 Home Situation

Model selection techniques generated

$$\ln\left(\frac{p}{1-p}\right) = -0.3400 - 0.6899\texttt{RACE2} - 1.5018\texttt{BRTHORDR} + 0.6113\texttt{TOTCHILD}$$

where `RACE2` took on 1 if the student was *white* and 0 otherwise. While the overall model was significant ($-2\ln(l) = 16.440$, $df = 3$, $p = 0.0009$; score=15.753, $df = 3$, $p = 0.0013$) these variables were not predictive in the presence of other factors—see below.

## 3.3 Academic Indicators

The stepwise and backward variable selection methods produced the following model:

$$\ln\left(\frac{p}{1-p}\right) = 1.2442 - 0.4910\texttt{LETREC1C} + 3.7168\texttt{ROUND2}$$

[1] Logistic regression is a technique which allows for the fitting of multidimensional contingency tables. While dichotomous dependent variables are the norm, logistic regression may also be used to fit ordinal outcomes. Another advantage of the use of logistic regression is its ability to incorporate interval-ratio (discrete or continuous) covariates into otherwise purely categorical models.

Fleiss[2] discusses the relationship of logistic models and odds-ratios. For information and logistic regression modeling and validation see Hosmer and Lemeshow[3]. Agresti[1] gives a more technical, but thorough description of the various forms of logistic regression.

As will later be shown more completely, this model indicates that the odds of those students who are in kindergarten for a second time needing retention at the end of the year are approximately 40 times higher than otherwise equivalent students. In addition, an increase of a single recognized capital letter is associated with an approximate 1.5 *decrease* in the odds of retention.

# 4 Final Model

Using the variables selected in the above models, a general main effects model was fitted. Both forward and backward selection methods generated:

$$\ln\left(\frac{p}{1-p}\right) = 1.2442 - 0.4910\texttt{LETREC1C} + 3.7168\texttt{ROUND2}$$

This model is obviously identical to the model generated in the Academic Indicators section above.

Additional models which included various interaction effects showed no additional significant effects. Both model significance statistics ($-2\ln(l) = 64.329$,$df = 2$,$p = 0.0001$;score=39.971,$df = 2$,$p = 0.0001$) and residual analysis suggested that the model was valid.

This model also supports some reasonable descriptive probabilities.

- Correct: $P(R+\cap M+) + P(R-\cap M-) = 0.8378$.
- Sensitivity: $P(M+|R+) = 0.8378$.
- Specificity: $P(M-|R-) = 0.8378$.
- False Positive: $P(R-|M+) = 0.2781$.
- False Negative: $P(R+|M-) = 0.0882$.

# 5 Interpretation

The final model generates the following statistics:

| | df | $\hat{\beta}$ | $\widehat{se(\beta)}$ | $\chi^2$ | $p$ | $\hat{\omega}$ | 95% CI for $\omega$ |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 1.2242 | 0.4220 | 8.4178 | 0.0037 | | |
| LETREC1C | 1 | -0.4910 | 0.1348 | 13.2604 | 0.0003 | 0.612 | $[0.47, 0.80]$ |
| ROUND2 | 1 | 3.7168 | 1.4105 | 6.9439 | 0.0084 | 41.133 | $[2.59, 652.87]$ |

The estimated odds-ratios[2] indicate that, all other things being equal, the odds of a child who is going into kindergarten for a second time being retained are 41.133 times higher than a child who was not previously retained. This indicates that either these children have severe learning disabilities, kindergarten is not a useful learning experience, ROUND2 students don't get enough home support, or something is weird in Moreno Valley.

Similarly, all other things being equal, the odds of a being retained are *decreased* by a factor of $\frac{1}{0.612} = 1.634$ for every capital letter they recognize at the time they enter kindergarten. This is especially important because it is a simple indicator which parents can use for assessing their child's preparation and probability of success (Table 1).

A bit of a warning is needed here. Use of these estimated probabilities is based upon the child's attributes at the time that Table 1 is used. A parent who then goes home and forces his/her child to learn a few more letters will not be able to use the table to reassess the child's probability of success.

[2] See Fleiss[2] for a complete description of the odd-ratio. For now we will define it to be $\omega = \frac{P(R+|M+)P(R-|M-)}{P(R+|M-)P(R-|M+)}$.

| Letters | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ROUND2: No | 0.68 | 0.56 | 0.44 | 0.32 | 0.23 | 0.15 | 0.10 | 0.06 | 0.04 | 0.02 | 0.02 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| ROUND2: Yes | 0.99 | 0.98 | 0.97 | 0.95 | 0.92 | 0.88 | 0.82 | 0.73 | 0.63 | 0.51 | 0.39 | 0.28 | 0.19 | 0.13 | 0.08 | 0.05 | 0.03 | 0.02 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

Table 1: Probability of need for retention by number of capital letters recognized and previous retention

## 6 Future Work

Categorical data analysis is inherently sensitive to sparseness (low cell counts) in the design. While the methods used in analyzing the data presented herein were chosen to help reduce sparseness, obtaining additional observations would be helpful.

Only one teacher from Ramona was available. Because of this, teacher and school effects were confounded. Obtaining information from another teacher at the same school would alleviate this problem.

The predictive ability of the final model cannot be accurately assessed when using the data which generated the model. Collecting more data to be used solely for confirmation of the predictive qualities of the model is advised.

## References

[1] A. Agresti. *Categorical Data Analysis.* John Wiley & Sons, New York, 1990.

[2] J. L. Fleiss. *Statistical Methods for Rates and Proportions.* John Wiley & Sons, New York, 1973.

[3] David W. Hosmer, Jr. and Stanley Lemeshow. *Applied Logistic Regression.* John Wiley & Sons, New York, 1989.