ISE 537

Yuhan Hu, Naimur Rahman Chowdhury

Title- Meal gap analysis: Predicting needs for food-insecure persons in US

Abstract

Hunger is one of the major global issues and is critical to achieving United Nation's Sustainable Development Goals (SDG). About four million people in the US face hunger, including one in five children. In order to tackle this global problem, food banks, food pantries, and community organizations unite to help millions of people access affordable, nutritious food for themselves and their families. Although hunger can affect people from all walks of life, it affects different groups differently based on their geographic locations, race, ethnicity, and age. In this project, we use a historical dataset that combines Feeding America's (the largest non-profit hunger-relief organization in the US) analysis of food insecurity and food cost in the United States and develop a prediction model for estimating the amount of money needed by a food-insecure person to meet weekly food needs. Our prediction model includes critical decision variables, such as food insecurity rate, and food insecurity rate among different ethnicities and different ages. The prediction model can help hunger relief organizations make strategic decisions to allocate donations according to geographic, demographic locations, and racial characteristics. The results can thus improve equity and fairness in hunger relief operations and help achieve zero hunger in the US.

1 Problem Description

This project aims to develop a prediction model using US food insecurity data (FANO, 2023) based on the demographic and racial composition of people in different regions. The aim of the project is to analyze the impact of the composition of racial factors (e.g., food insecurity rate among white, black people, etc.) and demographic factors (food insecurity rate among older people, younger people, etc.) on the amount of money needed by a food-insecure person to meet weekly food needs. Using county-level aggregated data, we focus on developing a prediction model to estimate the monetary value for required food. The ultimate goal of this project is to provide decision-makers with a model that can be used to determine how much food to distribute in a region with a particular composition of racial and demographic characteristics. The result will help hunger relief organizations to ensure

better fairness and equity on distributing food.

2 Data Description

For the analysis, we use **Map the meal gap** data from Feeding America (FANO, 2023). The dataset includes food insecurity estimates disaggregated by race and ethnicity of groups: people who identified themselves as Black (all ethnicities), Latino or Hispanic (all races), and white non-Hispanic in the federal data, and different age groups. Unfortunately, results are not available for other groups, such as Asian, Native American, Pacific Islander, or multiple races. The unavailability of these federal data prevented the data set from making estimations on those groups. The variables used in our model from the data are listed as follows:

1. **Weighted weekly $ needed by FI (y)**: The amount of money needed by a food-insecure person to meet weekly food needs (numerical, Y)
2. **insec_rate**: Percentage of inhabitants who are food insecure (numerical)
3. **rate_black**: Percentage of Black inhabitants (all ethnicities) who are food insecure (numerical)
4. **rate_hispanic**: Percentage of Hispanic inhabitants (all races) who are food insecure (numerical)
5. **rate_non_hispanic**: Percentage of white, non-Hispanic inhabitants who are food insecure (numerical)
6. **rate_senior**: Percentage of seniors (60+) who are food insecure (numerical)
7. **rate_senior_very_low**: Percentage of seniors (60+) who are very low food secure (numerical)
8. **rate_older**: Percentage of older adults (50-59) who are food insecure (numerical)
9. **rate_older_very_low**: Percentage of older adults (50-59) who are very low food secure (numerical)
10. **rate_child**: Percentage of children (under 18) who are food insecure (numerical)
11. **percent_children**: Percent of food insecure children who live in households with income below 185% of the federal poverty line (the cutoff for many child nutrition programs) (numerical)

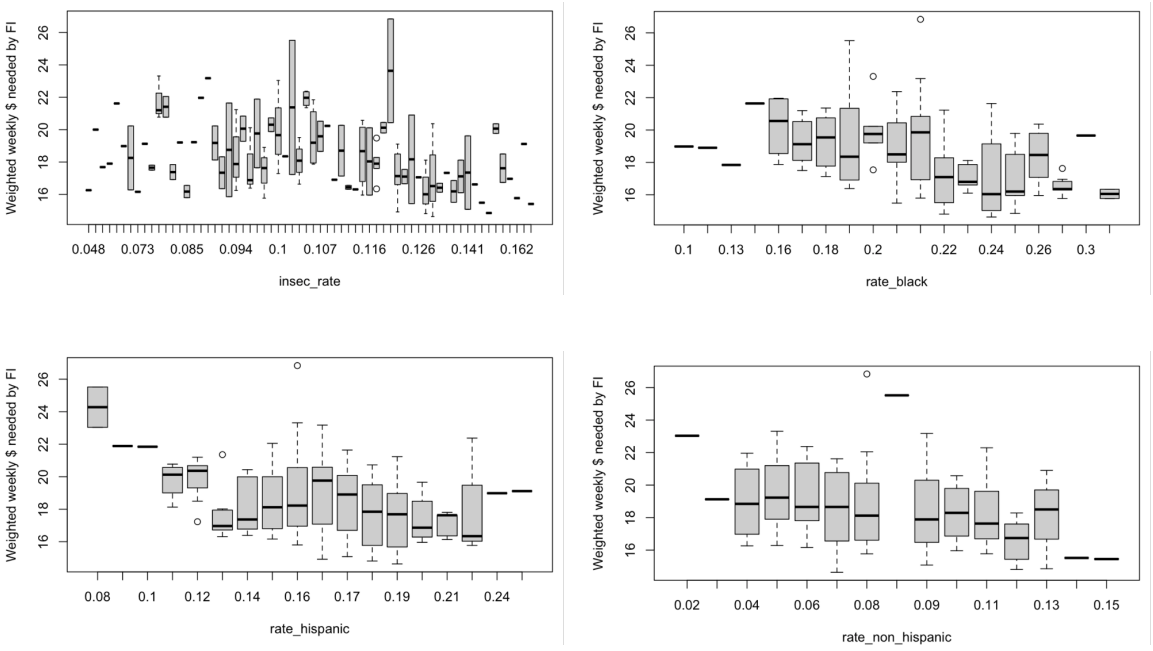
12. **cost_per_meal**: The average dollar amount spent on food per meal by food-secure individuals (numerical)

