Updated Statistical Models for the Florida APD's iBudget Algorithm

Xu-Feng Niu Ph.D
Professor and Chair
Department of Statistics
Florida State University
Tallahassee, FL 32306-4330

Minjing Tao, Ph.D
Assistant Professor
Department of Statistics
Florida State University
Tallahassee, FL 32306-4330

December, 2014 – June, 2015

Task-Two Technical Report Submitted to the
Florida Agency for Persons with Disabilities

Note: Final recommended model for the new algorithm: Model 5b, which was previously labeled as Model 7b in other documents.

I. Background and Introduction

The Florida Agency for Persons with Disabilities (APD) serves individuals with intellectual disabilities, autism, cerebral palsy, spina bifida, Down syndrome, and Prader-Willi syndrome as well as children at risk of a developmental disability. Based on the information provided by the APD, "The majority of individuals served live in the community with family members, in their own home, or in a congregate living setting such as a group home. APD provides services such as physical therapy, respite, residential habilitation, and supported employment to support these individuals in living, learning, and working in their communities. Approximately 30,600 individuals receive services through a Medicaid waiver, and over 20,800 are waiting for waiver services."

In response to increasing need, concerns about the old budget allocation system, and a mandate from the Florida Legislature, in 2010 APD developed a new plan for serving its waiver-enrolled consumers. The new plan was based on the statistical algorithm developed by Niu and Bell (2010), which is currently used to calculate individuals' base budget allocations for services received under the Developmental Disabilities Home and Community Based Services (HCBS) Individual Budgeting Waiver, also known as iBudget Florida.

On May 1, 2011, APD began implementation of iBudget Florida. The iBudget Florida waiver uses an individual budgeting approach and is intended to enhance the simplicity, sustainability, and equity of the system while also increasing individuals' opportunities for self-direction. As of July 1, 2013 APD transitioned all waiver clients to the iBudget Florida waiver statewide, completing the phasing out of the HCBS Tier Waiver system.

Niu and Tao (2014) evaluated the current iBudget algorithm based on the FY 13-14 data. Specifically, they examined the Florida iBudget algorithm using the baseline data from July 1, 2013 to June 30, 2014, where the Square-Root of the FY 13-14 Claim is the response variable, and the independent variables (predictors) are 1) Living setting with four levels; 2) A two-level dummy variable for Age with **AgeI=0** for consumers 20-year-old or younger; **AgeI=1** for consumers 21 and over; 3) Sum of behavioral status raw score with individual item scores (Q14-Q24) from the Consumers' Questionnaire for Situational Information (QSI); 4) Sum of functional status raw score

(Q25-Q30); 5) Q18 (supports needed for transfer); 6) Q20 (supports needed to maintain hygiene); and 7) Q23 (supports needed for self-protection). Niu and Tao (2014) found that the current iBudget algorithm fitted very well to the FY 13-14 Claim data. Results from the statistical models showed that the R-squared values of the regression models based on the FY 13-14 Claim data (both before and after removing about 10% potential outliers) are significantly higher than those based on the FY 07-08 Claim data. On the other hand, since Niu and Tao (2014) only examined the iBudget model developed by Niu and Bell (2010), it is not sure whether the current algorithm is the best model for the new data or not. For further evaluation and modification of the iBudget Florida program, we have completed the following analysis and model development based on the FY 13-14 Claim data:

Task2:

- Determine and refine dependent variables;
- Determine and refine independent variables;
- Develop a model(s) that achieves Agency goals and objectives;
- Develop a method for identifying outliers;
- Assess and provide recommendations for improving data integrity;
- Test the accuracy and reliability of the model(s) and provide recommendations for improving accuracy and reliability;
- Perform other statistical analyses as needed to develop a model that achieves Agency goals and objectives;
- Review, evaluate and provide recommendations for improving the final model recommended by the Agency;
- Additional tasks as may be required.

This report is organized as follows. Section II introduces statistical methods used in this analysis, which include multiple linear regression models with transformations, model selection techniques, and outlier detection. Section III discusses all the dependent variable and independent variables considered in the iBudget algorithm. Specifically, two main independent variables, living setting and age, are examined carefully first; then a full analysis of independent variables is conducted. The main dependent variable used in this study is the APD consumers' FY 2013-2014 expenditures with the some adjustments. Section IV presents the transformation of the response variable, model selection using the Bayesian Information Criterion (SBC), and the final model selected for the new iBudget Algorithm based on the FY 13-14 Claim data. The weights of different predictors for the response are given in Section V, along with some examples. In Section VI, fractions of the total variation in the response variable explained by different groups of independent variable are given. Section VII assesses the robustness of the final model by utilizing the bootstrapping method, in which the average weights of independent variables based on 10,000 bootstrapping samples are presented, along with the 95% bootstrapping confidence intervals for the coefficients in the selected final model. Finally, Section VIII presents our final recommended model and conclusions.

II. Statistical Methods

1. Multiple Linear Regression Models with Transformations

Linear regression is probably the most popular statistical method used in data analysis, which models the relationship between the dependent variable (or response) y and independent variables (predictors). In this study, the dependent variable is the APD consumers' FY 2013-2014 expenditures, and the independent variables include consumers' age, living setting status, and individual characteristics and support needs specified in QSI. In this section, we will discuss linear regression models and their basic statistical properties. Transformation of the response variable will be selected in order to make the response approximately normal and with constant variance.

Consider a classical multiple linear regression model with the form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i, \qquad i=1, 2, \ldots, n, \tag{1}$$

where y_i is the dependent variable, $\{x_{1i}, x_{2i}, \cdots, x_{pi}\}$ are independent variables or predictors, β_0 is the intercept, and $\{\beta_0, \beta_1, \cdots, \beta_p\}$ are unknown coefficients. The random error terms $\{\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n\}$ should satisfy the following assumptions:

1) Each term ε_i has a normal distribution.
2) $\{\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n\}$ are independent with each other.
3) Each term ε_i has the same variance σ^2.

When the assumptions on $\{\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n\}$ are satisfied, the responses $\{y_1, y_2, \cdots, y_n\}$ are also independent and have normal distributions with constant variance, σ^2. However, in practice it may be the case that one or more of these assumptions are not valid and transformations on the responses are needed to ensure the assumptions are approximately satisfied.

Consider random variables $\{y_1, y_2, \cdots, y_n\}$ with variances $\{Var(y_i) = \sigma_i^2, i=1, 2, ..., n\}$. That is, the variances of $\{y_1, y_2, \cdots, y_n\}$ are not constant. We want to find a transformation $z_i = f(y_i)$ such that the distribution of z_i is approximately normal and with constant variance $Var(z_i) = \sigma^2$. The popular **Box-Cox Power Transformation Family** will be used for this purpose. Similarly, independent variables can also be transformed to make the relationship between response and predictors linear (Weisberg, 2005; Chapter 7).

First we suppose that the observations $\{y_1, y_2, \cdots, y_n\}$ are all positive. Otherwise, we may add a positive number to each of the observations, making all observations positive. **(This operation changes the mean values of the observations, a level shift, but will not change the variance and covariance structure of the data.)**

The Box-Cox Power Transformation Family is

$$z_i^{(\lambda)} = \frac{y_i^{\lambda} - 1}{\lambda}, \quad \text{if } \lambda \neq 0; \qquad z_i^{(\lambda)} = \log(y_i), \quad \text{if } \lambda = 0. \tag{2}$$

The Box-Cox Power transformation family given in (2) is continuous about real numbers λ since we have $\lim_{\lambda \to 0} \frac{y_i^{\lambda} - 1}{\lambda} = \log(y_i)$.

When we know that a transformation is needed for the responses $\{y_1, y_2, \cdots, y_n\}$, one natural question will be how to choose a transformation in the Box-Cox Power Transformation Family. For a given λ, define

$$z_i^{(\lambda)} = \frac{y_i^{\lambda} - 1}{\lambda [GM(y)]^{\lambda - 1}}, \quad \text{if } \lambda \neq 0; \qquad z_i^{(\lambda)} = \log(y_i)[GM(y)], \quad \text{if } \lambda = 0, \tag{3}$$

where $GM(y)$ is the **geometric mean** of the observations $\{y_1, y_2, \cdots, y_n\}$, calculated $GM(y) = \left[\prod_{i=1}^{n} y_i\right]^{1/n}$ with n being the sample size. The scale adjustment by $GM(y)$ in (3) guarantees that the units of $\{z_i^{(\lambda)}, i=1, 2, \ldots, n\}$ are similar to each other for all values of λ so that different transformations can be compared. For each given λ, fit the linear model

$$z_i^{(\lambda)} = \beta_0^{(\lambda)} + \beta_1^{(\lambda)} x_{1i} + \cdots + \beta_p^{(\lambda)} x_{pi} + \varepsilon_i^{(\lambda)}, \qquad i=1, 2, \ldots, n, \tag{4}$$

obtain the residuals $\left\{\hat{\varepsilon}_1^{(\lambda)}, \hat{\varepsilon}_2^{(\lambda)}, \cdots, \hat{\varepsilon}_n^{(\lambda)}\right\}$, and calculate the Residual Sum of Squares, $RSS(\lambda) = \sum_{i=1}^{n}\left(\hat{\varepsilon}_i^{(\lambda)}\right)^2$. Then the best transformation for the responses $\{y_1, y_2, \cdots, y_n\}$ will choose λ such that $RSS(\lambda)$ reaches its minimum (Weisberg, 2005; Chapter 7).

The Box-Cox power transformation method chooses λ such that residuals from the linear model are as close to normally distributed with constant variance as possible. Therefore, after the transformation when the normality and constant variance assumptions are valid, the residual sum of squares from the model should be smaller than that based on untransformed data.

In practice, $RSS(\lambda) = \sum_{i=1}^{n}\left(\hat{\varepsilon}_i^{(\lambda)}\right)^2$ is calculated only for some special cases, such as $\lambda \in \{-3, -2.5, -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, 2.5, 3\}$.

2. Model Selection

Model selection is an important topic in statistical data analysis, which chooses one or more models from a set of candidate models. In this study, 125 independent variables are considered for building the best model. We need to identify independent variables that have significant power for predicting consumers' expenditures. The Bayesian Information Criterion (SBC) will be used to compare models and choose the best prediction model for the Florida iBudget algorithm.

Consider the linear regression model specified in (1) with y_i as the dependent variable and $\{x_{1i}, x_{2i}, \cdots, x_{pi}\}$ as the independent variables (or predictors). In practice, one or more predictors in model (1) may not be statistically significant and lack prediction power for the response, y_i.

Keeping non-significant or borderline predictors in a model will bring additional sources of noise and reduce the accuracy of predictions. When different models are fit to the observations $\{y_1, y_2, \cdots, y_n\}$, model selection techniques should be used to decide which model fits the data best. Statistical inferences such as estimation and prediction will then be based on the best model selected.

The Bayesian Information Criterion (SBC) suggested by Schwartz (1978) is one popular criterion for model comparison. For a fitted model (linear or nonlinear) with p parameters, SBC is defined as SBC(p) = -2 log(maximum likelihood function) + $p \times$ log(n). The likelihood function is based on the distribution assumption of the model such as normal, log-normal, or other distribution families. n is the sample size. When the random errors have a normal distribution, the SBC(p) has the simplified form

$$\text{SBC(p)} = n \times \log\left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / (n-p-1)\right) + p \times \log(\text{n}), \tag{5}$$

where $\hat{Y}_i$ is the fitted value based on one of the candidate models and $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ is the **Residual Sum of Squares (RSS)** based on the fitted candidate model.

Intuitively, there are two parts in (5), the first part is

$$n \times \log\left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / (n-p-1)\right) = n \times \log \hat{\sigma}^2,$$

which is a measure of the goodness-of-fit of the candidate model. In general, increasing the number of parameters in a model will improve the goodness-of-fit of the model to the data regardless how many parameters are in the **true model** that generated the data. When a model with too many predicators (significant or not significant ones) is fit to a data set, we may get a perfect fit but the model will be useless for inference such as prediction. In statistics, fitting a model with too many unnecessary parameters is called *over-fitting*. The second part in SBC, $p \times$ log(n), places a penalty term on the complexity of a candidate model, which will increase when the number of parameters in a candidate model increases. Thus the criterion SBC requires a candidate model fitting the data well and penalizing the complexity of the model.

For a group of candidate models, the SBC(p) value will be calculated for each of the models and the best model is the one with the lowest SBC value.

3. Detecting potential outliers.

In regression analysis, outliers are observations with the response variable predicted poorly by a given model. In other words, outliers are observations that do not follow the same model as majority cases of the response variable. In this study, outliers are generally consumers with extremely high or extremely low expenditures. Including outliers in a model increases the sensitivity of the regression model, and sometimes reduces the precision of the model estimation and prediction ability. Hence in practice, outliers commonly need to be detected and removed for the data. Outlier identification techniques will be discussed in this section.

