**Exploration of center-based childcare costs for households with different numbers of working parents**

**Background/Significance**

Historically, studies of childcare prices in the United States have been limited by a lack of local-level price data. To better understand how childcare prices affect women’s labor supply, the Women’s Bureau contracted ICF International to compile a database of county-level childcare prices from state studies conducted between 2008 and 2018. We chose to use this data set to examine the across-county relationship between center-based childcare costs for non-school-aged children and households with different numbers of working parents in 2018. We chose to do this to try to identify the types of working parent dynamics most affected by high childcare costs across counties in the United States.

A previous study done by the Women’s Bureau (Landivar et al., 2023) found that median childcare costs range from between 8.0% and 19.3% of the median family income in the United States. Landivar et al. also found that counties with childcare prices that were twice as expensive as the median national childcare price had a 4 percent lower maternal employment rate. However, this study did not investigate the number of working parents in a household and how this variable could relate to childcare costs for households with children under 6 years old. Suppose we can find a good model to visualize this relationship. In that case, it can be useful to know which counties would need more resources to help with the burden of childcare costs based on the percentages of households with different numbers of working parents in those counties. Our research question aims to specifically explore the across-county relationship between center-based childcare costs and the percentage of households with varying numbers of working parents.

**Data Preparation and Methodology**

To begin our analysis we accessed the public-use data set from states’ child care market price surveys, in addition to county-level demographic and labor market data from the American Community Survey (ACS) 5-year estimates which originated from the National Database of Childcare Prices (NDCP). The childcare price information requested for this database came entirely from the market rate survey (MRS) final reports that states produced and/or the data files used to prepare the reports. The database offers childcare price data by childcare provider type, age of children, and county characteristics. Data is available from 2008 to 2018, though we focused exclusively on data from 2018 for this project. We chose the data from counties in 2018 because that was the most recent year that data was collected from counties, as well as the year with the most data available across all counties. There are a total of 34,567 rows in the original data set, where each row represents a given county’s childcare price data for a given year. There are 61 columns of variables related to demographic and labor market data.

We wanted to focus on examining childcare costs for households with children under 6 years old. Initially, we were unable to aggregate the data on center-based childcare costs for the different age groups described in the data set into one column. This was because the rest of the data was not split into the age groups, rather the age groups were organized into columns. To resolve this issue, we decided to investigate the correlation between each of the three age groups that spanned children under 6: infant, toddler, and preschool-aged children. The correlations between the three age groups were all approximately 0.96. Since they are highly correlated, we chose one age-group variable, *infant*, to represent the pattern for center-based childcare costs for households with children under 6 years of age.

Then, we had to begin the process of addressing the missing data in the data set. First, we chose to remove data from Colorado, New Mexico, and Indiana as those entire states were missing all center-based childcare cost data for 2018 (ICF, 2020). We still had missing values in the outcome variable even after removing data from these states so we decided to make a ‘missing’ column to indicate whether a county was missing childcare cost data. There were a variety of reasons why counties had missing data. Some counties collected data on childcare costs for different ages by month, so those counties’ values were made NA. There were also cases of low response rates, too few providers, or center-based care was not offered (ICF, 2020). Some counties did not indicate why they did not have data for some age groups. Therefore, we assumed that the data is missing not at random, and hypothesized that the missingness could be explained by the state where a county is located.

This hypothesis led us to make a regression tree *(Figure 1*) that would show whether *state* is an important variable for predicting if the childcare cost for infants is missing. We removed the fips county code since it is an ID variable and added a column for the state that the county is in. We generated a tree that predicts the missingness of center-based childcare costs based on all the predictor variables (replacing county fips code with the state name). From this tree, we found that the state *is* an indicator of missingness. We used the results from the tree to define the subset of counties that we would use in our analysis: *all* counties from AL, AR, AZ, CA, CT, DE, ID, IL, KS, KY, MA, MD, ME, MI, MN, MS, ND, NE, NH, NJ, NY, OH, OK, OR, PA, RI, SC, SD, TN, TX, UT, VA, VT, WA, WV, and WY; *some* counties from NC, WI, AK, MO, and HI.

Our variables of interest, those concerning the number of working parents in a household: households with children under 6 where both parents work, only the father works, only the mother works, or households with a single mother, are in terms of the total number of households in a county. We transformed these columns to be the percentage of households in the county so that these percentages could be compared across counties of different sizes.