13. **percent_FI**: The percentage of food insecure individuals who live in households with income at or below low threshold in state (numerical)

3 Exploratory Data Analysis

3.1 *Box plot for each variable*

For exploratory data analysis, first, we observe a box plot for the response variable (*Weighted weekly $ needed by FI (y)*) against each variable. Figure 1 presents the box plots. We can observe from the plot that **cost_per_meal** has a very strong positive relationship with the response. We also observe that **rate_black**, **rate_hispanic**, **rate_child** and **percent_children** have somewhat negative linear relationship with the response. The relationship with **rate_senior** and **rate_senior_very_low** is somewhat positively related to the response. However, it is very difficult to interpret the other relationship without further analysis as the data points are ambiguous from visualization.



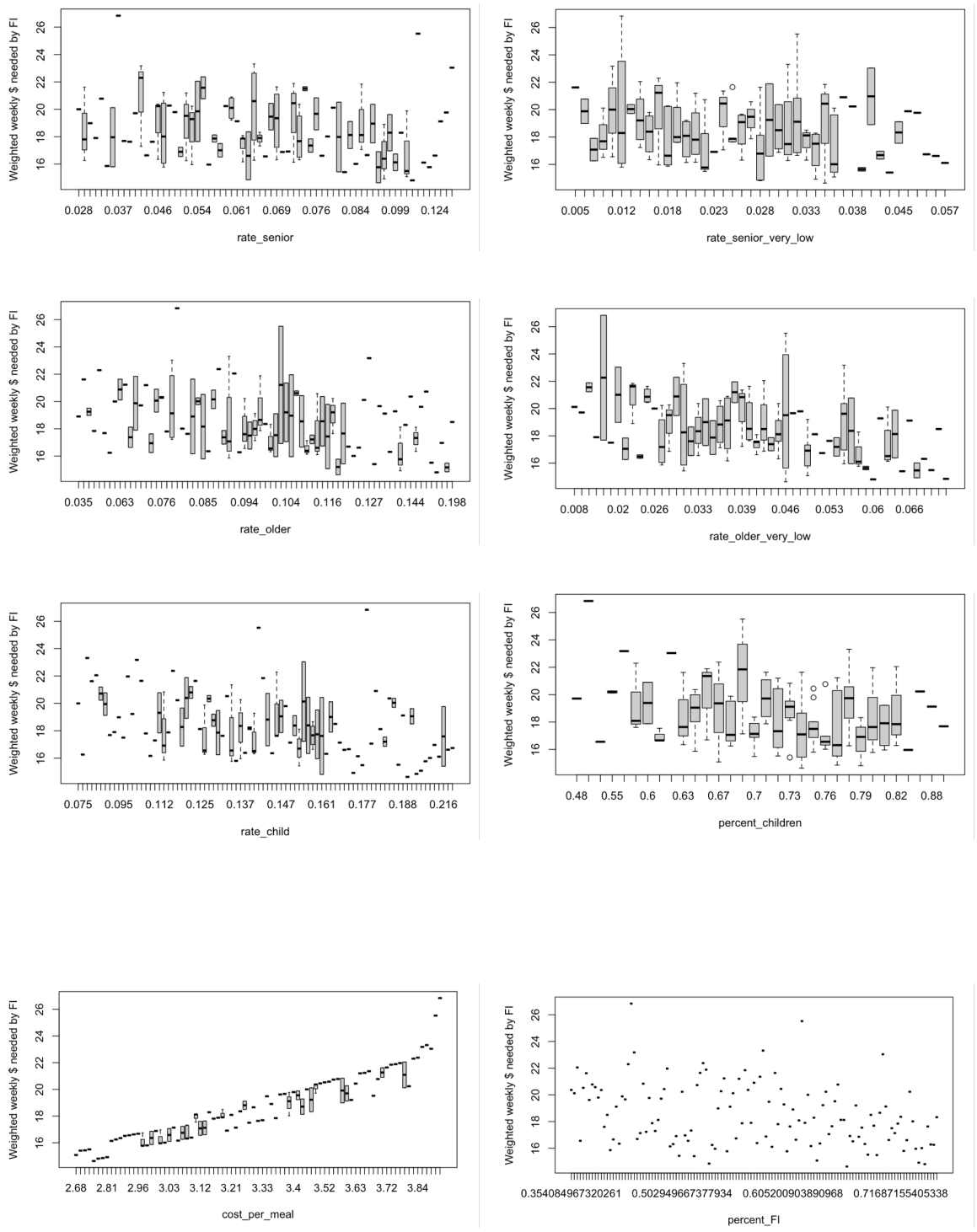


Figure 1: Box plot for response variable against each variable

3.2 *Correlation measure*

To better understand the relationship, we calculate the correlation coefficient for all variables. The correlation coefficient presented in Figure 2 reassures the positive relationship between the response variable and **cost_per_meal**. However, we can see a marginal to mild negative relationship with other variables. We also observe multicollinearity between some variables, such as **rate_older** and **rate_older_very_low**, etc. As a result, we further investigate the model.