Suppose that after an appropriate transformation $z_i = f(y_i)$, the transformed response variable z_i follows the linear regression model of the form

$$z_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i, \qquad i=1, 2, \ldots, n, \tag{6}$$

where $\{\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n\}$ satisfy the three assumptions given in the last section. Define

$$\mathbf{z} = (z_1, z_2, \cdots, z_n)', \quad \mathbf{x}_j = (x_{j1}, x_{j2}, \cdots, x_{jn})', \quad \varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n), \quad \beta = (\beta_0, \beta_1, \cdots, \beta_p)'$$

and

$$\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p), \qquad \text{where } \mathbf{1} = (1, 1, \cdots, 1)'.$$

Then model (5) can be expressed in the vector-matrix form:

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{7}$$

and the least-squares estimate of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}$. The fitted values based on the model are $\hat{\mathbf{z}} = \mathbf{X}\mathbf{z} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}$ where matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is called the projection matrix or the **hat** matrix. Moreover, the residuals can be expressed as $\hat{\boldsymbol{\varepsilon}} = (\mathbf{I} - \mathbf{H})\mathbf{z}$ (Weisberg, 2005; Chapter 8).

Let h_{ii} denote the i^{th} diagonal element of the matrix $\mathbf{H}$. The variance of residual $\hat{\varepsilon}_i$ is actually $(1 - h_{ii})\sigma^2$. The studentized residuals are defined as (Weisberg, 2005; Chapter 9)

$$r_i = \frac{\hat{\varepsilon}_i}{(1-h_{ii})\hat{\sigma}}, \qquad i=1,\ 2,\ \ldots,\ n, \tag{8}$$

When the sample size n is large, the studentized residual r_i has an approximately normal distribution with mean zero and variance one. In other words, r_i approximately has a standard normal distribution.

One important assumption in linear model analysis is that the model in (8) is appropriate for all cases $\{(z_i, x_{1i}, \cdots, x_{pi}),\ i=1,\ 2, \ldots,\ n\}$ in the given data set. Cases that follow a different model than the rest of the data are called outliers. In our analysis, outliers correspond to APD consumers whose waiver expenditures were not predicted well by the iBudget algorithm. Generally they are consumers with extremely high or extremely low expenditures.

In this study, outliers are defined as these cases with $|r_i| \geq 1.645$, corresponding to extreme values outside a 90% interval of the residual population, each tail with 5% of the theoretical normal population with mean zero and variance one.

III. Independent and Dependent Variable Analysis.

1) Dependent Variable Analysis:

The main dependent variable used in this study is the APD consumers' FY 2013-2014 expenditures with the following adjustments:

1) Removed expenditures for individuals who had fewer than 12 months' of claims in FY 13-14;
2) Removed expenditures for individual who were not actively enrolled on January 1, 2013.

After the first two adjustments, the sample size is n=25,625, i.e., 25,625 consumers' expenditures will be used in this study.

Besides the first two adjustments, the following potential adjustments were considered in this analysis as well:

3) Include or remove expenditures for waiver support coordination (WSC), dental services, environmental adaptations, durable medical equipment, transportation, and geographic rate differentials for all waiver services;
4) Include expenditures for dental ($601 and up), environmental modifications, and durable medical equipment;
5) Include or exclude WSC, transportation, and dental (less than $600).

After examining all the potential adjustments and fitting regression models under different scenarios, we conclude that the selected final model for the algorithm is essentially the same (with almost identical weights for independent variables and R-squared values) with or without adjustments 3)-5). **Therefore in this study, the APD consumers' FY 2013-2014 expenditures with adjustments 1) and 2) will be used as the dependent variable, referred as "FY13-14 claim."**

2) Independent Variable Analysis

The main purposes of developing a statistical algorithm for calculating APD consumers' individual budgets are: 1) increasing the fairness of resource distribution based on consumers' individual characteristics and assessment results; 2) predicting resource needs before services are decided upon and managing funds scientifically; and 3) enhancing transparency of the fund distribution process and sustainability of APD's programs and services. Independent variables used in this study, including living setting, age, and consumers' individual characteristics, are essential for developing the algorithm and for achieving these main purposes.

In this section, we first examine the two main independent variables, living setting and age, and then we move to other independent variables.

a). Living Setting Variable.

As we mentioned in Section I, the majority of consumers served by Florida APD live in the community with family members, their own home, or a congregate living setting such as a group home. The following 22 levels of living settings in Table 1a are initially examined.

Table 1a. Living Setting Levels (22 Levels)

Level	Descriptions
FH	Family Home **(n=12810)**
ILSL	Independent Living & Supported Living **(n=4658)**
NC	Not Classified (GH and 5 RH) **(n=234).** Renamed to "Long-Term Residential Care (LTRC)"
RH1	Residential Habilitation - Basic (month/day) **(n=272)**
RH2	Residential Habilitation - Minimal (month/day) **(n=1705)**
RH3	Residential Habilitation - Moderate (month/day) **(n=2798)**
RH4	Residential Habilitation - Extensive 1 (month/day) **(n=981)**
RH5	Residential Habilitation - Extensive 2 (month/day) **(n=158)**
RHBF1	Residential Habilitation - Behavioral Focus - Minimal (month/day) **(n=62)**
RHBF2	Residential Habilitation - Behavioral Focus - Moderate (month/day) **(n=565)**
RHBF3	Residential Habilitation - Behavioral Focus -Extensive 1 (month/day) **(n=493)**
RHBF4	Residential Habilitation - Behavioral Focus -Extensive 2 (month/day) **(n=222)**
RHCTEP1	Residential Habilitation - Intensive Behavioral - Comprehensive Transitional Education Program - Day Level 3/4 **(n=35)**
RHCTEP2	Residential Habilitation - Intensive Behavioral - Comprehensive Transitional Education Program - Day Level 5/6 **(n=100)**
RHCTEP3	Residential Habilitation - Intensive Behavioral - Trillium - Comprehensive Transitional Education Program - Day Adult **(n=11 consumers)**
RHCTEP4	Residential Habilitation - Intensive Behavioral - Trillium - Comprehensive Transitional Education Program - Day Child **(n=3 consumers)**
RHIB1	Residential Habilitation - Intensive Behavioral - Day Level 1 **(n=4 consumers)** Residential Habilitation - Intensive Behavioral - Day Level 2 **(n=16 consumers**)
RHIB2	Residential Habilitation - Intensive Behavioral - Day Level 3 **(n=71 consumers)**
RHIB3	Residential Habilitation - Intensive Behavioral - Day Level 4 **(n=145 consumers)**
RHIB4	Residential Habilitation - Intensive Behavioral - Day Level 5 **(n=108 consumers)** Residential Habilitation - Intensive Behavioral - Day Level 6 **(n=9 consumers)**
RHLI	Residential Habilitation - Live In day **(n=147 consumers)**
SHC	Special Medical Home Care **(n=18)**
Total Number:	**N=25,625**

Among the 22 categories listed in Table 1a, the NC (not classified) category has 234 consumers, with 5 consumers listed as Residential Habilitation (RH) and the others listed as Group Home (GH). APD found that "The Not Classified category is for individuals in facilities (GH, RH) that do not have Residential Habilitation as a service." A regression model (Regression model 1a on page 12) of FY13-14 Claim before transformation on the 22-level living setting variable is fitted. From the model we can find that the estimated coefficients for the ILSL (Independent Living &

Supported Living) and NC categories are 10504.4 and 9991.6, respectively. We feel that it is reasonable to assign consumers in the NC category to the ILSL category. APD suggests naming this combined category as " Independent-Living/Supported-Living and Long-Term Residential Care", or for short, "ILSL/LTRC." Since the category name "ILSL" has already been used in the legislature rules, we will keep using "ILSL" in this report.

Regression Model 1a: FY13-14 claims (unadjusted) as the dependent variable

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     17964.8      156.4 114.874  < 2e-16 ***
LiveILSL        10504.4      302.8  34.686  < 2e-16 ***
LiveNC           9991.6     1167.6   8.557  < 2e-16 ***
LiveRH1          3478.1     1084.6   3.207  0.00134 ** 
LiveRH2         15495.8      456.3  33.960  < 2e-16 ***
LiveRH3         30730.1      369.4  83.198  < 2e-16 ***
LiveRH4         49579.1      586.4  84.554  < 2e-16 ***
LiveRH5         68341.1     1416.8  48.236  < 2e-16 ***
LiveRHBF1       23152.8     2253.3  10.275  < 2e-16 ***
LiveRHBF2       40701.5      760.9  53.492  < 2e-16 ***
LiveRHBF3       56095.6      812.4  69.052  < 2e-16 ***
LiveRHBF4       73282.7     1198.2  61.161  < 2e-16 ***
LiveRHCTEP1    117307.6     2995.9  39.156  < 2e-16 ***
LiveRHCTEP2    133932.7     1776.9  75.374  < 2e-16 ***
LiveRHCTEP3    182485.8     5339.1  34.179  < 2e-16 ***
LiveRHCTEP4    163687.9    10220.3  16.016  < 2e-16 ***
LiveRHIB1       86430.4     3960.9  21.821  < 2e-16 ***
LiveRHIB2       81478.4     2106.4  38.681  < 2e-16 ***
LiveRHIB3       93561.1     1478.2  63.294  < 2e-16 ***
LiveRHIB4       95937.3     1643.8  58.362  < 2e-16 ***
LiveRHLI        25647.3     1468.2  17.468  < 2e-16 ***
LiveSHC        105236.9     4174.9  25.207  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17700 on 25603 degrees of freedom
Multiple R-squared: 0.5922,     Adjusted R-squared: 0.5919 
F-statistic:  1771 on 21 and 25603 DF,  p-value: < 2.2e-16
```

Comments:

1) The original Living Setting variable has 22 levels as defined on Table 1a;

   The degrees of Freedom of this variable is 21 because Level-FH is used as the base level.

2) SBC = 501,516 for this model.

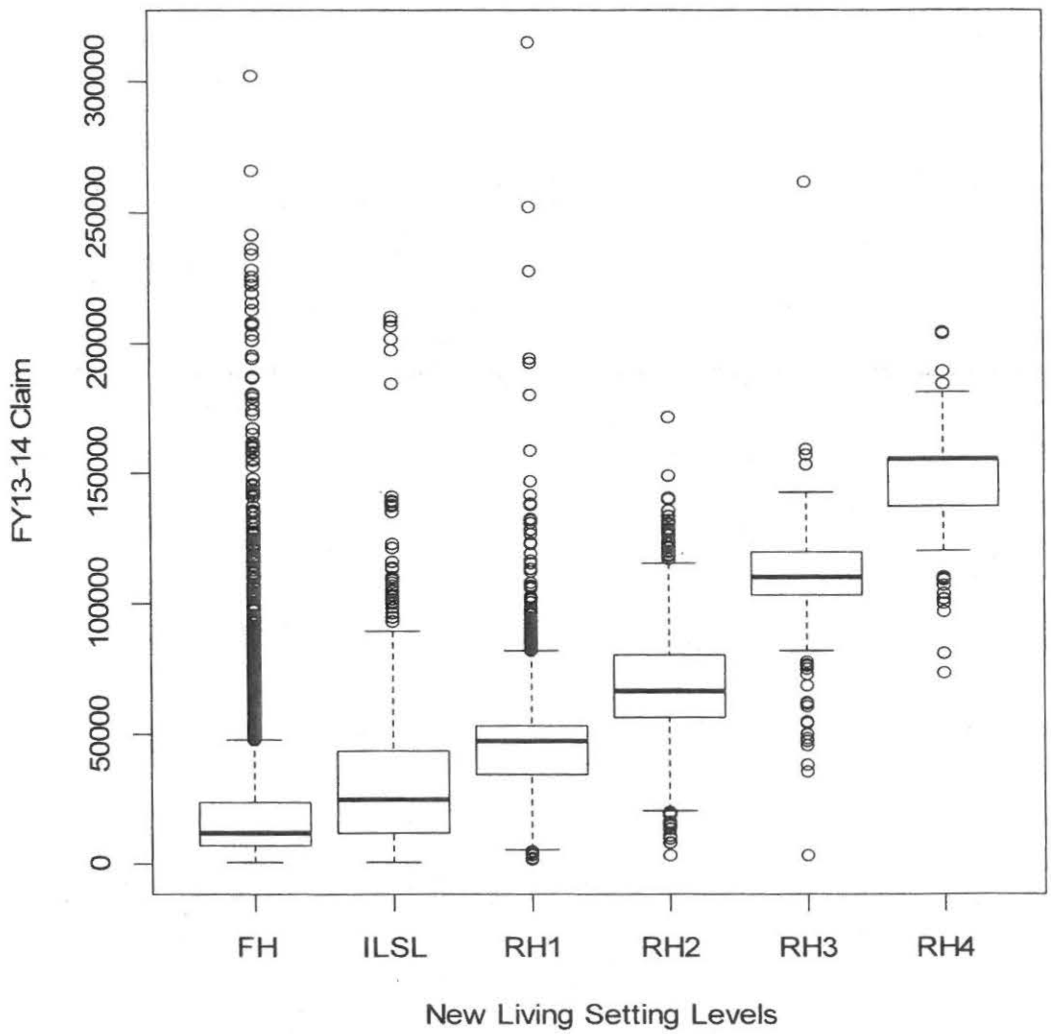
With feedbacks and suggestions from stakeholders and after discussing with APD, the 22 levels are further aggregated to 6 levels listed in the first column of Table 1b. Figure 1 shows the boxplots of the dependent variable FY13-14 Claim versus the 6 levels of the new living setting variable. As expected, the FY13-14 Claim increases along with the level of living setting. In addition, many outliers of expenditures appear in the Family Home (FH) group and in the Residential Habilitation 1 (RH1) group.

A regression model (Regression model 1b on page 15) of FY13-14 Claim before transformation on the new 6-level living setting variable is fitted, with R-squared value of 0.518. In other words, this new living setting variable alone explains about 51.8% of the total variation of the dependent variable (FY13-14 claim).

Table 1b. New Version of Living Setting Type (Decided on 2/24/15 meeting with APD group)

New Level	Level	Descriptions
FH	**FH**	Family Home **(n=12810)**
ILSL	**ILSL**	Independent Living & Supported Living **(n=4658)**
	NC	Not Classified (GH and 5 RH) **(n=234).** Renamed to LTRC
New-RH1 (Residential Habilitation, Standard and Live In)	**RHLI**	Residential Habilitation - Live In day **(n=147 consumers)**
	RH1	Residential Habilitation - Basic (month/day) **(n=272)**
	RH2	Residential Habilitation - Minimal (month/day) **(n=1705)**
	RH3	Residential Habilitation - Moderate (month/day) **(n=2798)**
	RH4	Residential Habilitation - Extensive 1 (month/day) **(n=981)**
	RH5	Residential Habilitation - Extensive 2 (month/day) **(n=158)**
New-RH2 (Residential Habilitation, Behavior Focus)	**RHBF1**	Residential Habilitation - Behavioral Focus - Minimal (month/day) **(n=62)**
	RHBF2	Residential Habilitation - Behavioral Focus - Moderate (month/day) **(n=565)**
	RHBF3	Residential Habilitation - Behavioral Focus -Extensive 1 (month/day) **(n=493)**
	RHBF4	Residential Habilitation - Behavioral Focus -Extensive 2 (month/day) **(n=222)**
New-RH3 (Residential Habilitation, Intensive Behavior)	**RHIB1**	Residential Habilitation - Intensive Behavioral - Day Level 1 **(n=4 consumers)** Residential Habilitation - Intensive Behavioral - Day Level 2 **(n=16 consumers)**
	RHIB2	Residential Habilitation - Intensive Behavioral - Day Level 3 **(n=71 consumers)**
	RHIB3	Residential Habilitation - Intensive Behavioral - Day Level 4 **(n=145 consumers)**
	RHIB4	Residential Habilitation - Intensive Behavioral - Day Level 5 **(n=108 consumers)** Residential Habilitation - Intensive Behavioral - Day Level 6 **(n=9 consumers)**
New-RH4 (Residential Habilitation, CTEP and Special Medical Home Care)	**RHCTEP1**	Residential Habilitation - Intensive Behavioral - Comprehensive Transitional Education Program - Day Level 3/4 **(n=35)**
	RHCTEP2	Residential Habilitation - Intensive Behavioral - Comprehensive Transitional Education Program - Day Level 5/6 **(n=100)**
	RHCTEP3	Residential Habilitation - Intensive Behavioral - Trillium - Comprehensive Transitional Education Program - Day Adult **(n=11 consumers)**
	RHCTEP4	Residential Habilitation - Intensive Behavioral - Trillium - Comprehensive Transitional Education Program - Day Child **(n=3 consumers)**
	SHC	Special Medical Home Care **(n=18)**
	Total Number:	**N=25,625**

Figure 1: FY13-14 Claim vs New Living Setting Levels



Comments:

1) As expected, expenditures increase along with the 6 levels.

2) Many outliers present for consumers living in Family Home (FH) and in Residential Habilitation1 (RH1) levels.

Regression Model 1b: FY13-14 Claim as the dependent variable
Regress Model on the New Version of the Living Setting Variable in Table 1b.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   17964.8       169.9  105.74   <2e-16 ***
 LiveILSL     10479.9       323.2   32.43   <2e-16 ***
 LiveRH1      29129.6       299.8   97.17   <2e-16 ***
 LiveRH2      50935.7       551.7   92.32   <2e-16 ***
 LiveRH3      91514.4      1037.5   88.21   <2e-16 ***
 LiveRH4     131088.1      1497.6   87.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19230 on 25619 degrees of freedom
Multiple R-squared:  0.5184,     Adjusted R-squared:  0.5183
F-statistic:  5516 on 5 and 25619 DF,  p-value: < 2.2e-16
```

