Household Pulse data on kids mental health

Introduction

In this analysis, we aim to identify the predictors of children's need for mental health services. We will utilize a linear modeling approach for continuous outcomes (OLS) and a logistic regression approach for binary outcomes (GLM with a binomial family) to ascertain the significant predictors and their interactions.

To predict whether children are assessed as needing mental health services using the provided dataset, we would focus on variables that are likely to be associated with mental health needs. Based on the variable definitions provided and general understanding from psychological and sociological research, here are some variables that could be important influences and moderators:

TBIRTH_YEAR / ABIRTH_YEAR: Age is a primary factor in mental health needs. Different age groups tend to have different mental health challenges.

RHISPANIC / AHISPANIC / RRACE / ARACE: Ethnicity and race can influence the prevalence and recognition of mental health issues due to cultural, socio-economic, and genetic factors.

EEDUC / AEDUC: The education level of household members can affect awareness and attitudes towards mental health and access to services.

MS: Marital status of parents can impact the mental health of children, with factors such as single parenting or divorce potentially playing a role.

THHLD_NUMPER / THHLD_NUMKID / AHHLD_NUMPER / AHHLD_NUMKID: The number of people and children in the household can influence the level of attention each child receives, which can affect mental health.

KIDS_LT5Y / KIDS_5_11Y / KIDS_12_17Y: These variables give an age breakdown that can be important, as mental health needs vary significantly with age.

CURFOODSUF: Food sufficiency status can impact mental health, with food insecurity potentially causing stress and anxiety.

CHILDFOOD: This variable specifically addresses whether children are not eating enough due to financial constraints, which could directly relate to stress and mental health.

ANXIOUS / WORRY / INTEREST / DOWN: Previous responses to questions about anxiety, worry, interest, and feeling down can indicate mental health status.

HLTHINS1 - HLTHINS8: Health insurance status affects access to mental health services.

MHLTH_NEED / MHLTH_GET / MHLTH_SATISFD / MHLTH_DIFFCLT: Direct indicators of whether mental health services are needed, have been received, the satisfaction with the services, and any difficulties obtaining them.

SOCIAL1 / SOCIAL2: Measures of social support and feelings of loneliness, which can be significant factors in mental health.

SUPPORT1 - SUPPORT4 / SUPPORT1EXP: These social support indicators can moderate the effects of various stressors on mental health.

INCOME: Household income level, which can affect access to resources, including mental health services.

EXPNS_DIF: Difficulty with expenses can be a source of stress that affects mental health.

Analysis

Data Cleaning

```
datta<-read.csv("hps_04_00_02_puf.csv")
# Load necessary libraries
library(dplyr)
library(tidyr)
library(ggplot2)
library(broom)
library(caret)

# Step 1: Subsetting
selected_vars <- c("TBIRTH_YEAR", "RHISPANIC", "RRACE", "EEDUC", "MS",
                   "THHLD_NUMPER", "THHLD_NUMKID", "KIDS_LT5Y", "KIDS_5_11Y", "KIDS_12_17Y",
                   "CURFOODSUF", "CHILDFOOD", "ANXIOUS", "WORRY", "INTEREST", "DOWN",
                   "HLTHINS1", "HLTHINS2", "HLTHINS3", "HLTHINS4", "HLTHINS5", "HLTHINS6",
                   "HLTHINS7", "HLTHINS8", "MHLTH_NEED", "MHLTH_GET", "MHLTH_SATISFD",
                   "MHLTH_DIFFCLT", "SOCIAL1", "SOCIAL2", "SUPPORT1", "SUPPORT2",
                   "SUPPORT3", "SUPPORT4", "INCOME", "EXPNS_DIF")

data <- datta %>% select(all_of(selected_vars))


# Step 2: Handle Missing Values
cleaned_data <- na.omit(data)
```

Split the data into training and testing sets

```
# Split the data into training and testing sets
set.seed(123) # for reproducibility
index <- createDataPartition(cleaned_data$MHLTH_NEED, p = .7, list = FALSE)
train_data <- cleaned_data[index, ]
test_data <- cleaned_data[-index, ]
dim(train_data)
```

```
## [1] 49808    36
```

```
dim(test_data)
```

```
## [1] 21344    36
```