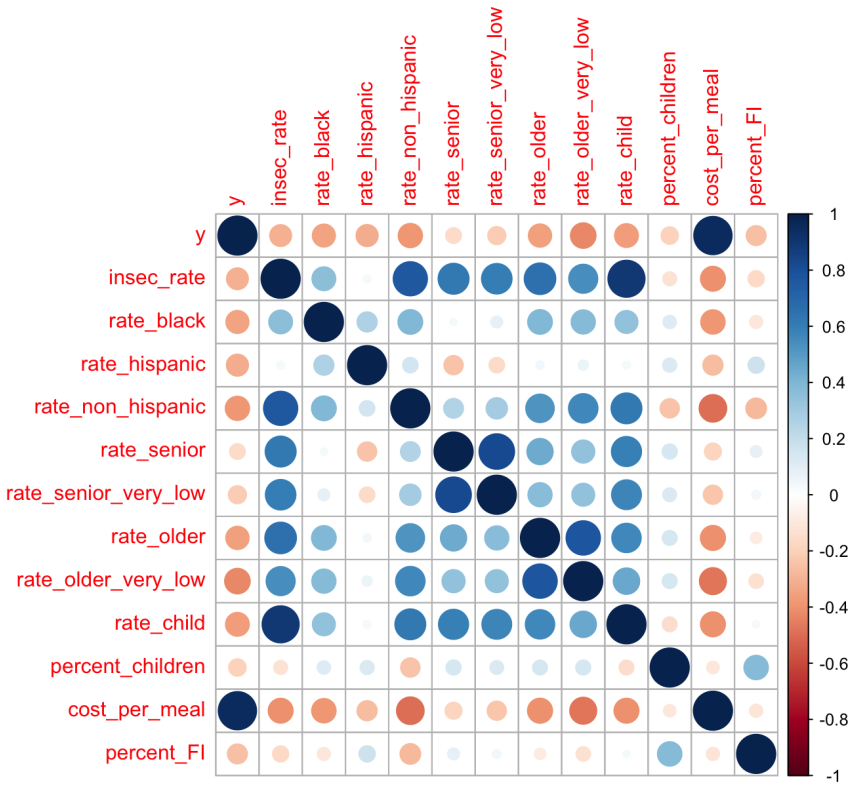


Figure 2: Correlation coefficient matrix

4 Model Fitting

4.1 *Full model*

For the first fitting, we use all variables to develop a full multiple regression model. The model summary is presented in Figure 3. From the model summary, it is clear that some of the variables (e.g., **insec_rate**, **cost_per_meal**) are statistically insignificant. How-

ever, most of the variables (6 out of 12) turned out to be statistically insignificant at 10% significance level.

```
## Call:
## lm(formula = y ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.38993 -0.36503  0.05839  0.39057  1.11398
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.05279    1.13766   0.046  0.96308
## insec_rate            23.40001    7.48255   3.127  0.00226 **
## rate_black             2.76913    1.85710   1.491  0.13880
## rate_hispanic         -3.55216    1.85703  -1.913  0.05837 .
## rate_non_hispanic     -0.68343    4.00327  -0.171  0.86476
## rate_senior           -0.52954    4.27753  -0.124  0.90170
## rate_senior_very_low   2.62418    9.09757   0.288  0.77355
## rate_older             0.18597    3.33611   0.056  0.95565
## rate_older_very_low   -1.57727    5.41934  -0.291  0.77157
## rate_child           -15.01448    3.94177  -3.809  0.00023 ***
## percent_children      -1.91352    0.83744  -2.285  0.02423 *
## cost_per_meal          6.15929    0.18116  33.999  < 2e-16 ***
## percent_FI            -1.50870    0.58567  -2.576  0.01132 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5475 on 110 degrees of freedom
## Multiple R-squared:  0.9539, Adjusted R-squared:  0.9488
## F-statistic: 189.5 on 12 and 110 DF,  p-value: < 2.2e-16
```