Comments:

1) The New Version of the Living Setting variable has 6 levels as defined on Table 1b;
   The degrees of freedom of this variable are 5 since the Level-FH is used as the base level.

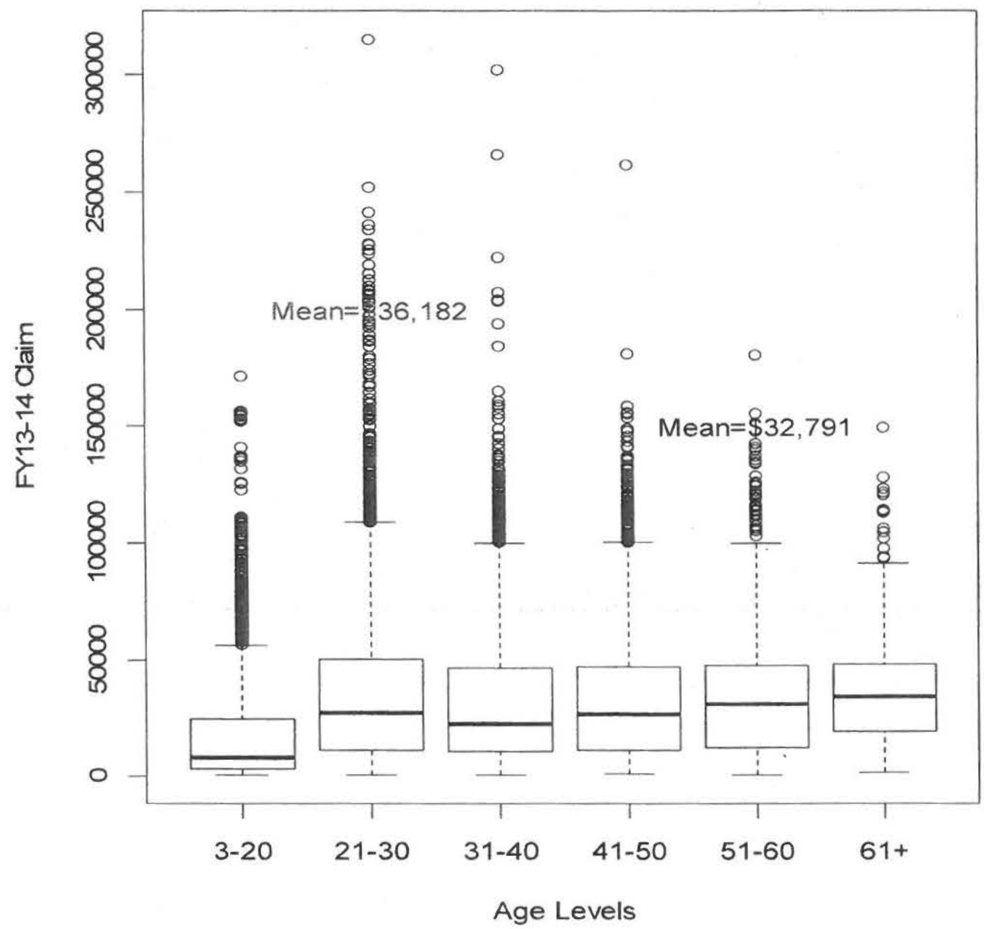
2) SBC = 505,600 for this model.

3) This New version of Living Setting Variable (Table 1b) will be used in this study.

b) Age Effects:

Consumers' age is another important predictor variable for determining annual expenditures. Based on suggestions provided by stakeholders, we first considered six age groups: 3-20 (the youngest consumer is 3 years old), 21-30, 31-40, 41-50, 51-60, and 61+. Figure 2 shows the boxplot of FY13-14 Claim vs the six age groups.

Figure 2: FY13-14 Claim vs Six Age Levels



Comment: Claim mean of the age 21-30 group is higher than Claim means for other groups, mainly due to outliers.

Besides the six levels for the age variable, the following three versions of the age variable are also considered:

1) A two-level dummy variable for Age with **Age2=0** for consumers 3-20; **Age2=1** for consumers 21+.
2) A three-level dummy variable for Age with **Age3=0** for consumers 3-20; **Age3=1** for consumers 21-30, and **Age3=2** for consumers who are 31+.
3) A four-level dummy variable for Age with **Age4=0** for consumers 3-12, **Age4=1** for consumers 13-20, **Age4=2** for consumers 21-30, and **Age4=3** for consumers who are 31+.

Regression models are fitted for these four versions of the age variable. Specifically, regression models 2a-2d present the results for FY13-14 Claim on the two-level, six-level, three-level and four-level age variables, respectively.

Regression Model 2a: FY13-14 claims as the dependent variable
Age21 = 0 if below 21; = 1 if over 21.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19417.9      525.6   36.95   <2e-16 ***
Age21        13668.2      555.8   24.59   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27390 on 25623 degrees of freedom
Multiple R-squared: 0.02305,    Adjusted R-squared: 0.02302
F-statistic: 604.7 on 1 and 25623 DF,  p-value: < 2.2e-16
```

SBC=523,681

Regression Model 2b: FY13-14 claims as the dependent variable,
Age with 6 levels as the independent variable.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19481.0      521.4   37.36   <2e-16 ***
Age21-30     16701.4      626.5   26.66   <2e-16 ***
Age31-40     11129.6      626.3   17.77   <2e-16 ***
Age41-50     12082.9      653.4   18.49   <2e-16 ***
Age51-60     13310.0      686.1   19.40   <2e-16 ***
Age61+       16050.1      806.6   19.90   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27300 on 25619 degrees of freedom
Multiple R-squared: 0.02919,    Adjusted R-squared: 0.029
F-statistic: 154.1 on 5 and 25619 DF,  p-value: < 2.2e-16
```

SBC=523,564

Regression Model 2c: FY13-14 claims as the dependent variable,
Age with 3 levels as the independent variable.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19481.0      521.9   37.33   <2e-16 ***
Age21-30     16701.4      627.1   26.63   <2e-16 ***
Age31+       12472.5      563.1   22.15   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27330 on 25622 degrees of freedom
Multiple R-squared:  0.02717,   Adjusted R-squared:  0.02709
F-statistic: 357.8 on 2 and 25622 DF,  p-value: < 2.2e-16
```

SBC=523,584

Regression Model 2d: FY13-14 claims as the dependent variable,
Age with 4 levels as the independent variable.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    13983       3156   4.431 9.40e-06 ***
Age13-20        5653       3200   1.767 0.0773 .
Age21-30       22200       3175   6.993 2.76e-12 ***
Age31+         17971       3163   5.682 1.34e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27330 on 25621 degrees of freedom
Multiple R-squared:  0.02729,   Adjusted R-squared:  0.02717
F-statistic: 239.6 on 3 and 25621 DF,  p-value: < 2.2e-16
```

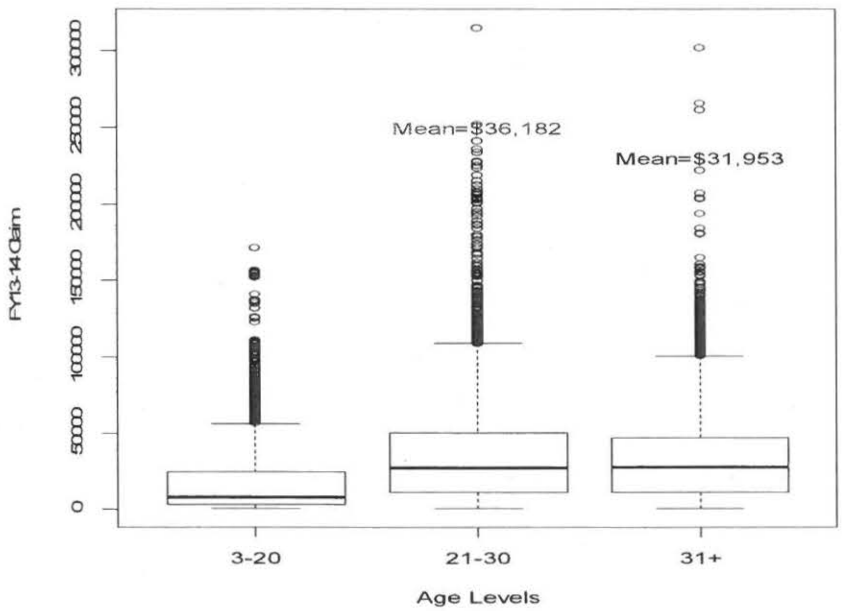
SBC=523,592

Comments:

1) SBC value of Model 2b is lower than the SBC values for Model 2a, Model 2c, and Model 2d;

2) In later model fitting after removing outliers, the estimated weights for the last four groups (31-40, 41-50, 51-60, and 60+) are almost identical. Thus Age with 3 levels will be used as an independent variable in this analysis.

Figure 3: FY13-14 Claim (unadjusted) vs Three Age Levels



c) Disability Type Analysis:

Next, we examined the disability types of the consumers, which consist of the following three variables (details of the disability types are provided on Page 29):

1) Primary Disability: a six-level categorical variable;
2) Secondary Disability: a ten-level categorical variable;
3) Other Disability: a three-level categorical variable.

Regression models of FY13-14 Claim on these three variables, after adjusting for other main predictors such as Living Setting, Age, functional status raw score (FSum), behavioral status raw score (BSum), Q18, Q20, and Q23, are fitted. The results are presented in Models 3b, 3c, and 3d, respectively. Model 3a shows the results of FY13-14 Claim on the main predictors that were used in the 2010 algorithm developed by Niu and Bell (2010). Fifty-three consumers have disability type not classified, which results in the sample size reduced from 26,625 to 26,572.