OLS model

```
# OLS model
ols_model <- lm(MHLTH_NEED ~ ., data = train_data)
summary(ols_model)
```

```
##
## Call:
## lm(formula = MHLTH_NEED ~ ., data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -157.638   -9.766   -5.461    3.795   94.880
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.829e+02  1.341e+01 -28.561  < 2e-16 ***
## TBIRTH_YEAR    1.980e-01  6.751e-03  29.333  < 2e-16 ***
## RHISPANIC     -1.023e+00  3.388e-01  -3.019  0.00254 **
## RRACE          6.597e-02  1.211e-01   0.545  0.58579
## EEDUC          9.685e-01  6.583e-02  14.712  < 2e-16 ***
## MS            -1.453e-01  1.163e-02 -12.492  < 2e-16 ***
## THHLD_NUMPER   4.458e-01  1.050e-01   4.246 2.18e-05 ***
## THHLD_NUMKID   1.571e+01  2.244e-01  69.998  < 2e-16 ***
## KIDS_LT5Y      2.149e-01  4.693e-03  45.784  < 2e-16 ***
## KIDS_5_11Y     1.472e-01  4.405e-03  33.420  < 2e-16 ***
## KIDS_12_17Y    2.560e-01  4.495e-03  56.954  < 2e-16 ***
## CURFOODSUF    -5.999e-02  1.320e-02  -4.544 5.52e-06 ***
## CHILDFOOD      2.829e-01  4.113e-03  68.789  < 2e-16 ***
## ANXIOUS        1.421e-01  2.043e-02   6.954 3.59e-12 ***
## WORRY          8.139e-03  2.152e-02   0.378  0.70532
## INTEREST       7.574e-02  2.457e-02   3.082  0.00205 **
## DOWN           7.677e-03  2.123e-02   0.362  0.71763
## HLTHINS1       6.992e-03  5.290e-03   1.322  0.18622
## HLTHINS2      -4.704e-03  5.773e-03  -0.815  0.41521
## HLTHINS3      -3.174e-02  5.903e-03  -5.377 7.62e-08 ***
## HLTHINS4      -1.317e-03  8.231e-03  -0.160  0.87285
## HLTHINS5      -9.815e-04  1.092e-02  -0.090  0.92839
## HLTHINS6       5.477e-03  1.173e-02   0.467  0.64063
## HLTHINS7       8.028e-03  1.166e-02   0.689  0.49102
## HLTHINS8       2.020e-03  5.611e-03   0.360  0.71880
## MHLTH_GET     -6.515e-02  5.731e-02  -1.137  0.25564
## MHLTH_SATISFD  1.083e-01  1.408e-02   7.690 1.50e-14 ***
## MHLTH_DIFFCLT  1.071e-01  5.662e-02   1.892  0.05852 .
## SOCIAL1       -9.133e-03  1.681e-02  -0.543  0.58688
## SOCIAL2        1.053e-01  1.763e-02   5.972 2.35e-09 ***
## SUPPORT1       7.710e-02  1.980e-02   3.894 9.86e-05 ***
## SUPPORT2      -3.687e-02  2.111e-02  -1.747  0.08069 .
## SUPPORT3       1.529e-02  1.703e-02   0.898  0.36938
## SUPPORT4      -4.694e-03  9.241e-03  -0.508  0.61147
## INCOME         1.460e-02  5.241e-03   2.786  0.00534 **
## EXPNS_DIF     -2.971e-02  1.440e-02  -2.063  0.03911 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.45 on 49772 degrees of freedom
## Multiple R-squared:  0.7192, Adjusted R-squared:  0.719
## F-statistic:  3642 on 35 and 49772 DF,  p-value: < 2.2e-16
```