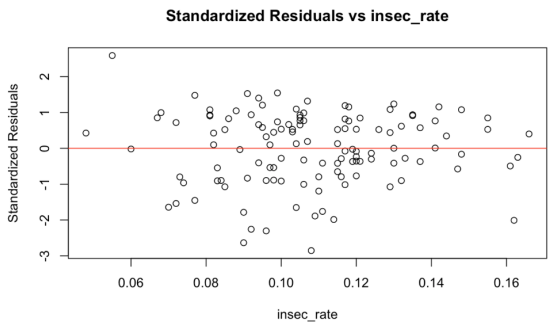
Figure 3: Full model summary

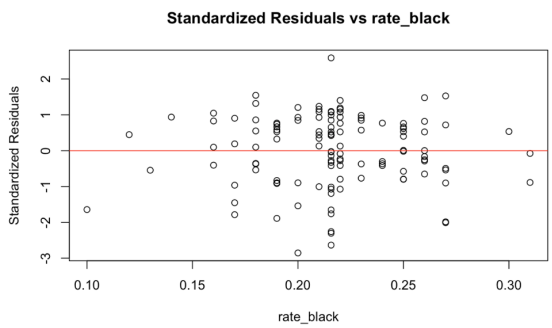
4.2 *Full model diagnostics*

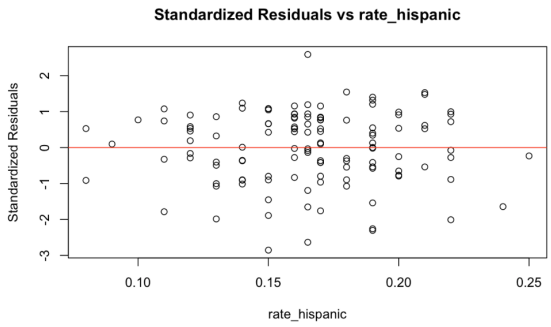
4.2.1 *Residual analysis*

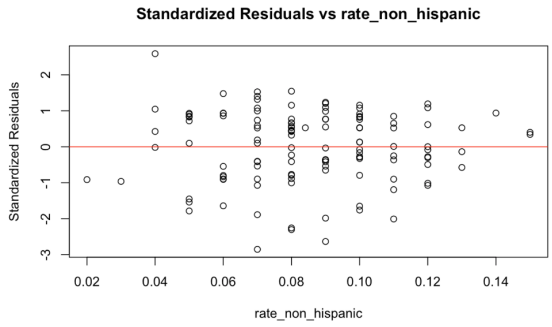
a. Standardized residuals vs individual predicting variables

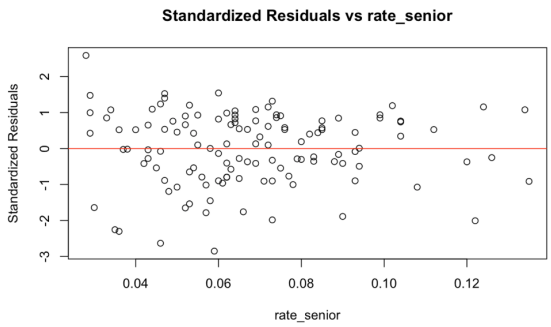
First, we calculate the standard residuals of the model and plot them against individual predicting variables. The plots shown in Figure 4 demonstrate that none of the plots have a U-like shape. Thus, we can roughly interpret that the linearity assumption is held.

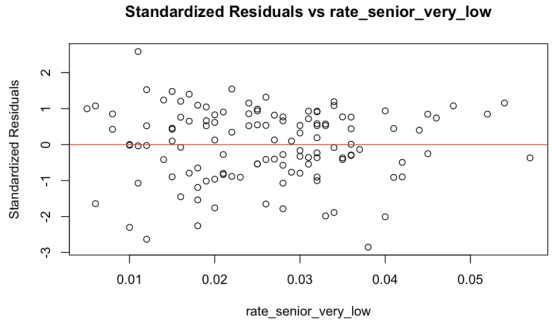

Standardized Residuals vs insec_rate
Standardized Residuals
-3
-2
-1
0
1
2
0.06
0.08
0.10
0.12
0.14
0.16
insec_rate


Standardized Residuals vs rate_black
Standardized Residuals
-3
-2
-1
0
1
2
0.10
0.15
0.20
0.25
0.30
rate_black


Standardized Residuals vs rate_hispanic
Standardized Residuals
-3
-2
-1
0
1
2
0.10
0.15
0.20
0.25
rate_hispanic


Standardized Residuals vs rate_non_hispanic
Standardized Residuals
-3
-2
-1
0
1
2
0.02
0.04
0.06
0.08
0.10
0.12
0.14
rate_non_hispanic


Standardized Residuals vs rate_senior
Standardized Residuals
-3
-2
-1
0
1
2
0.04
0.06
0.08
0.10
0.12
rate_senior


Standardized Residuals vs rate_senior_very_low
Standardized Residuals
-3
-2
-1
0
1
2
0.01
0.02
0.03
0.04
0.05
rate_senior_very_low