Regression Model 3a: Square-Root of FY13-14 claims as the dependent variable

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   25.65070    1.27429  20.129   <2e-16 ***
LiveILSL      50.97013    0.82974  61.429   <2e-16 ***
LiveRH1       78.06844    0.74254 105.138   <2e-16 ***
LiveRH2      120.11484    1.38349  86.820   <2e-16 ***
LiveRH3      191.72507    2.47028  77.613   <2e-16 ***
LiveRH4      253.12753    3.46464  73.060   <2e-16 ***
Age21-30      49.42857    1.03089  47.947   <2e-16 ***
Age31+        50.46685    0.98508  51.231   <2e-16 ***
BSum           1.53597    0.05604  27.409   <2e-16 ***
FSum           0.77823    0.08714   8.931   <2e-16 ***
Q18            7.87478    0.40814  19.294   <2e-16 ***
Q20            4.20690    0.44483   9.457   <2e-16 ***
Q23            7.25649    0.39767  18.248   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 44.01 on 25559 degrees of freedom
Multiple R-squared:  0.6416,    Adjusted R-squared:  0.6414
F-statistic:  3812 on 12 and 25559 DF,  p-value: < 2.2e-16
```

SBC=193,684

Regression Model 3b: Square-Root of FY13-14 claims as the dependent variable
Main Predictors and Primary Disability variable

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   26.00751    1.38106  18.832  < 2e-16 ***
LiveILSL      50.91577    0.83095  61.274  < 2e-16 ***
LiveRH1       78.19836    0.74415 105.084  < 2e-16 ***
LiveRH2      120.38133    1.38260  87.069  < 2e-16 ***
LiveRH3      191.13075    2.48324  76.968  < 2e-16 ***
LiveRH4      252.89565    3.45934  73.105  < 2e-16 ***
AgeLevel1     50.45288    1.05828  47.674  < 2e-16 ***
AgeLevel2     51.65373    1.06052  48.706  < 2e-16 ***
BSum           1.47527    0.05663  26.051  < 2e-16 ***
FSum           0.83559    0.08729   9.573  < 2e-16 ***
Q18            8.30123    0.41833  19.844  < 2e-16 ***
Q20            4.02880    0.44483   9.057  < 2e-16 ***
Q23            6.56043    0.40759  16.095  < 2e-16 ***
DisbP2        -2.02583    1.01798  -1.990 0.046596 *
DisbP4         5.66770    1.09822   5.161 2.48e-07 ***
DisbP8       -12.33817    1.95970  -6.296 3.10e-10 ***
DisbP9        20.46185    5.60367   3.652 0.000261 ***
DisbP10      -14.54256    6.04646  -2.405 0.016173 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43.94 on 25554 degrees of freedom
Multiple R-squared:  0.6428,    Adjusted R-squared:  0.6426
F-statistic:  2705 on 17 and 25554 DF,  p-value: < 2.2e-16
```

SBC=193,650

Comments (compared with Model 3a):

1) In this model six levels of the primary disability were considered: levels 1, 2, 4, 8, 9, and 10, where level 1 was used as the reference level. Relative to the reference level, all other levels are significant, where levels 2, 8, and 10 have negative coefficients.
2) The R-Squared value of this model is 0.6428, slightly higher than the value of 0.6416 in Model 3a.
3) The SBC value of 193,650 is lower than the value of 193,684 of Model 3a.

Regression Model 3c: Square-Root of FY13-14 claims as the dependent variable
Main Predictors and Secondary Disability variable

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   26.93045    1.30493  20.637  < 2e-16 ***
LiveILSL      51.10115    0.83024  61.550  < 2e-16 ***
LiveRH1       78.06686    0.74289 105.085  < 2e-16 ***
LiveRH2      120.10586    1.38344  86.817  < 2e-16 ***
LiveRH3      191.16069    2.47842  77.130  < 2e-16 ***
LiveRH4      252.73644    3.46805  72.876  < 2e-16 ***
Age21-30      49.12761    1.03367  47.527  < 2e-16 ***
Age31+        50.16419    0.99279  50.529  < 2e-16 ***
BSum           1.52884    0.05618  27.214  < 2e-16 ***
FSum           0.75754    0.08728   8.680  < 2e-16 ***
Q18            7.89325    0.41277  19.123  < 2e-16 ***
Q20            4.15155    0.44489   9.332  < 2e-16 ***
Q23            7.16474    0.39803  18.001  < 2e-16 ***
DisbS1         2.44288    1.70628   1.432  0.15224
DisbS2         1.13170    1.11485   1.015  0.31006
DisbS3         0.67909    1.22278   0.555  0.57865
DisbS4         5.19403    1.68966   3.074  0.00211 **
DisbS6        -2.51884    0.64462  -3.907 9.35e-05 ***
DisbS7         0.50410    1.88976   0.267  0.78966
DisbS8         9.08134    5.70278   1.592  0.11130
DisbS9        10.00041    5.63615   1.774  0.07602 .
DisbS10        2.64369    8.99501   0.294  0.76883
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43.98 on 25550 degrees of freedom
Multiple R-squared:  0.6421,     Adjusted R-squared:  0.6418
F-statistic:  2183 on 21 and 25550 DF,  p-value: < 2.2e-16
```

SBC=193,742

Comments (compared with Model 3a):

1) In this model the secondary disability category zero was used as the reference level. Relative to the reference level, level 6 has significant effect with negative coefficient; level 4 has

positive significant coefficient; the effects of all other levels (1-3, and 7-10) are not statistically different from the effect of level zero.

2) The R-Squared value of this model is 0.6421, slightly higher than the value of 0.6416 for Model 3a.
3) The SBC value of 193,742 is higher than the value of 193,684 of Model 3a.

Regression Model 3d: Square-Root of FY13-14 claims as the dependent variable Main Predictors and Other Disability variable

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   25.68384    1.27373  20.164  < 2e-16 ***
LiveILSL      50.72481    0.83077  61.058  < 2e-16 ***
LiveRH1       77.89820    0.74298 104.846  < 2e-16 ***
LiveRH2      119.89798    1.38387  86.640  < 2e-16 ***
LiveRH3      191.38272    2.47000  77.483  < 2e-16 ***
LiveRH4      252.57500    3.46467  72.900  < 2e-16 ***
Age20-30      49.31776    1.03096  47.837  < 2e-16 ***
Age31+        50.21233    0.98601  50.925  < 2e-16 ***
BSum           1.50453    0.05635  26.701  < 2e-16 ***
FSum           0.79121    0.08733   9.060  < 2e-16 ***
Q18            7.79595    0.40845  19.086  < 2e-16 ***
Q20            4.21724    0.44482   9.481  < 2e-16 ***
Q23            7.23309    0.39759  18.192  < 2e-16 ***
Disb07         0.96435    1.19349   0.808    0.419
Disb09         7.29999    1.43285   5.095 3.52e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43.99 on 25557 degrees of freedom
Multiple R-squared:  0.6419,    Adjusted R-squared:  0.6417
F-statistic:  3273 on 14 and 25557 DF,  p-value: < 2.2e-16
F-statistic:  2183 on 21 and 25550 DF,  p-value: < 2.2e-16
```

SBC=193,680

Comments (compared with Model 3a):

1) In this model three levels of other disability were considered: level 0, 7, and 9, where level 0 was used as the reference level. Relative to the reference level, only the effect of level 9 is significant with positive coefficient.
2) The R-Squared value of this model is 0.6419, slightly higher than the value of 0.6416 for Model 3a.
3) The SBC value of 193,680 is slightly lower than the value of 193,684 of Model 3a.

Overall Comment: These three disability-type variables are not statistically predictive for the response variable because (1) the estimated coefficients for some categories of the three variables are negative and/or (2) the estimated coefficients are not statistically different from zero. Thus these three variables will not be used in this study as predictor variables, even though Models 3b and 3d give slightly smaller SBC values (and slightly higher R-squared values) than Model 3a.

d) Full Independent Variable Analysis:

During the period of 2007-2009, APD developed the Florida Questionnaire for Situational Information (QSI) for assessing its consumers' individual characteristics and support needs. The QSI (Version 4, Revised on 02/15/2008) consists of the following three main parts:

- Part 1: Functional Status, with 11 elements (Q14-Q24) focusing on a person's needs for assistance during the normal course of a routine day;
- Part 2: Behavioral Status, with 6 elements (Q25-Q30) focusing on major behavioral issues requiring support, assistance or intervention;
- Part 3: Physical Status, with 19 elements (Q32-Q50) focusing on health and physical concerns.

Elements in the three parts (Q14-Q50) are listed in Table 2. Each element listed in Table 2 has five levels (level 0 to level 4), from basic to intensive (detailed description of the levels can be found in the QSI document). Based on the "Report to the Legislature on the Agency's Implementation of the Questionnaire for Situational Information (QSI) Assessment" provided by APD on March 5 of 2009, APD had conducted five studies in 2008-2009 (Susan M. Havercamp, Florida Center for Inclusive Communities, USF, Technical reports submitted to APD: 2008, 2009a, 2009b, 2009c, 2009d) "on the Questionnaire for Situational Information (QSI) to determine its reliability and validity in meeting the needs of APD customers. In general, the studies have found that the QSI is a good assessment tool with reliability and validity. As with any assessment tool, its validity and reliability should be continually reviewed. The agency is committed to ongoing improvement of the QSI so it meets both the needs of the agency and its customers."

Similarly, items and their descriptions for QSI Questions 8 and 9 are listed in Table 3. However, items in QSI Questions 8, 9, 12, and 13 were not validated and the reliability of these items was not examined. Therefore items in QSI Questions 8, 9, 12, and 13 will not be used in this study.

Table 2. Elements in the three parts of QSI (Version 4)

Part 1. Functional Support Status		Part 2. Behavioral Support Status		Part 3. Physical Support Status	
Item Number	**Item Description**	**Item Number**	**Item Description**	**Item Number**	**Item Description**
Q14	Vision	Q25	Hurtful to Self/Self Injurious Behavior	Q32	Injury to the Person caused by Self-Injurious Behavior
Q15	Hearing	Q26	Aggressive/Hurtful to Others	Q33	Injury to the Person Caused by Aggression toward Others or Property
Q16	Eating	Q27	Destructive to Property	Q34	Use of Mechanical Restraints or Protective Equipment for Maladaptive Behavior
Q17	Ambulation	Q28	Inappropriate Sexual Behavior	Q35	Use of Emergency Chemical Restraint
Q18	Transfers	Q29	Running Away	Q36	Use of Psychotropic Medications
Q19	Toileting	Q30	Other Behaviors that May Result In Separation from Others	Q37	Gastrointestinal Conditions (includes vomiting, reflux, heartburn, or ulcer)
Q20	Hygiene			Q38	Seizures
Q21	Dressing			Q39	Anti-Epileptic Medication use
Q22	Communications			Q40	Skin Breakdown
Q23	Self-Protection			Q41	Bowel Function
Q24	Ability to Evacuate (place of residence)			Q42	Nutrition
				Q43	Treatment (physician prescribed)
				Q44	Assistance in meeting Chronic Healthcare Needs
				Q45	Individual's Injuries
				Q46	Falls
				Q47	Physician Visits/Nursing Services
				Q48	Emergency Room Visits
				Q49	Hospital Admission
				Q50	Days missed- illness

Table 3. QSI Questions 8 and 9

Item Number	Item Description
Q8a1	LifeChangeOverPast12Months
Q8a2	DeathLossPrimCaregiverSeenDaily
Q8a3	DeathLossSignificantOtherSpouseFriend
Q8a4	ChildRemovedForMaltreatment
Q8a5	DeathLossCloseFamilyMemberNonCustodial
Q8a6	SurvivorMajorPhysicalAssaultNearDeathExperience
Q8a7	DetentionJailInstitutionMore3Days
Q8a8	MajorIllnessInjurySurgeryHospitalMore3Days
Q8a9	PregnancyChildBirth
Q8a10	NewFamilyMemberRoomMate
Q8a11	MajorChangeLivingConditionLifestyle
Q8a12	ChangeInPlaceOfResidence
Q8a13	MajorChangeTypeAmtRecreationalActivity
Q8a14	MajorChangePositiveSocialInteraction
Q8a15	MajorChangeADT_Work
Q8a16	MajorChangeSleeepingHabits
Q8a 17	MajorChangeEatingHabits
Q8b1	NoneApply
Q8b2	SadnessCryingSpells
Q8b3	AvoidanceActivitiesFriends
Q8b4	FeelOverwhelmedDisorientedLost
Q8b5	MajorWeightGainLoss
Q8b6	AccidentInjuryUnknownOrigin
Q8b7	SuicidalThoughtsPlansAttempts
Q8b8	PropertyDestruction
Q8b9	NervousnessAnxietyWorryDesperation
Q8b10	DeclineWorkAttendancePerformance
Q8b11	AgitationIrritabilityRestlessness
Q8b12	SelfInjuriousBehaviors
Q8b13	ReturnIncreaseSeizures
Q8b14	AggressiveBehaviorOthers
Q8b15	UseAlcoholOrIllegalDrugs

Table 3 (Continued). QSI Questions 8 and 9

Q8c1	NoneApply
Q8c2	AdjustmentDisorder
Q8c3	Depression
Q8c4	AnxietyDisorder
Q8c5	SuicideHomicideRisk
Q8c6	PostTraumaticStressDisorder
Q8c7	RiskVictimizationOrRevictimization
Q8d1	NoneApply
Q8d2	LossChangePrimaryCaregiver
Q8d3	LossChangeSignificantOther
Q8d4	MajorSurgeryHospitilization
Q8d5	BirthLossChild
Q8d6	DetentionJailOrInstitution
Q8d7	NewFamilyMemberOrRoomMate
Q8d8	MajorChangeLivingConditionLifestyle
Q8d9	ChangePlaceOfResidence
Q8d10	MajorChangeRecreationalActivities
Q8d11	MajorChangeSocialActivities
Q8d12	MajorChangeWorkDayTimeActivities
Q8d13	OtherMajorChange
Q9	PlanMoveNext12Months

In total 125 independent variables are considered in the model building and analysis. These independent variables are:

1) **Independent Variables 1-36 (Q14-Q24, Q25-Q30, Q32-Q50):** The 36 elements in the QSI survey, including 11 elements (Q14-Q24) for the functional status support part, 6 elements (Q25-Q30) for the behavioral status support part, and 19 elements (Q32-Q50) for the physical status support part. Each score has 5 levels ranging from 0 to 4.