GLM model

```
dat<-cleaned_data
dat$MHLTH_NEED<-ifelse(dat$MHLTH_NEED%in%1:2,0,ifelse(dat$MHLTH_NEED==3,1,NA))
dat <- dat[!is.na(dat$MHLTH_NEED), ]


set.seed(123) # for reproducibility
index <- createDataPartition(dat$MHLTH_NEED, p = .7, list = FALSE)
train_data <- dat[index, ]
test_data <- dat[-index, ]

# GLM model with logistic regression
glm_model <- glm(MHLTH_NEED ~ ., data = train_data, family = binomial)
summary(glm_model)
```

```
##
## Call:
## glm(formula = MHLTH_NEED ~ ., family = binomial, data = train_data)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -6.811e+01  4.171e+01  -1.633  0.10249
## TBIRTH_YEAR    2.320e-02  1.730e-02   1.340  0.18010
## RHISPANIC     -8.290e-01  5.663e-01  -1.464  0.14322
## RRACE         -1.592e-01  2.297e-01  -0.693  0.48821
## EEDUC         -1.405e-02  1.454e-01  -0.097  0.92300
## MS            -1.615e-01  6.178e-02  -2.614  0.00894 **
## THHLD_NUMPER  -2.662e-01  1.862e-01  -1.430  0.15269
## THHLD_NUMKID   2.906e-01  3.218e-01   0.903  0.36644
## KIDS_LT5Y     -2.725e-03  6.006e-03  -0.454  0.65005
## KIDS_5_11Y     2.060e-03  5.123e-03   0.402  0.68755
## KIDS_12_17Y   -7.366e-03  5.291e-03  -1.392  0.16382
## CURFOODSUF    -6.643e-03  8.621e-02  -0.077  0.93858
## CHILDFOOD      1.844e-02  2.128e-02   0.866  0.38626
## ANXIOUS       -1.097e-01  1.826e-01  -0.601  0.54789
## WORRY         -2.964e-02  1.091e-01  -0.272  0.78590
## INTEREST       4.917e-03  6.205e-02   0.079  0.93684
## DOWN          -5.073e-02  1.503e-01  -0.338  0.73565
## HLTHINS1      -4.679e-03  3.696e-02  -0.127  0.89925
## HLTHINS2      -4.626e-03  6.900e-02  -0.067  0.94654
## HLTHINS3       1.725e-02  8.227e-02   0.210  0.83396
## HLTHINS4       3.493e-03  1.010e-01   0.035  0.97241
## HLTHINS5      -2.651e-03  1.320e-01  -0.020  0.98397
## HLTHINS6       3.916e-03  1.330e-01   0.029  0.97651
## HLTHINS7      -5.833e-03  1.495e-01  -0.039  0.96887
## HLTHINS8      -2.055e-02  6.176e-02  -0.333  0.73938
## MHLTH_GET     -1.521e-01  2.842e-02  -5.352 8.72e-08 ***
## MHLTH_SATISFD -8.189e-02  2.398e-01  -0.342  0.73268
## MHLTH_DIFFCLT -1.567e-01  2.782e-02  -5.634 1.76e-08 ***
## SOCIAL1       -5.947e-04  3.154e-02  -0.019  0.98495
## SOCIAL2        3.245e-02  2.997e-02   1.082  0.27904
## SUPPORT1       3.358e-02  4.181e-02   0.803  0.42183
```

```
## SUPPORT2      -2.446e-02  3.686e-02  -0.664  0.50692
## SUPPORT3      -6.205e-03  4.139e-02  -0.150  0.88084
## SUPPORT4       3.164e-02  1.464e-02   2.161  0.03070 *
## INCOME         7.803e-03  1.875e-02   0.416  0.67733
## EXPNS_DIF      2.426e-02  8.511e-02   0.285  0.77558
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12867.68  on 13600  degrees of freedom
## Residual deviance:   164.96  on 13565  degrees of freedom
## AIC: 236.96
##
## Number of Fisher Scoring iterations: 16
```

Interaction models

```
glm_interaction_model <- glm(MHLTH_NEED ~ TBIRTH_YEAR * ANXIOUS + RHISPANIC * RRACE + EEDUC * MS,
                             data = train_data, family = binomial)
summary(glm_interaction_model)
```

```
##
## Call:
## glm(formula = MHLTH_NEED ~ TBIRTH_YEAR * ANXIOUS + RHISPANIC *
##     RRACE + EEDUC * MS, family = binomial, data = train_data)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -5.644e+01  4.725e+00 -11.944  < 2e-16 ***
## TBIRTH_YEAR            2.942e-02  2.405e-03  12.229  < 2e-16 ***
## ANXIOUS                1.256e+01  1.252e+00  10.034  < 2e-16 ***
## RHISPANIC              2.175e-01  1.254e-01   1.735 0.082788 .
## RRACE                  3.314e-01  8.710e-02   3.805 0.000142 ***
## EEDUC                 -3.347e-02  1.646e-02  -2.033 0.042019 *
## MS                     2.023e-02  1.490e-02   1.357 0.174678
## TBIRTH_YEAR:ANXIOUS -6.482e-03  6.417e-04 -10.101  < 2e-16 ***
## RHISPANIC:RRACE     -1.627e-01  6.839e-02  -2.379 0.017357 *
## EEDUC:MS            -8.885e-03  4.550e-03  -1.953 0.050872 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12868  on 13600  degrees of freedom
## Residual deviance: 12627  on 13591  degrees of freedom
## AIC: 12647
##
## Number of Fisher Scoring iterations: 7
```