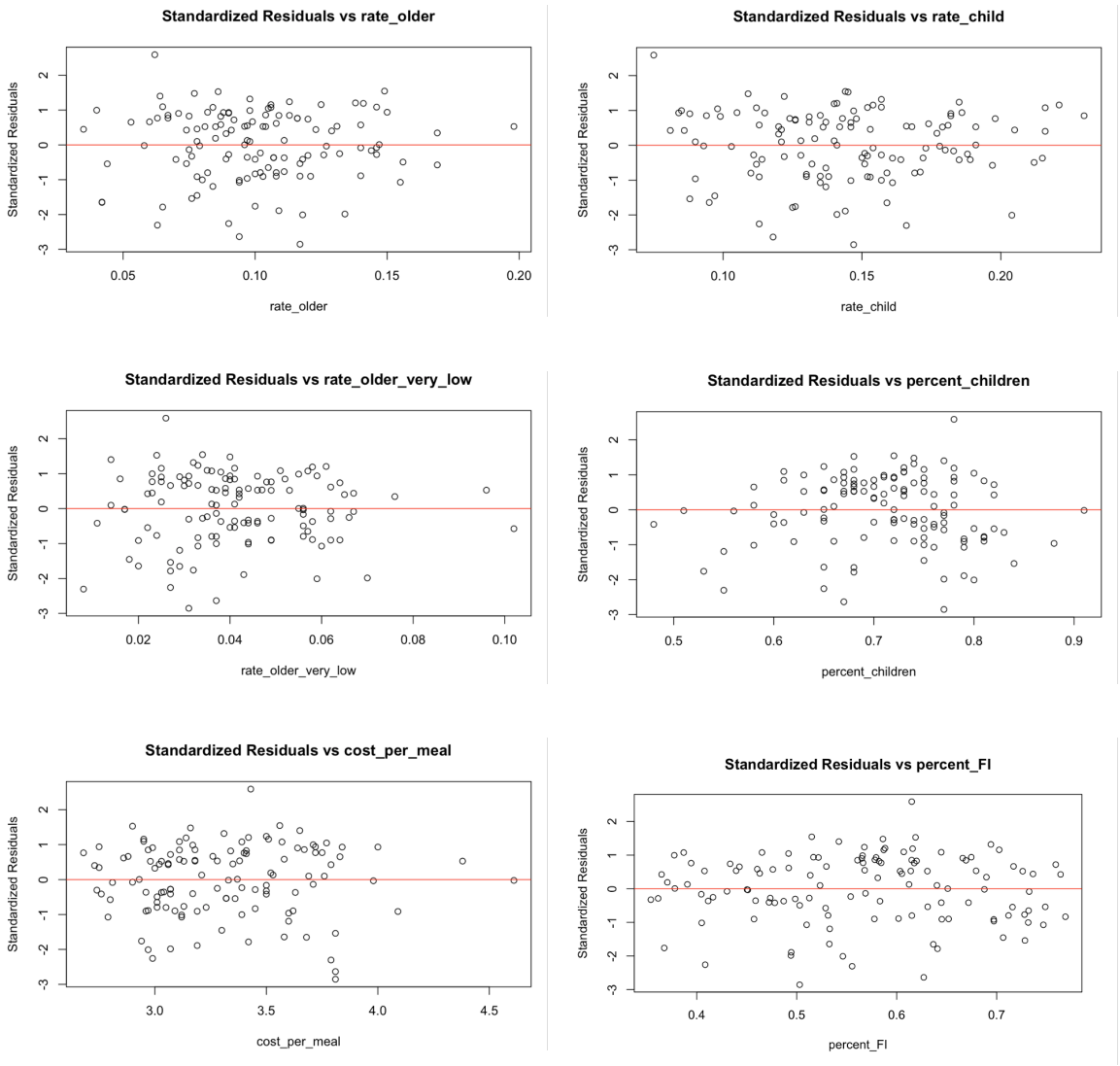


Figure 4: Standard residuals individual variable

b. Residuals vs Fitted values and Q-Q plot

We plot Residuals vs Fitted values to confirm the linearity assumption and constant variance assumption. The plot in Figure 5 shows clearly that there is no curved-like shape in the residuals (confirming linearity assumption), and the residuals are evenly distributed on both sides for most fitted values. Hence we can roughly estimate that the model holds the constant variance assumption. So, we do not do any transformation to retrieve the constant variance assumption.

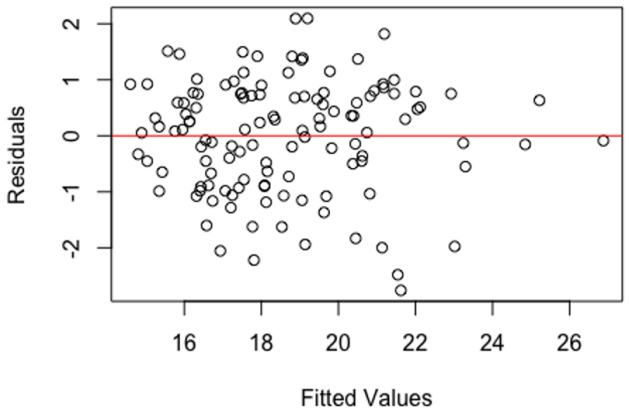


Figure 5: Residuals vs Fitted values

Furthermore, we performed Q-Q plotting in Figure 6 and saw that the normality assumption is held. The histogram also roughly depicts a normal distribution in Figure 7 also roughly depicts normal distribution.

To ensure that the model contains no outliers, we also calculate Cook's distance, and find out that the distances are significantly small (shown in Figure 8). Since all the assumptions hold, we do not perform any transformation on the dataset.