2) **Independent Variables 37-39 (BSum, FSum, PSum):** Sums of raw scores for the three sections in the QSI, named functional status raw score (FSum), behavioral status raw score (BSum), and physical status raw score (PSum). Specifically, the functional status raw score (FSum) is the sum of scores of the 11 elements (Q14-Q24) for the functional status support part, ranging from 0 to 44; behavioral status raw score (BSum) is the sum of scores of the 6 elements (Q25-Q30) for the

behavioral status support part, ranging from 0 to 24; and physical status raw score (PSum) is the sum of scores of the 19 elements (Q32-Q50) for the physical status support part, ranging from 0 to 76.

3) Independent Variable 40 (Live): Living setting with 6 levels defined in Table 1b.

4) Independent Variable 41 (Age) with the following 4 versions:

- **Age-2**: A two-level dummy variable for Age with **Age21=0** for consumers 3-20; **Age21=1** for consumers 21+;
- **Age-3:** A three-level categorical variable for Age with **Age3-20, Age21-30, and Age31+;**
- **Age-4:** A four-level categorical variable for Age with **Age-4=0** for consumers 3-12, **Age-4=1** for consumers 13-20, **Age-4=2** for consumers 21-30, and **Age-4=3** for consumers who are 31+.
- **Age-6:** A six-level categorical variable for Age with **Age-6=0** for consumers 3-20, **Age-6=1** for consumers 21-30, **Age-6=2** for consumers 31-40, **Age-6=3** for consumers 41-50, **Age-6=4** for consumers 51-60, and **Age-6=5** for consumers 61+.

5) Independent Variable 42-50 (Interaction terms between Living Setting and BSum, FSum, and PSum):

- FH = 1 for Family Home; 0 for others;
- SL = 1 for Independent Living & Supported Living and Long-Term Residential Care; 0 for others;
- RH = 1 for New RH1-RH4; 0 for others;

Thus nine interaction terms are considered: FH-BSum, FH-FSum, FH-PSum; SL-BSum, SL-FSum, SL-PSum; and RH-BSum, RH-FSum, RH-PSum.

Besides the above 50 independent variables, the following variables were also considered during the analysis. But these variables were dropped from the algorithm development since they were not validated by professionals in the field (Susan M. Havercamp, Florida Center for Inclusive Communities, USF, Technical reports submitted to APD: 2008, 2009a, 2009b, 2009c, 2009d).

6) Independent Variables 51-100: {Dummy variables (0 or 1) in QSI Questions 8 and 9}, except Q8b1, Q8c1, and Q8d1 (None Apply).

7) Independent Variables 101-113 (Q12a to Q12o): Community Inclusion and Fulfillment of Valued Adult Roles for Persons 18 years and older. Each variable with scores 1-4.

Level 1: You do not need any personal support;
Level 2: You need personal support and it is limited to occasional reminders or verbal prompts and/or physical assistance;
Level 3: You need personal support and require daily reminders, verbal and/or physical prompts;
Level 4: You need personal support from someone and require supervision to complete.

Note: In QSI version 4, there was Level 5 that is now identical to Level 4. Therefore both Level 5 and Level 4 are recoded as Level 4.

Score 0: Don't know (treated as missing);

Score 9: Not applicable (treated as Level 0).

8) Independent Variable 114 (CBC): Dummy variable indicating whether the individual is a child involved in the Community Based Care system, 1 if Yes, 0 if No.

9) Independent Variable 115 (Safety): Community Safety indicator. The default value is zero. Value is set to 1 if there is a record of the consumer ever having been in any one of the following program components (i.e., residential living settings).

- 71 = Adult Mentally Retarded Defendant Program
- 72 = Juvenile Mentally Retarded Defendant Program
- 95 = Jail pre-sentencing (all jail and prison situations prior to May 2007)
- 98 = Jail post-sentencing
- 99 = Prison

10) Independent Variable 116 (MenH) (Note: APD decided in December of 2009 not to use this variable in the analysis since it is not reliable): Dummy variable indicating participation in Florida Medicaid Pre-Paid Mental Health Plan, 1 if Yes, 0 if No.

11) **Independent Variable 117 (DMYN):** Dummy variable indicating participation in Florida Medicaid Chronic Disease Management Program, 1 if Yes, 0 if No.

12) **Independent Variable 118-120 (Disability Type):** Besides "No Disability", the following ten types of primary and secondary disability are defined for consumers

- 0='No Disability'.
- 1='Intellectual Disabilities'.
- 2='Cerebral Palsy'.
- 3='Epilepsy'.
- 4='Autism'.
- 5='High Risk'.
- 6='DD PL Eligible'.
- 7='Other'.
- 8='Spina Bifida'.
- 9='Prader Willi'.
- 10='Down syndrome'.

Primary Disability Type in the file: Levels: 1 2 4 8 9 10;

Secondary Disability Type in the file: Levels: 0 1 2 3 4 6 7 8 9 10;

Besides primary and secondary disability, the following type of disability is considered too:

- 0='No Other Disability'.
- 6='Public Law Eligibility'.
- 7='Chronic Health'.
- 9='Mental Health'.

Other Disability Type in the file: Levels: 0 7 9.

13) **Independent Variable 121 (Nursing):** Dummy variable indicating whether nursing is needed, 1 if Yes, 0 if No.

14) **Independent Variable 122 (CDC):** Dummy variable indicating whether the individual is involved in the Consumer Directed Care system, 1 if Yes, 0 if No.

15) **Independent Variable 123-125 (QSI 13):** Employment Information Q13a, Q13b and Q13c.

IV. Results for the FY 13-14 Claim and Final Model Selected for the Algorithm

1) Transformation of the dependent variable: FY13-14 Claim (minimum-Claim=$734, maximum-Claim=$315,081).

Before the algorithm development and model selection, we first examine whether a transformation in the Box-Cox power transformation family is needed for the dependent variable. The method discussed in Section II is used to choose the transformation power, λ. A regression model with main independent variables Living Setting, Age-3, Functional Status raw score, Behavioral Status raw score, Physical Status raw score, Q18, Q20, and Q23 is fitted.

Regression Model 4a (Before Transformation): FY13-14 Claim as the dependent variable with the main independent variables:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -12144.41     483.58 -25.114  < 2e-16 ***
LiveILSL     15546.06     313.42  49.602  < 2e-16 ***
LiveRH1      24938.75     279.58  89.200  < 2e-16 ***
LiveRH2      45223.51     520.82  86.832  < 2e-16 ***
LiveRH3      86415.81     930.67  92.854  < 2e-16 ***
LiveRH4     127174.96    1304.58  97.483  < 2e-16 ***
Age21-30     14594.01     387.87  37.626  < 2e-16 ***
Age31+       13667.14     370.90  36.849  < 2e-16 ***
PSum           134.40      15.11   8.892  < 2e-16 ***
BSum           455.93      22.90  19.910  < 2e-16 ***
FSum           352.99      33.11  10.662  < 2e-16 ***
Q18           3035.34     154.91  19.594  < 2e-16 ***
Q20            567.68     167.41   3.391 0.000698 ***
Q23           2140.35     149.61  14.307  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16570 on 25611 degrees of freedom
Multiple R-squared:  0.6424,    Adjusted R-squared:  0.6422
F-statistic:  3539 on 13 and 25611 DF,  p-value: < 2.2e-16
```

Figure 4a presents the histogram of the FY 13-14 claim, from which we can see that the distribution of the FY 13-14 Claim is highly skewed. Transformation of the response variable needs to be selected to make the FY 13-14 Claim closer to the bell-shaped normal distribution.

Figure 4a: Histogram of the Response Variable FY 13-14 Claim.

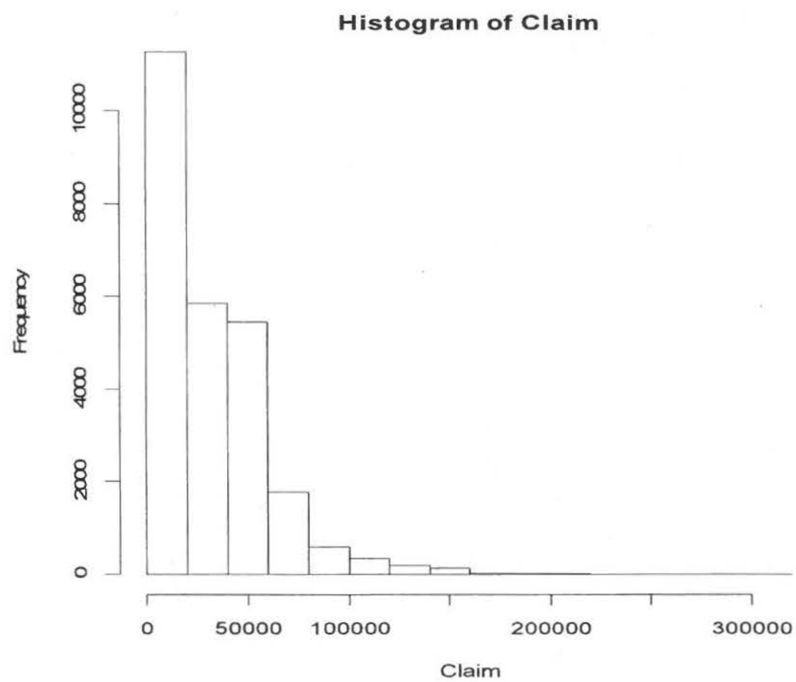
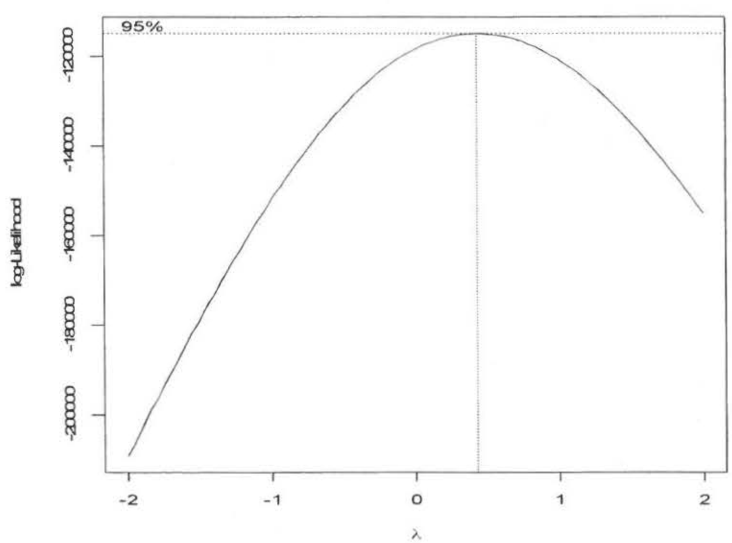


Figure 4b: Transformation for the Response Variable FY 13-14 Claim. The Estimated Power is 0.4. Square-root Transformation Will Be Used.



Regression Model 4b (After Transformation): Square-Root of FY13-14 Claim as the dependent variable with the main independent variables:

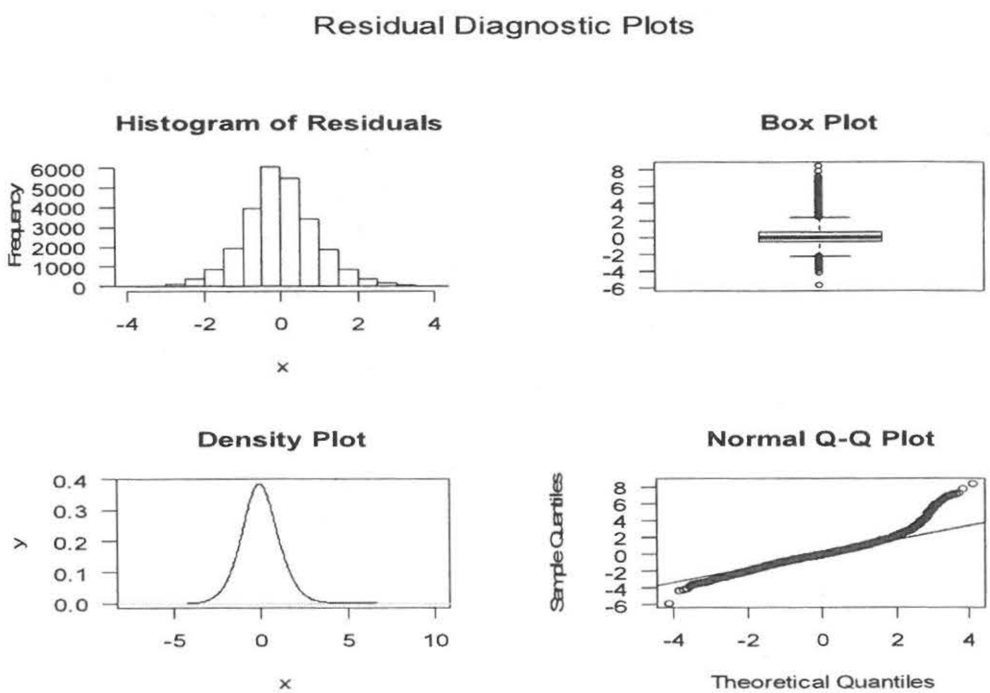
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	24.85794	1.28434	19.355	< 2e-16	***
LiveILSL	50.66126	0.83241	60.861	< 2e-16	***
LiveRH1	77.94992	0.74256	104.975	< 2e-16	***
LiveRH2	120.00970	1.38325	86.759	< 2e-16	***
LiveRH3	191.37613	2.47179	77.424	< 2e-16	***
LiveRH4	253.02480	3.46488	73.026	< 2e-16	***
Age21-30	49.40990	1.03016	47.963	< 2e-16	***
Age31+	50.20032	0.98508	50.961	< 2e-16	***
PSum	0.19136	0.04014	4.767	1.88e-06	***
BSum	1.41661	0.06082	23.293	< 2e-16	***
FSum	0.72272	0.08793	8.220	< 2e-16	***
Q18	7.61319	0.41144	18.504	< 2e-16	***
Q20	4.22085	0.44464	9.493	< 2e-16	***
Q23	7.34014	0.39734	18.473	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 44.01 on 25611 degrees of freedom
Multiple R-squared: 0.6416, Adjusted R-squared: 0.6414
F-statistic: 3526 on 13 and 25611 DF, p-value: < 2.2e-16

Figure 4c: Diagnostic Plots of the Studentized Residuals from Model 4b.



Comments: Histogram (top left) and density plot (bottom left) of the studentized residuals from Regression model 4b are symmetric. But the Q-Q normal plot (bottom right) of the residuals show that the distribution is away from the normal distribution, especially the upper tail part caused by outliers.

Another popular transformation of a skewed distribution is the natural logarithm transformation. The following model shows the results after the logarithm transformation.

Regression Model 4c (Natural Logarithm Transformation): Natural Logarithm of FY13-14 Claim as the dependent variable with the main independent variables:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.8815264  0.0193796 406.692  < 2e-16 ***
LiveILSL    0.7380523  0.0125603  58.761  < 2e-16 ***
LiveRH1     1.0673492  0.0112045  95.261  < 2e-16 ***
LiveRH2     1.4252992  0.0208721  68.287  < 2e-16 ***
LiveRH3     1.9377003  0.0372971  51.953  < 2e-16 ***
LiveRH4     2.3835162  0.0522819  45.590  < 2e-16 ***
Age21-30    0.8660944  0.0155443  55.718  < 2e-16 ***
Age31+      0.9304408  0.0148640  62.597  < 2e-16 ***
PSum        0.0002427  0.0006057   0.401 0.688683
BSum        0.0201844  0.0009177  21.995  < 2e-16 ***
FSum        0.0046411  0.0013267   3.498 0.000469 ***
Q18         0.0845306  0.0062083  13.616  < 2e-16 ***
Q20         0.0896016  0.0067092  13.355  < 2e-16 ***
Q23         0.1184292  0.0059956  19.753  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6641 on 25611 degrees of freedom
Multiple R-squared:  0.5672,    Adjusted R-squared:  0.567
F-statistic:  2582 on 13 and 25611 DF,  p-value: < 2.2e-16
```