Model evaluation with the test set

```
# Model evaluation with the test set
# Predict on test data
predictions_ols <- predict(ols_model, test_data)
predictions_glm <- predict(glm_model, test_data, type = "response")

# For OLS, we may look at the RMSE
rmse_ols <- RMSE(predictions_ols, test_data$MHLTH_NEED)

# For GLM, since it is a classification, we may look at the accuracy or AUC
confusion_matrix <- confusionMatrix(as.factor(ifelse(predictions_glm > 0.5, 1, 0)),
                                    as.factor(test_data$MHLTH_NEED))
accuracy_glm <- confusion_matrix$overall['Accuracy']

# Output the results
print(paste("RMSE for OLS model:", rmse_ols))
```

```
## [1] "RMSE for OLS model: 33.935393290247"
```

```
print(paste("Accuracy for GLM model:", accuracy_glm*100))
```

```
## [1] "Accuracy for GLM model: 99.7940974605353"
```

Interpretation

Data Preparation and Model Fitting

The data was split into a training set (70%) and a test set (30%) using `createDataPartition` to ensure reproducibility, with `set.seed(123)`. The dimensions of the training and test sets are 49808x36 and 21344x36, respectively, indicating 36 variables were used, including the response variable `MHLTH_NEED`.

Ordinary Least Squares (OLS) Model

An OLS regression model was fit to predict `MHLTH_NEED` using all other variables as predictors. The model's residuals range significantly, indicating possible outliers or a non-linear relationship not captured by the model.

Key significant predictors include `TBIRTH_YEAR`, `RHISPANIC`, `EEDUC`, `MS`, household variables (`THHLD_NUMPER`, `THHLD_NUMKID`, `KIDS_LT5Y`, `KIDS_5_11Y`, `KIDS_12_17Y`), `CURFOODSUF`, `CHILDFOOD`, `ANXIOUS`, `INTEREST`, and several health insurance variables. `MHLTH_SATISFD`, `SOCIAL2`, `SUPPORT1`, and `INCOME` also show significant effects on mental health needs.

The coefficients for these variables suggest that both demographic factors (such as race and education) and current living conditions (like household size and income level) play roles in mental health service needs. Importantly, variables like `ANXIOUS` indicate a direct relationship between mental health symptoms and the assessed need for mental health services.

The model explains a substantial portion of the variance in the response variable (`Multiple R-squared: 0.7192`), suggesting a good fit. However, the significance of `EXPNS_DIF` hints at financial strain's impact on mental health needs.

Generalized Linear Model (GLM) for Binary Outcomes

A logistic regression model was also fit with `MHLTH_NEED` as a binary outcome. A warning about fitted probabilities indicates extreme predictions, which could suggest overfitting or separation in the data.

The coefficients from this model differ from the OLS model, with `MS`, `MHLTH_GET`, `MHLTH_DIFFCLT`, and `SUPPORT4` as significant predictors. This change reflects the binary nature of the response variable in the GLM model compared to the continuous scale in the OLS model.

`MHLTH_GET` and `MHLTH_DIFFCLT` are particularly noteworthy, as they reflect the challenges in accessing mental health services and satisfaction with those services, respectively. Their significance aligns with the premise that difficulty in accessing and satisfaction with mental health services are critical factors in the assessed need for such services.

Interaction Models

An interaction model was also examined, looking specifically at the interaction between `TBIRTH_YEAR` and `ANXIOUS`, `RHISPANIC` and `RRACE`, and `EEDUC` and `MS`. The interactions between `TBIRTH_YEAR` and `ANXIOUS`, and `RHISPANIC` and `RRACE` were significant, suggesting that the effect of anxiety on mental health need varies by birth year and that the effect of race on mental health need varies by Hispanic ethnicity.

Model Evaluation

The RMSE for the OLS model was approximately 33.94, which is quite high, indicating poor model performance in predicting the exact levels of mental health needs.

The accuracy of the GLM model was exceptionally high (approximately 99.79%), which may indicate overfitting, especially given the warning about probabilities of 0 or 1 during model fitting.

Conclusion

The predictive models have yielded several significant insights into the factors influencing the assessment of children's need for mental health services. The OLS model identified key predictors such as birth year, ethnicity, education level, household composition, food security, and mental health symptoms, suggesting that these factors are influential in predicting mental health service needs. However, the relatively high RMSE indicates that the OLS model's predictions are not very close to the actual values on average.

The GLM model, on the other hand, shows extremely high accuracy, which may suggest overfitting. Such a high accuracy rate is unusual and warrants further investigation. Variables like access to mental health services and satisfaction levels are significant, underscoring the importance of service availability and quality in assessing mental health needs. Interaction effects in the GLM model also reveal complex relationships between variables, which indicates that the impact of certain factors on the need for mental health services is conditional upon other variables.

Overall, the analysis highlights the multifaceted nature of assessing children's mental health service needs and emphasizes the role of demographic, socio-economic, and mental health-related variables in this context.