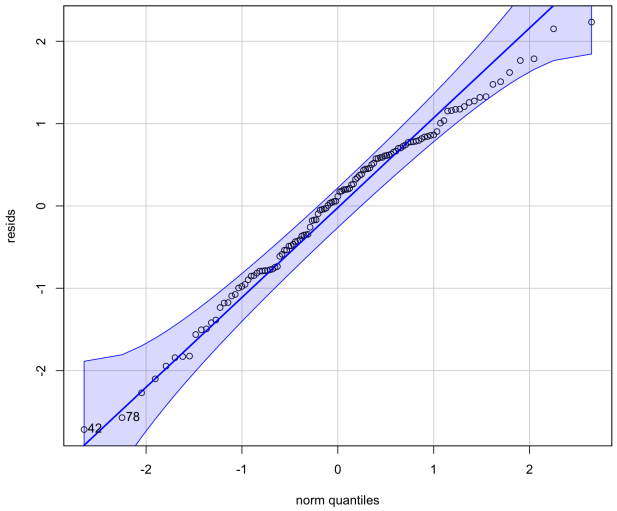


Figure 6: Q-Q Plot

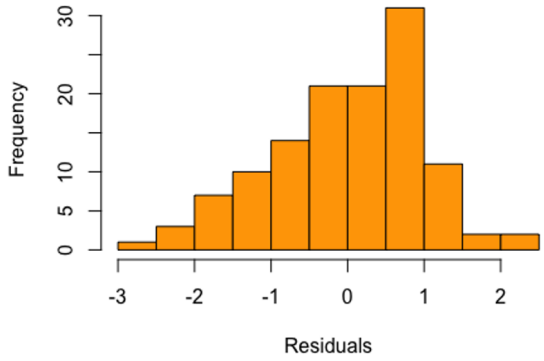


Figure 7: Histogram

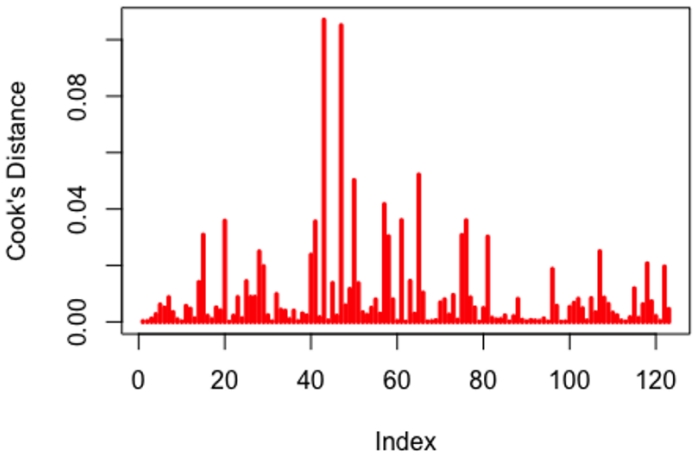


Figure 8: Cook's distance

4.3 Reduced model with significant predictors

Based on the results in 4.1, we extract the significant predictors at the 99% significance level, including insec_rate, rate_child, cost_per_meal, and percent_FI, to fit a multiple linear regression model, as shown in Figure 9. We also perform an ANOVA test to compare the reduced model with the full model, and the result is shown in Figure 10. It shows that the *p-value* is 0.11, which suggests that we fail to reject the null hypothesis that all the additional predictors in the full model are zero. Thus, the reduced model is preferred over the full model. However, since selecting variables based on the significance of individual coefficients is not accurate, we further conduct variable selection methods on the dataset. To make the model more interpretative, we perform variable selection. All the variable selection methods are described in this section.

```
Call:
lm(formula = y ~ insec_rate + rate_child + cost_per_meal + percent_FI,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.26248 -0.38812  0.09678  0.42781  1.11046

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.5775     0.8174  -1.930  0.05601 .
insec_rate     18.6569     5.3257   3.503  0.00065 ***
rate_child    -10.2219     3.5290  -2.897  0.00450 **
cost_per_meal   6.3350     0.1609  39.384  < 2e-16 ***
percent_FI     -2.4710     0.5168  -4.782 5.07e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5559 on 118 degrees of freedom
Multiple R-squared:  0.9472,    Adjusted R-squared:  0.9454
F-statistic: 529.6 on 4 and 118 DF,  p-value: < 2.2e-16
```

Figure 9: Reduced model with significant predictors

```
Analysis of Variance Table

Model 1: y ~ insec_rate + rate_child + cost_per_meal + percent_FI
Model 2: y ~ insec_rate + rate_black + rate_hispanic + rate_non_hispanic +
    rate_senior + rate_senior_very_low + rate_older + rate_older_very_low +
    rate_child + percent_children + cost_per_meal + percent_FI
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    118 36.465
2    110 32.485  8      3.98 1.6846   0.11
```

Figure 10: Reduced model with significant predictors

5 Variable selection

5.1 Stepwise regression model

5.1.1 *Forward selection*

First, we use the forward selection model to see how the model selects the variables. The summary of the forward selection model is presented in Figure 11. In the forward selection model, six variables are statistically significant including **cost_per_meal, percent_FI, insec_rate, rate_Hispanic, rate_child, and percent_children**.