Comments:

1) The R-squred value of Model 4c is 0.5672, much lower than the value of 0.6426 for Model 4b. This indicates that the square-root transformation is better than the natural-log transformation.

2) In this analysis, the square-root transformation of the dependent variable will be used.

2) Model Selection. Square-Root of FY13-14 claims as the dependent variable (minimum-Claim=$734, and maximum-Claim=$315,081).

After transforming the response variable by square-root, we now consider regression models with the first 50 independent variables listed on Pages 26-27. The stepwise method, an automatic model selection procedure (e.g., Weisberg 2005), and the Bayesian Information Criterion (SBC) are combined to select the best model for the iBudget algorithm. An initially selected model is given in Model 5a1.

Regression Model 5a1: Square-root of FY13-14 claims as the dependent variable
Model Selected by the stepwise method.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   28.94807    1.36364  21.229  < 2e-16 ***
LiveILSL      39.37091    1.24275  31.681  < 2e-16 ***
LiveRH1       89.67802    1.32062  67.906  < 2e-16 ***
LiveRH2      130.38304    1.75824  74.156  < 2e-16 ***
LiveRH3      202.79715    2.65173  76.477  < 2e-16 ***
LiveRH4      258.53844    3.69903  69.894  < 2e-16 ***
Age21-30      49.47629    1.01895  48.556  < 2e-16 ***
Age31+        51.18360    0.98463  51.983  < 2e-16 ***
BSum           0.77438    0.08499   9.111  < 2e-16 ***
FSum          -0.93722    0.14947  -6.270 3.66e-10 ***
FHFSum         0.73115    0.06843  10.684  < 2e-16 ***
SLFSum         1.71123    0.09993  17.125  < 2e-16 ***
SLBSum         0.85839    0.13717   6.258 3.96e-10 ***
PSum          -0.13319    0.04568  -2.916 0.003549 **
Q16            4.17493    0.39637  10.533  < 2e-16 ***
Q18            7.68356    0.42227  18.196  < 2e-16 ***
Q20            3.53683    0.46698   7.574 3.75e-14 ***
Q21            3.65838    0.46892   7.802 6.34e-15 ***
Q23            7.01105    0.40924  17.132  < 2e-16 ***
Q24            1.58948    0.40468   3.928 8.60e-05 ***
Q28            2.20847    0.32821   6.729 1.75e-11 ***
Q33            1.88509    0.49918   3.776 0.000159 ***
Q34            2.71950    0.60500   4.495 6.99e-06 ***
Q36            2.23616    0.30023   7.448 9.77e-14 ***
Q43            3.99532    0.30250  13.208  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43.36 on 25600 degrees of freedom
Multiple R-squared:  0.6523,     Adjusted R-squared:  0.652
F-statistic:  2001 on 24 and 25600 DF,  p-value: < 2.2e-16
```

Comment:

1) Both the estimated coefficients of FSum and PSum are negative, which implies that the higher of the two scores, the less fund consumers will receive. This does not make sense because we do not want consumers with worse conditions, either functionally or physically, to receive less support. Due to this reason, FSum and PSum will be removed from the analysis in the next step.

Regression Model 5a2: Square-root of FY13-14 claims as the dependent variable
FSum and PSum are removed from Model 5a1.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   28.85174    1.35565  21.283  < 2e-16 ***
LiveILSL      38.79130    1.24121  31.253  < 2e-16 ***
LiveRH1       87.44011    1.27928  68.351  < 2e-16 ***
LiveRH2      128.58365    1.73862  73.957  < 2e-16 ***
LiveRH3      200.80786    2.63942  76.080  < 2e-16 ***
LiveRH4      257.63000    3.68848  69.847  < 2e-16 ***
Age21-30      49.42278    1.01991  48.458  < 2e-16 ***
Age31+        50.81861    0.98341  51.676  < 2e-16 ***
BSum           0.74929    0.08427   8.891  < 2e-16 ***
FHFSum         0.58671    0.06501   9.024  < 2e-16 ***
SLFSum         1.58662    0.09823  16.152  < 2e-16 ***
SLBSum         0.85251    0.13730   6.209 5.40e-10 ***
Q16            2.82431    0.34303   8.233  < 2e-16 ***
Q18            6.04498    0.35207  17.170  < 2e-16 ***
Q20            2.38989    0.43413   5.505 3.73e-08 ***
Q21            2.43698    0.42869   5.685 1.32e-08 ***
Q23            6.04823    0.38118  15.867  < 2e-16 ***
Q24            0.21270    0.33927   0.627 0.530710
Q28            2.28844    0.32814   6.974 3.15e-12 ***
Q33            1.63101    0.49219   3.314 0.000922 ***
Q34            2.31113    0.59747   3.868 0.000110 ***
Q36            2.10835    0.28813   7.317 2.60e-13 ***
Q43            3.56558    0.28235  12.628  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43.4 on 25602 degrees of freedom
Multiple R-squared:  0.6516,     Adjusted R-squared:  0.6513
F-statistic:  2177 on 22 and 25602 DF,  p-value: < 2.2e-16
```

SBC = 193,472

Comment:

After removing FSum and PSum from Model 5a1, Q24 becomes not significant (p-value = 0.53 > 0.05). Predictor Q24 will be removed from Model 5a2 in the next step.

Regression Model 5a3: Square-root of FY13-14 claims as the dependent variable
Final model Selected before removing outliers.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   28.86442    1.35548  21.295  < 2e-16 ***
LiveILSL      38.73034    1.23738  31.300  < 2e-16 ***
LiveRH1       87.45002    1.27916  68.365  < 2e-16 ***
LiveRH2      128.59510    1.73850  73.969  < 2e-16 ***
LiveRH3      200.77660    2.63892  76.083  < 2e-16 ***
LiveRH4      257.71202    3.68611  69.914  < 2e-16 ***
Age21-30      49.39938    1.01922  48.468  < 2e-16 ***
Age31+        50.79827    0.98286  51.684  < 2e-16 ***
BSum           0.75146    0.08420   8.925  < 2e-16 ***
FHFSum         0.59081    0.06468   9.134  < 2e-16 ***
SLFSum         1.59380    0.09756  16.336  < 2e-16 ***
SLBSum         0.84837    0.13714   6.186 6.25e-10 ***
Q16            2.84565    0.34133   8.337  < 2e-16 ***
Q18            6.06602    0.35046  17.309  < 2e-16 ***
Q20            2.44140    0.42628   5.727 1.03e-08 ***
Q21            2.50630    0.41418   6.051 1.46e-09 ***
Q23            6.09430    0.37402  16.294  < 2e-16 ***
Q28            2.28147    0.32794   6.957 3.56e-12 ***
Q33            1.63661    0.49210   3.326 0.000883 ***
Q34            2.31274    0.59745   3.871 0.000109 ***
Q36            2.09799    0.28765   7.294 3.11e-13 ***
Q43            3.56285    0.28231  12.620  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43.4 on 25603 degrees of freedom
Multiple R-squared:  0.6516,    Adjusted R-squared:  0.6513
F-statistic:  2280 on 21 and 25603 DF,  p-value: < 2.2e-16
```

SBC = 193,461.

Comment:

The SBC value of Model 5a3 is 193, 461, which is smaller than the value of 193,472 of Model 5a2.

Conclusion: Model 5a3 is selected as the final model before removing outliers.

Regression Model 5b: Square-root of FY13-14 claims as the dependent variable
Final model Selected after removing 9.40% outliers (2,410 consumers).

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   27.5720    1.02385  26.930  < 2e-16 ***
LiveILSL      35.8220    0.91949  38.958  < 2e-16 ***
LiveRH1       90.6294    0.94365  96.041  < 2e-16 ***
LiveRH2      131.7576    1.28906 102.212  < 2e-16 ***
LiveRH3      209.4558    1.93208 108.409  < 2e-16 ***
LiveRH4      267.0995    2.71191  98.491  < 2e-16 ***
Age21-30      47.8473    0.79766  59.985  < 2e-16 ***
Age31+        48.9634    0.76383  64.103  < 2e-16 ***
BSum           0.4954    0.06304   7.859 4.05e-15 ***
FHFSum         0.6349    0.04891  12.980  < 2e-16 ***
SLFSum         2.0529    0.07452  27.550  < 2e-16 ***
SLBSum         1.4501    0.10411  13.929  < 2e-16 ***
Q16            2.4984    0.25754   9.701  < 2e-16 ***
Q18            5.8537    0.26477  22.109  < 2e-16 ***
Q20            2.6772    0.31360   8.537  < 2e-16 ***
Q21            2.7878    0.30608   9.108  < 2e-16 ***
Q23            6.3555    0.27706  22.939  < 2e-16 ***
Q28            2.2803    0.24347   9.366  < 2e-16 ***
Q33            1.2233    0.36703   3.333  0.00086 ***
Q34            2.1764    0.46642   4.666 3.09e-06 ***
Q36            2.6734    0.21641  12.353  < 2e-16 ***
Q43            1.9304    0.22799   8.467  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.82 on 23193 degrees of freedom
Multiple R-squared:  0.7998,	Adjusted R-squared:  0.7996
F-statistic:  4412 on 21 and 23193 DF,  p-value: < 2.2e-16
```

SBC = 159,394.3

Comments:

1) Model 5b is the same model as Model 7b in some task-2 earlier reports that were presented in stakeholder meetings. In earlier reports, many different models had been examined, such as (1) models involving different versions of the living setting variable, (2) models with and without nursing variable, and (3) models with and without QSI questions 8-13. This results in many more models. In this final report, the model selection process was streamlined and fewer models were presented.
2) As we discuss later, Model 5b will be recommended to APD as the final model for the new Florida iBudget Algorithm.