5.1.2 *Backward selection*

Similarly, we perform backward selection and present the summary in Figure 12. The same predictors are significant for this method as in the forward selection.

```
summary(forward)

##
## Call:
## lm(formula = y ~ cost_per_meal + percent_FI + insec_rate + rate_child +
##     percent_children + rate_hispanic + rate_black, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3495 -0.3761  0.0869  0.3943  1.1187
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.08465    1.03582  -0.082  0.93501
## cost_per_meal      6.18745    0.15881  38.961  < 2e-16 ***
## percent_FI        -1.48417    0.56011  -2.650  0.00919 **
## insec_rate        22.46017    5.14725   4.364 2.81e-05 ***
## rate_child       -14.56810    3.69732  -3.940  0.00014 ***
## percent_children  -1.86570    0.73298  -2.545  0.01224 *
## rate_hispanic     -3.65845    1.69016  -2.165  0.03249 *
## rate_black         2.66139    1.75962   1.512  0.13315
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5363 on 115 degrees of freedom
## Multiple R-squared:  0.9537, Adjusted R-squared:  0.9509
## F-statistic: 338.6 on 7 and 115 DF,  p-value: < 2.2e-16
```

Figure 11: Forward selection model

```
summary(backward)

##
## Call:
## lm(formula = y ~ insec_rate + rate_black + rate_hispanic + rate_child +
##     percent_children + cost_per_meal + percent_FI, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3495 -0.3761  0.0869  0.3943  1.1187
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.08465    1.03582  -0.082  0.93501
## insec_rate        22.46017    5.14725   4.364 2.81e-05 ***
## rate_black         2.66139    1.75962   1.512  0.13315
## rate_hispanic     -3.65845    1.69016  -2.165  0.03249 *
## rate_child       -14.56810    3.69732  -3.940  0.00014 ***
## percent_children  -1.86570    0.73298  -2.545  0.01224 *
## cost_per_meal      6.18745    0.15881  38.961  < 2e-16 ***
## percent_FI        -1.48417    0.56011  -2.650  0.00919 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5363 on 115 degrees of freedom
## Multiple R-squared:  0.9537, Adjusted R-squared:  0.9509
## F-statistic: 338.6 on 7 and 115 DF,  p-value: < 2.2e-16
```

Figure 12: Backward selection model

5.1.3 *Step-wise regression*

Finally, we perform step-wise regression and present the summary in Figure 13.

```
summary(stepwise)

##
## Call:
## lm(formula = y ~ cost_per_meal + percent_FI + insec_rate + rate_child +
##     percent_children + rate_hispanic + rate_black, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3495 -0.3761  0.0869  0.3943  1.1187
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.08465    1.03582  -0.082  0.93501
## cost_per_meal      6.18745    0.15881  38.961  < 2e-16 ***
## percent_FI        -1.48417    0.56011  -2.650  0.00919 **
## insec_rate        22.46017    5.14725   4.364 2.81e-05 ***
## rate_child       -14.56810    3.69732  -3.940  0.00014 ***
## percent_children  -1.86570    0.73298  -2.545  0.01224 *
## rate_hispanic     -3.65845    1.69016  -2.165  0.03249 *
## rate_black         2.66139    1.75962   1.512  0.13315
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5363 on 115 degrees of freedom
## Multiple R-squared:  0.9537, Adjusted R-squared:  0.9509
## F-statistic: 338.6 on 7 and 115 DF,  p-value: < 2.2e-16
```

Figure 13: Step wise regression model

We compare the models based on different parameters and present them in Table 1. We can see that the forward, backward, and stepwise; three models improve from the full model and converge to the same coefficient values. In all three stepwise regression, there are 6 variables that are selected including cost_per_meal, percent_FI, insec_rate, rate_Hispanic, rate_child, and percent_children.

However, since we found high multicollinearity between some of the variables from the

Model	*Adjusted R^2*	*Mallow's Cp*	*AIC*	*BIC*
Model1 (full model)	0.943	13	201.633	241.004
Model2 (reduced model)	0.940	9.590	198.962	215.835
Model3 (forward)	0.944	3.751	192.672	215.169
Model4 (backward)	0.944	3.751	192.672	215.169
Model5 (stepwise)	0.944	3.751	192.672	215.169

Table 1: Model Comparison

correlation matrix, we performed Ridge regression. In addition to that, we also use LASSO and elastic net for regularization before concluding on the final model selected.

6 Model Regularization

6.1 Ridge regression

We perform Ridge regression to reduce multicollinearity and prevent coefficients from being too large. We First apply ridge regression for a range of penalty constants, and finally, λ is selected to minimize the Cross-validation (CV) score. We use 10-fold CV to find the optimal λ that minimizes the cross-validation error. At the optimal $\lambda = 1.25$, all 12 predictors are selected since ridge regression does not perform variable selection but shrinkage. The selected coefficients for the optimal λ are presented in Figure 14.

```
13 x 1 sparse Matrix of class "dgCMatrix"
                                    s1
(Intercept)           3.38178433
insec_rate            8.10177405
rate_black            1.18833803
rate_hispanic        -3.90289946
rate_non_hispanic     0.04929012
rate_senior           3.13662306
rate_senior_very_low -0.44132006
rate_older            0.85127369
rate_older_very_low  -7.33302504
rate_child           -7.16109949
percent_children     -1.93268815
cost_per_meal         5.52190936
percent_FI           -2.19004279
```