Regression Model 5b1 (2010 Algorithm but with the new living setting variable and the new age groups). Square-root of FY13-14 claims as the dependent variable after removing 9.40% outliers (2,410 consumers).

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   21.88395    0.98213   22.28   <2e-16 ***
LiveILSL      53.43289    0.63024   84.78   <2e-16 ***
LiveRH1       80.26754    0.55607  144.35   <2e-16 ***
LiveRH2      121.30582    1.04120  116.51   <2e-16 ***
LiveRH3      198.18801    1.83471  108.02   <2e-16 ***
LiveRH4      259.63198    2.58056  100.61   <2e-16 ***
Age21-30      48.82486    0.82037   59.52   <2e-16 ***
Age31+        50.21821    0.77899   64.47   <2e-16 ***
BSum           1.43767    0.04279   33.60   <2e-16 ***
FSum           0.88170    0.06644   13.27   <2e-16 ***
Q18            6.91175    0.31581   21.89   <2e-16 ***
Q20            4.29588    0.33551   12.80   <2e-16 ***
Q23            7.72666    0.29954   25.80   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.81 on 23202 degrees of freedom
Multiple R-squared:  0.7867,     Adjusted R-squared:  0.7866
F-statistic:  7130 on 12 and 23202 DF,  p-value: < 2.2e-1616
```

SBC= 160,769.3

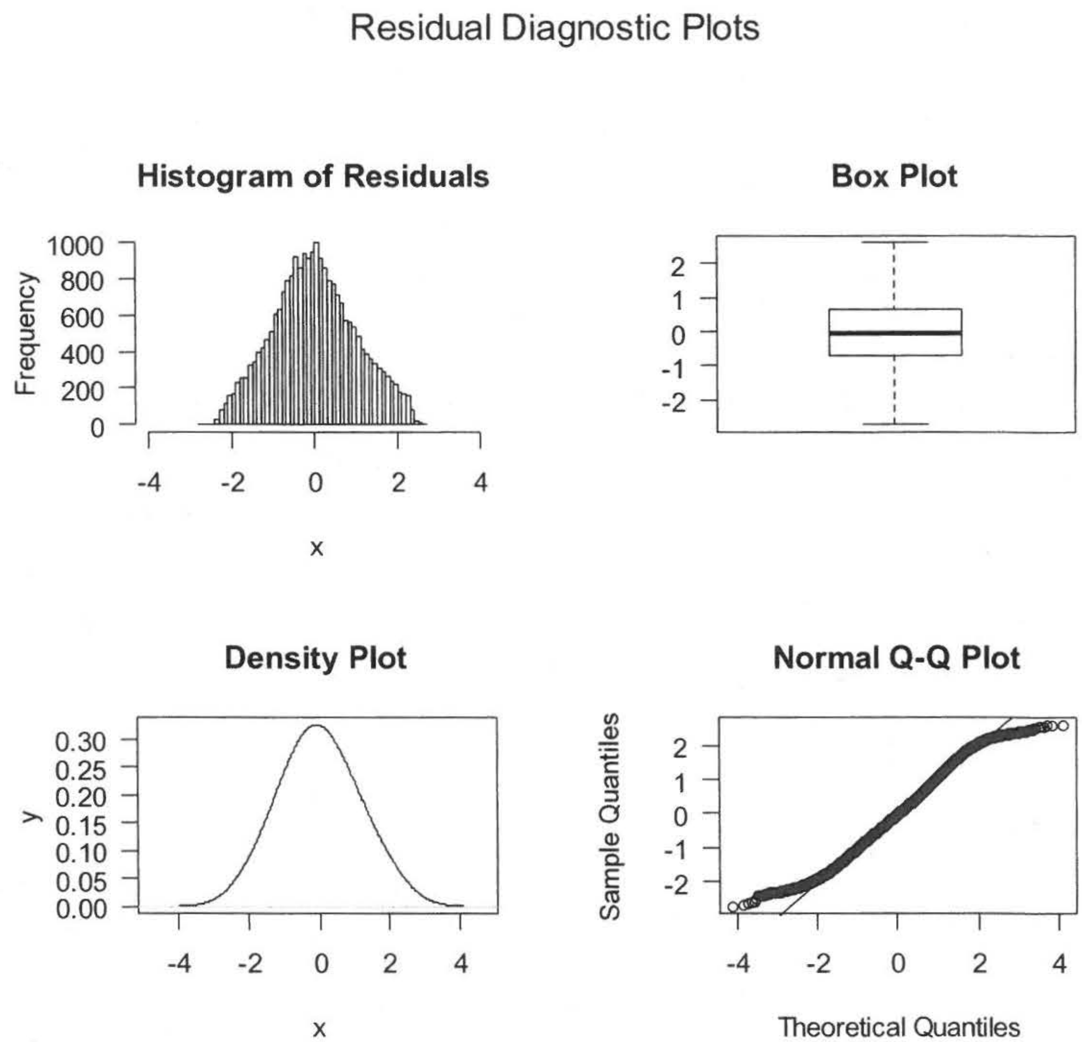
Comments:

1) The R-squared value of Model 5b1 is 0.7867, smaller than the value of 0.7998 of Model 5b.

2) SBC value of 160,769 for Model 5b1 (2010 algorithm but with the new living setting variable and the new age groups) is much larger than the SBC value of 159,394 for Model 5b.

Conclusion: Model 5b is a better choice for the final model than Model 5b1, the model for the 2010 algorithm but with the new living setting variable and the new age groups

Figure 5: Diagnostic Plots of the Studentized Residuals from Model 5b.



Comment: Histogram (top left), Boxplot (top right), and density plot (bottom left) of the studentized residuals from Regression model 4b are symmetric. But the Q-Q normal plot (bottom right) of the residuals shows that the distribution is still away from the normal distribution in the two tails.

Regression Model 5c: Square-root of FY13-14 claims as the dependent variable
Model Selected after removing 4.96% outliers (1,270 consumers).

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    28.38318    1.11888  25.368  < 2e-16 ***
LiveILSL       37.07476    1.00511  36.886  < 2e-16 ***
LiveRH1        89.51319    1.04052  86.027  < 2e-16 ***
LiveRH2       131.27332    1.41953  92.476  < 2e-16 ***
LiveRH3       207.63918    2.14877  96.631  < 2e-16 ***
LiveRH4       265.75236    2.99812  88.640  < 2e-16 ***
Age21-30       47.91222    0.85611  55.965  < 2e-16 ***
Age30+         49.84830    0.82144  60.684  < 2e-16 ***
BSum            0.55595    0.06896   8.062 7.83e-16 ***
FHFSum          0.56743    0.05348  10.610  < 2e-16 ***
SLFSum          1.94349    0.08153  23.837  < 2e-16 ***
SLBSum          1.20150    0.11353  10.583  < 2e-16 ***
Q16             2.34810    0.28106   8.355  < 2e-16 ***
Q18             5.81242    0.28850  20.147  < 2e-16 ***
Q20             2.83477    0.34570   8.200 2.52e-16 ***
Q21             2.66914    0.33674   7.926 2.35e-15 ***
Q23             6.12122    0.30433  20.114  < 2e-16 ***
Q28             2.29379    0.26799   8.559  < 2e-16 ***
Q33             1.19875    0.40269   2.977 0.002915 **
Q34             1.90605    0.50538   3.772 0.000163 ***
Q36             2.71833    0.23640  11.499  < 2e-16 ***
Q43             1.51392    0.24440   6.194 5.95e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.61 on 24333 degrees of freedom
Multiple R-squared:  0.7549,    Adjusted R-squared:  0.7546
F-statistic:  3568 on 21 and 24333 DF,  p-value: < 2.2e-16
```

Comment: If only about 5% outliers are removed, the R-squared value of Model 5c reduced to 0.7549, which is lower than the value of 0.7998 of Model 5b.

Conclusion: In this study, Model 5b is selected as the final model for the new algorithm and will be recommended to APD.

V. Weights for the Final Algorithm and Examples.

When Model 5b is used as the final model, the weights for calculating consumers' individual budget are actually the estimated coefficients of the independent variables in Model 5b and listed in Table 4. In addition, an example iBudget is calculated using the given weights. For example, in Model 5b family_home (FM) of the living setting variable is treated as the reference level with weight zero, and the weight of ILSL (Independent Living & Supported Living) is 35.822 in the square-root scale. Similarly for the age variable, the group of age below 21 is treated as the reference level with weight zero, while the weights of Age 21-30 and Age older than 31 are assigned as 47.8473 and 48.9634 in the square-root scale, respectively.

Table 4. Weights used to Calculate Consumers' Needs (Square-Root Scale)

Variable Name	Weights	Example Level	Example Value
Intercept	27.5720	1	27.5720
LiveILSL	35.8220	1	35.8220
LiveRH1	90.6294	0	0
LiveRH2	131.7576	0	0
LiveRH3	209.4558	0	0
LivePH4	267.0995	0	0
Age21-30	47.8473	0	0
Age31+	48.9634	1	48.9634
BSum	0.4954	0	0
FHFSum	0.6349	0	0
SLFSum	2.0529	23	47.2170
SLBSum	1.4501	0	0
Q16	2.4984	3	7.4953
Q18	5.8537	1	5.8537
Q20	2.6772	3	8.0315
Q21	2.7878	3	8.3633
Q23	6.3555	3	19.0666
Q28	2.2803	0	0
Q33	1.2233	0	0
Q34	2.1764	0	0
Q36	2.6734	0	0
Q43	1.9304	0	0
Total in the square Root Scale			208.3847
Predicted Support			$43424.18

Table 5 contains example iBudget calculation for 50 randomly chosen consumers. A full list of predicted expenditures for over 23,000 consumers will be submitted to APD in an Excel file. Note that due to the square-root transformation on the response, predicted values in the model must be squared back to attain the iBudget amount as seen below. Predicted support would then be adjusted further based on appropriations and set-asides for exceptional needs, changed needs, and one-time needs.

Table 5. iBudget Calculation for Fifty Randomly Selected Individuals Based on the Chosen Model 5b.

Example	Prediction	Claim	Live	AgeLevel	BSum	FHFSum	SLFSum	SLBSum
1	34493	22946	RH1	2	1	0	0	0
2	45787	49716	RH1	2	0	0	0	0
3	15968	23913	FH	1	2	18	0	0
4	66972	59594	RH2	1	8	0	0	0
5	12090	11198	FH	2	1	9	0	0
6	31299	45989	RH1	0	18	0	0	0
7	15204	11186	FH	0	11	28	0	0
8	37819	34560	RH1	2	2	0	0	0
9	16887	11439	ILSL	2	0	0	2	0
10	6372	8094	FH	0	5	18	0	0
11	11204	18269	FH	1	6	9	0	0
12	20056	19177	FH	2	0	24	0	0
13	26417	23041	FH	1	0	28	0	0
14	12628	2901	FH	2	0	15	0	0
15	9582	7861	FH	2	0	5	0	0
16	12120	13819	FH	2	0	10	0	0
17	38675	29714	RH1	2	0	0	0	0
18	22757	30768	FH	2	6	26	0	0
19	45432	75471	ILSL	2	3	0	21	3
20	13353	13196	FH	1	3	11	0	0
21	35266	35725	RH1	1	0	0	0	0
22	48416	77771	RH1	2	23	0	0	0
23	57834	90045	RH1	2	3	0	0	0
24	31111	43828	ILSL	2	10	0	6	10
25	8804	12960	FH	2	0	3	0	0
26	33087	31336	ILSL	2	6	0	9	6
27	18399	24692	FH	1	11	22	0	0
28	67590	63879	RH2	2	15	0	0	0
29	17706	10292	ILSL	2	2	0	2	2
30	41618	32346	RH1	2	9	0	0	0
31	13764	23665	FH	2	8	9	0	0
32	11842	21653	FH	2	0	8	0	0
33	54484	51107	RH1	2	0	0	0	0
34	15356	19666	FH	2	5	15	0	0
35	15194	4941	FH	1	15	6	0	0
36	13288	11841	FH	2	0	6	0	0
37	11255	9753	FH	0	23	15	0	0
38	48986	51211	RH1	2	0	0	0	0
39	26588	22364	ILSL	2	3	0	13	3
40	9342	1884	FH	1	0	10	0	0
41	6274	7723	FH	2	0	0	0	0
42	24619	11528	ILSL	1	0	0	8	0
43	9811	2265	FH	1	0	10	0	0
44	28320	43775	ILSL	2	4	0	7	4
45	55365	50973	RH1	2	11	0	0	0
46	20691	24597	FH	1	11	23	0	0
47	34985	44613	FH	1	2	35	0	0
48	8542	8171	FH	2	0	5	0	0
49	28984	52064	FH	1	0	32	0	0
50	10999	6773	FH	2	0	6	0	0

VI. Fractions of Variation in the Response explained by Predictors

In this section, regression models with different groups of predictors are fitted and fractions of total variation in the response variable, the square-root of the FY 13-14 claims, are calculated.

Regression Model 6a: Square-root of FY13-14 claims as the dependent variable after removing 9.40% outliers (2,410 consumers).