Figure 14: Ridge model

6.2 LASSO regression

After that, we performed LASSO regression for feature selection by making some coefficients zero. For LASSO, we get optimal $\lambda = 0.0032$, and the selected coefficients are presented in Figure 15. From LASSO, we can see that the coefficients for all predictors but **rate_older** are selected, and the coefficient path is shown in 16.

```
13 x 1 sparse Matrix of class "dgCMatrix"
                              s1
(Intercept)          -0.3385018
insec_rate           15.3329098
rate_black            1.7396979
rate_hispanic        -2.6331955
rate_non_hispanic     0.2827329
rate_senior           2.0199096
rate_senior_very_low  0.0130322
rate_older            .
rate_older_very_low  -2.0187633
rate_child           -9.6794259
percent_children     -1.8483310
cost_per_meal         6.2835205
percent_FI           -1.6533003
```

Figure 15: LASSO model

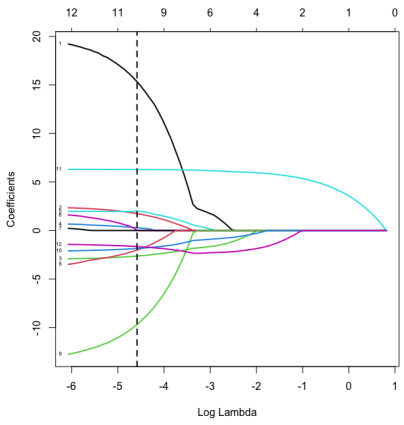


Figure 16: Coefficient path for LASSO

6.3 Elastic Net

Finally, we use the elastic net for feature selection and handle multicollinearity. The selected coefficients for scaled data are presented in Figure 17. Similar to LASSO, here, all predictors but **rate_older** are selected.

```
13 x 1 sparse Matrix of class "dgCMatrix"
                              s1
(Intercept)          -0.1669394
insec_rate           14.8301491
rate_black            1.7570153
rate_hispanic        -2.7071095
rate_non_hispanic     0.4056122
rate_senior           2.1730064
rate_senior_very_low  0.0175783
rate_older            .
rate_older_very_low  -2.3419758
rate_child           -9.5409106
percent_children     -1.8543301
cost_per_meal         6.2488562
percent_FI           -1.6841551
```

Figure 17: Net elastic model

7 Final model selection

7.1 *Prediction results*

We use all models to test our separated dataset and calculate errors from each model. For error measurement, we calculate mean squared prediction error (MSPE), mean absolute prediction error (MAE), mean absolute percentage error(MAPE), and prediction measure (PM). In all cases, the stepwise regression model results in the lowest error (shown in Table 3). So, we choose the stepwise regression model as the best model to interpret out data.

Model	MSPE	MAE	MAPE	PM
Model1 (full)	0.287	0.460	0.024	0.053
Model2 (reduced)	0.290	0.463	0.023	0.049
Model3 (forward)	0.264	0.443	0.023	0.049
Model4 (backward)	0.264	0.443	0.023	0.049
Model5 (stepwise)	0.264	0.443	0.023	0.049
Model6 (ridge)	0.271	0.449	0.023	0.051
Model7 (Lasso)	0.287	0.460	0.024	0.053
Model8 (Elastic net)	0.285	0.458	0.024	0.053

Table 2: Variable mapping across all models

7.2 *Comparison of all variable selection approaches*

We compare all the variable selection methods and explore the variables that are selected by all methods. As shown in Table 3, we find there are three predictors that are significant including **rate_child**, **cost_per_meal**, **percent_FI** for all models. Thus, we interpret that the amount of money needed by a food-insecure person to meet weekly food needs heavily depends on child food insecurity rate, the cost per meal, and the percentage of food-insecure individuals who live in households with income at or below a low threshold in the state.

Predictor	Model1 (full model)	Model2 (reduced)	Model3 (forward)	Model4 (backward)	Model5 (stepwise)	Model6 (ridge)	Model7 (Lasso)	Model8 (elastic net)
insec_rate	X	X	X	X	X	X	X	X
rate_black	X					X	X	X
rate_hispanic	X		X	X	X	X	X	X
rate_non_hispanic	X					X	X	X
rate_senior	X					X	X	X
rate_senior_very_low	X					X		
rate_older	X					X	X	X
rate_older_very_low	X					X	X	X
rate_child	X	X	X	X	X	X	X	X
percent_children	X		X	X	X	X	X	X
cost_per_meal	X	X	X	X	X	X	X	X
percent_FI	X	X	X	X	X	X	X	X

Table 3: Prediction Errors

8 Discussion and Conclusion

In this project, we tried to develop a prediction model to estimate the weekly monetary value of required food for food-insecure people. The dataset for this study included food insecurity estimates disaggregated by race and ethnicity of groups, different age groups, and rate of overall food insecurity in different samples, etc. The stepwise regression model provided the best model with the lowest error among all models analyzed. From the best model, we can see that the response variable heavily depends on the overall percentage of insecure inhabitants and the rate of food-insecure children, the proportion of Hispanic people, and the meal cost. The results practically depict the real-life scenario as the number of children, people in a particular racial group, and cost difference can heavily influence the monetary value of required food for people in need. So, we can conclude that the result of the model analysis provides a good estimation for the response variable.

References

FANO (2023). Map the meal gap data. Technical report, Feeding America, https://www.feedingamerica.org/research/map-the-meal-gap/by-county.