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 116.0392     0.3956  293.34   <2e-16 ***
 LiveILSL    41.5641     0.7484   55.54   <2e-16 ***
 LiveRH1     97.5407     0.6774  143.99   <2e-16 ***
 LiveRH2    143.1878     1.2486  114.68   <2e-16 ***
 LiveRH3    216.7217     2.3211   93.37   <2e-16 ***
 LiveRH4    271.8728     3.3637   80.83   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.99 on 23209 degrees of freedom
Multiple R-squared:  0.6282,    Adjusted R-squared:  0.6282
F-statistic:  7844 on 5 and 23209 DF,  p-value: < 2.2e-16
```

Regression Model 6b: Square-root of FY13-14 claims as the dependent variable after removing 9.40% outliers (2,410 consumers).

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  112.686      1.423   79.19   <2e-16 ***
Age21-30      57.373      1.693   33.88   <2e-16 ***
Age31+        52.218      1.520   34.34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 67 on 23212 degrees of freedom
Multiple R-squared:  0.05316,    Adjusted R-squared:  0.05308
F-statistic: 651.6 on 2 and 23212 DF,  p-value: < 2.2e-16
```

Regression Model 6c: Square-root of FY13-14 claims as the dependent variable after removing 9.40% outliers (2,410 consumers).

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 144.29359    0.63563  227.01   <2e-16 ***
BSum          4.95920    0.06510   76.18   <2e-16 ***
FHFSum       -1.60826    0.04041  -39.80   <2e-16 ***
SLFSum        2.68106    0.10206   26.27   <2e-16 ***
SLBSum       -3.58077    0.17495  -20.47   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 58.64 on 23210 degrees of freedom
Multiple R-squared:  0.2748,    Adjusted R-squared:  0.2747
F-statistic:  2199 on 4 and 23210 DF,  p-value: < 2.2e-16
```

Regression Model 6d: Square-root of FY13-14 claims as the dependent variable after removing 9.40% outliers (2,410 consumers).

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  95.2881     1.0289  92.613  < 2e-16 ***
Q16           2.0804     0.4651   4.473 7.75e-06 ***
Q18           9.4475     0.4590  20.581  < 2e-16 ***
Q20           6.1446     0.5575  11.022  < 2e-16 ***
Q21          -1.0762     0.5563  -1.934   0.0531 .
Q23           8.8348     0.4501  19.629  < 2e-16 ***
Q28          10.2707     0.3909  26.275  < 2e-16 ***
Q33           9.9610     0.6189  16.095  < 2e-16 ***
Q34           7.0303     0.8199   8.575  < 2e-16 ***
Q36          14.2799     0.3368  42.396  < 2e-16 ***
Q43           3.3620     0.4195   8.014 1.17e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 56.86 on 23204 degrees of freedom
Multiple R-squared:  0.3183,    Adjusted R-squared:  0.318
F-statistic:  1084 on 10 and 23204 DF,  p-value: < 2.2e-16
```

Comments on the Coefficient of Determination of Models 6a – 6d:

1) From Model 6a, the Living Setting variable alone explains about 62.8% of the total variation of the response variable FY13-14 Claim in the square-root scale.

2) From Model 6b, the Age variable alone explains about 5.3% of the total variation of the response variable FY13-14 Claim in the square-root scale.

3) From Model 6c, BSum and the three interaction variables (FHFSum, SLFSum, and SLBSum) together explain about 27.5% of the total variation of the response variable FY13-14 Claim in the square-root scale .

4) From Model 6d, the ten QSI individual questions (Q16, Q18, Q20, Q21, Q23, Q28, Q33, Q34, Q36, and Q43) together explain about 31.8% of the total variation of the response variable FY13-14 Claim in the square-root scale.

5) It should be pointed out that the predictor variables in the final selected model are dependent with each other. The fractions of total variation explained by different groups of independent variables are not additive.

VII. Robustness Test of the Final Model by the Bootstrapping Method

In this section, we use the bootstrapping method to assess the robustness of the selected final model. Bootstrapping is a statistical method for estimating the sampling distribution of an estimator by resampling with replacement from the original sample (Efron, 1979; Efron and Tibshirani, 1993). This method is widely used in parameter estimation, regression, time series, and other statistical problems.

In regression analysis, there are two basic approaches to perform bootstrapping. The first approach is bootstrapping cases. Assume that we want to fit a regression model with response variable Y and predictors $\{ x_1, \cdots, x_p \}$, and we have a sample of n cases $\{Z_i = (Y_i, x_{1i}, \cdots, x_{pi}), i = 1, \cdots, n\}$. In bootstrapping, we select n cases randomly from the original observations with replacement. For each bootstrapping sample, we fit the regression model and save the coefficients. Averages of the estimated coefficients and confidence intervals for the regression coefficients can be constructed based on M (a given large number) bootstrapping samples. The second approach in bootstrapping regression analysis is resampling the residuals. In this way a regression model is fitted to the original data first. The fitted values and residuals are calculated based on the fitted model. The residuals are randomly resampled and added back to the fitted values. After generating a new set of bootstrapping data, the regression model is refitted.

Efron and Tibshirani (1993, Chapter 9, p. 113) pointed out that bootstrapping cases was less sensitive to model assumptions such as normality and constant variance than bootstrapping residuals. In this study, the bootstrapping cases method was used to find the averages of the estimated coefficients in the final model.

There are several approaches to constructing bootstrap confidence intervals for the estimated coefficients in a regression model. The normal-theory interval assumes that an estimated coefficient $\hat{\theta}$ based on the original data is normally distributed and uses the bootstrap estimate of sampling variance of the estimated coefficient to construct a 100(1 − α)-percent confidence interval of the form:

$$\hat{\theta}^* \pm z_{1-\alpha/2} SE(\hat{\theta}^*) \quad (9)$$

where $\hat{\theta}^*$ is the bootstrapping average estimate of the coefficient with $\hat{\theta}^* = \frac{1}{M}\sum_{i=1}^{M}\hat{\theta}_i^*$, $SE(\hat{\theta}^*)$ is the bootstrap estimate of the standard error of $\hat{\theta}$, and $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard-normal distribution (e.g., 1.96 for a 95-percent confidence interval, where $\alpha = .05$).

An alternative approach, called the bootstrap percentile interval, is to use the empirical quantiles of $\hat{\theta}$ to form a confidence interval $[\hat{\theta}^*_{Lower}, \hat{\theta}^*_{Upper}]$ for the coefficient θ, where $\{\hat{\theta}_{(1)}^*, \hat{\theta}_{(2)}^*, \cdots, \hat{\theta}_{(M)}^*\}$ are the ordered bootstrap estimates of the coefficient; $\hat{\theta}^*_{Lower}$ and $\hat{\theta}^*_{Upper}$ are the $\alpha/2$ percentile and the $(1-\alpha/2)$ percentile of the bootstrap estimates of the coefficient, respectively.

Table 6. Weights in the Final Model, Average Weights based on 10,000 Bootstrapping Samples, and the 95% Lower Limit and Upper Limit for the Average Weights.

Coefficient	Weights in Model 5b	Bootstrapping Average Weights	95% Bootstrapping Lower Limit	Bootstrapping Median Weights	95% Bootstrapping Upper Limit
(Intercept)	27.57	27.55	25.49	27.55	29.60
LiveILSL	35.82	35.84	33.94	35.85	37.68
LiveRH1	90.63	90.63	88.92	90.63	92.34
LiveRH2	131.76	131.76	129.27	131.77	134.26
LiveRH3	209.46	209.44	206.26	209.46	212.59
LiveRH4	267.10	267.15	261.88	267.10	272.44
Age21-30	47.85	47.85	46.16	47.86	49.56
Age31+	48.96	48.97	47.31	48.97	50.62
BSum	0.50	0.49	0.37	0.49	0.62
FHFSum	0.63	0.63	0.54	0.63	0.73
SLFSum	2.05	2.05	1.90	2.05	2.20
SLBSum	1.45	1.45	1.22	1.45	1.67
Q16	2.50	2.50	1.99	2.50	3.02
Q18	5.85	5.85	5.32	5.85	6.39
Q20	2.68	2.67	2.07	2.67	3.29
Q21	2.79	2.79	2.21	2.79	3.38
Q23	6.36	6.36	5.79	6.36	6.92
Q28	2.28	2.28	1.80	2.28	2.76
Q33	1.22	1.23	0.50	1.23	1.97
Q34	2.18	2.17	1.12	2.18	3.23
Q36	2.67	2.67	2.23	2.68	3.11
Q43	1.93	1.93	1.45	1.93	2.41

Comment: The Bootstrapping average weights are almost identical to the weights in Model 5b, indicating that the selected final model and weights of different predictors are very robust when consumers' information changes in the future.

VIII. Conclusions and Recommended Final Model for the iBudget Algorithm Based on the FY 13-14 Claim

The main purpose of Task two is to develop a new model for the Florida iBudget Algorithm using the baseline data from July 1, 2013 to June 30, 2014. Dependent variables and independent variables in this study are carefully examined in Section III. Before selecting the final model between the response variable and the predictors, a Square-Root transformation of the FY 13-14 Claim was chosen based on the Box-Cox procedure.

Multiple linear regression models are used to build the relationship between the dependent variable and all the 125 independent variables. This type of model is in line with typical practices used in many areas, including health and human services. For the FY 13-14 Claim data, 26,625 consumers' information is used in the analysis. We first examined two main independent variables, living setting and age. Then we conducted a full analysis of the independent variables, in which 125 variables were involved and many of them, including items in QSI questions 8 to 13, were eliminated from the study since these variables were neither validated by APD nor by other research agencies.

The final model was selected by the Bayesian Information Criterion, which explained about 80% of the total variation in the response variable. The final model contains the following independent variables:

Table 7. Independent Variables used in the Final Model and Weights

Variable Name	Weights	Variable Description
Intercept	27.5720	Intercept of the model
LiveILSL	35.8220	Independent Living and Supported Living
LiveRH1	90.6294	Residential Habilitation, Standard and Live In
LiveRH2	131.7576	Residential Habilitation, Behavior Focus
LiveRH3	209.4558	Residential Habilitation, Intensive Behavior
LiveRH4	267.0995	Residential Habilitation, CTEP and Special Medical Home Care
Age21-30	47.8473	Consumer age between 21 and 30
Age31+	48.9634	Consumer age 31 and older
BSum	0.4954	Behavioral status sum score
FHFSum	0.6349	Interaction term between Family Home and Functional status sum score
SLFSum	2.0529	Interaction term between ILSL and Functional status sum score
SLBSum	1.4501	Interaction term between ILSL and Behavioral status sum score
Q16	2.4984	Eating
Q18	5.8537	Transfers
Q20	2.6772	Hygiene
Q21	2.7878	Dressing
Q23	6.3555	Self-protection
Q28	2.2803	Inappropriate Sexual Behavior
Q33	1.2233	Injury to the Person Caused by Aggression toward Others or Properties
Q34	2.1764	Use of Mechanical Restraints or Protective Equipment for Maladaptive Behavior
Q36	2.6734	Use of Psychotropic Medications
Q43	1.9304	Treatment (Physician Prescribed)

After the selection of the final model for the Florida iBudget Algorithm, the bootstrapping method was used to assess the robustness of the final model. The results of this assessment showed that average weights of the independent variables based on 10,000 bootstrapping samples are almost identical to these estimated in the final model, indicating the selected model is very stable against changes of consumers' information in the future.

References:

DeBeaugrine, J. Director of APD, 2009. "Report to the Legislature on the Agency's Implementation of the Questionnaire for Situational Information (QSI) Assessment."

Efron, B. 1979. "Bootstrap Methods: Another Look at the Jackknife." *Annals of Statistics* 7:1-26.

Efron, B. and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Havercamp, Susan M., 2008. "Assessing the Face, Content and Construct Validity of the Florida Questionnaire for Situational Information: Observations of Two Expert Panels." Technical Report Submitted to the Florida Agency for Persons with Disabilities.

Havercamp, Susan M., 2009a. "Evaluating the Reliability and Validity of the Questionnaire for Situational Information: Test-Retest Reliability." Technical Report Submitted to the Florida Agency for Persons with Disabilities.

Havercamp, Susan M., 2009b. "Evaluating the Reliability and Validity of the Questionnaire for Situational Information: Inter - Interviewer Reliability." Technical Report Submitted to the Florida Agency for Persons with Disabilities.

Havercamp, Susan M., 2009c. "Evaluating the Reliability and Validity of the Questionnaire for Situational Information: Factor Analysis." Technical Report Submitted to the Florida Agency for Persons with Disabilities.

Havercamp, Susan M., 2009d. "Evaluating the Reliability and Validity of the Questionnaire for Situational Information: Item Analyses." Technical Report Submitted to the Florida Agency for Persons with Disabilities.

Niu, X-F., and Bell, L. 2010. "Statistical Models for Predicting Resource Needs and Establishing Individual Budgets for Individuals served by the Florida Agency for Persons with Disabilities. " Technical Report Submitted to the Florida Agency for

Persons with Disabilities.

Niu, X-F., and Tao, M. Nov. 2014 to Jan. 2015 "Evaluate and Refine the Florida APD's iBudget Algorithm. " Task-One Technical Report Submitted to the Florida Agency for Persons with Disabilities.

Schwartz, G. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics*, 6: 461-464

Weisberg, S. 2005. *Applied Linear Regression.* John Wiley & Sons Inc. New York.