Then, we investigated the correlation of our variables by examining scatterplots that show the relationships between the predictors and the response variable. Two of the scatterplots indicated that there might be another predictor at play, as points seemed to diverge into two lines as seen in scatter plots 2, 3, and 4 from *Figure 2*. To examine the counties with higher versus lower infant center-based childcare costs, we set a threshold at 220, as that seemed to be where the distribution of points split off into two directions. When looking at the geographic distribution of the states with larger versus lower childcare costs, we did not find any relationship between geographic region and whether the variable value exceeded our set threshold. We did find that many states did not have any counties with childcare costs exceeding the threshold and some had a few that did. Only Massachusetts and Connecticut had all infant childcare costs exceed this threshold. California had only 1/8 of counties with infant childcare costs below the threshold. Since these are relatively costly states to live in, we can consider an income-related variable in our model, we chose this as the median household income (MHI). After the entire data preparation process, we ended up with a dataset with 2458 rows and 65 columns.

**Results**

In the first-order linear regression model with all variables of interest, all predictors seem to be significant as evidenced by their low p-values (below the threshold of 0.05), as seen in *Figure 3*. To investigate if interaction terms would improve the model, we examined the model with *all* possible interaction terms but found that four interaction terms might be insignificant since they resulted in high p-values (*Figure 3*). Comparing the full interaction model and the reduced interaction model without the four “insignificant” terms, we observed no significant difference in how the two models fit the data through their residuals vs. fitted values plots (*Figure 4).* We then conducted an F-test comparing the two models and concluded that we did not have sufficient evidence (p-value = 0.2797) to argue that the added complexity of those additional interaction terms in the full model is worth analyzing. Looking at the model selection criteria, we saw that there is no change in R² and adjusted R² values between the two models (*Figure 6*). The reduced interaction model also has the smallest AIC and BIC values by a significant amount (*Figure 6*). Thus, we chose the reduced interaction model as our best interaction model. Comparing the best first-order model and the best interaction model, the models don’t seem to fit the data any differently in the residuals vs. fitted values plots (*Figure 5*). The F-test comparing the models gave sufficient evidence to reject the null hypothesis that the first-order model was best (p-value 0). Looking at the model selection criteria, the interaction model has the highest adjusted R² value, which attempts to avoid overfitting, and the smallest AIC and BIC values (*Figure 6*). Thus, this model has the smallest residuals and the fewest parameters. Therefore we chose the reduced interaction model as our final model.

The regression equation for the reduced interaction model can be described as (*Figure 3*):

Mean(*infant childcare cost*) = 33.92 - 6.226 (*both work*) - 44.42 (*mother works*) + 3.532 (*father works*) - 4.639 (*single mother works*) + 0.0032 (*median household income*) + 0.5312 (*both work\*single mom work*) + 2.85 (*mother work\*single mom work*) - 0.26 (*father work\*single mom work*) + 0.0008 (*mother work \* MHI*) - 0.00008 (*father work\*MHI*) + 0.00006 (*single mom work\*MHI*)

To understand the relationship between each of these variables and their interactions with the cost of center-based childcare for infants, we set all of the variables equal to their medians in the regression equation except for either *both work*, *mother works*, *father works*, or *single mother works* depending on which variable we wanted to investigate (*Figure 7*). We found that when the percentage of households where both parents work increases by 1, the cost of childcare decreases by $4.11 (*Scatterplot 1)*; when the percentage of households where only the mother works increases by 1, the cost of childcare increases by $5.48 (*Scatterplot 2*); when the percentage of households where only the father works increases by 1, the cost of childcare decreases by $1.79 (*Scatterplot 3*); and when the percentage of households with only a single mother working increases by 1, the cost of childcare increases by $1.90 (*Scatterplot 4*).

**Conclusions and Discussion**

Based on our final model, we concluded that when the percentage of households where only the mother works increases or when the percentage of households where only a single mother works increases, there is an increase in center-based childcare costs across counties. However, when the percentage of households where only the father works increases or when the percentage of households where both parents work increases, the center-based childcare costs decrease across counties. These results suggest that counties with higher percentages of working single mothers or working mothers only; may require more resources to help mitigate the economic burden of paying for center-based childcare for their infants.

Limitations in this study include the removal of a large subset of the data due to counties’ missing records of center-based childcare costs. Thus, our results are only generalizable to *all* the counties from AL, AR, AZ, CA, CT, DE, ID, IL, KS, KY, MA, MD, ME, MI, MN, MS, ND, NE, NH, NJ, NY, OH, OK, OR, PA, RI, SC, SD, TN, TX, UT, VA, VT, WA, WV, and WY; and *some* counties from NC, WI, AK, MO, and HI. In future explorations of this data set, the scope of this study can be expanded to consider the changes in county demographics across the 10 years that the data was collected to give a more complete picture of the relationship between these variables.

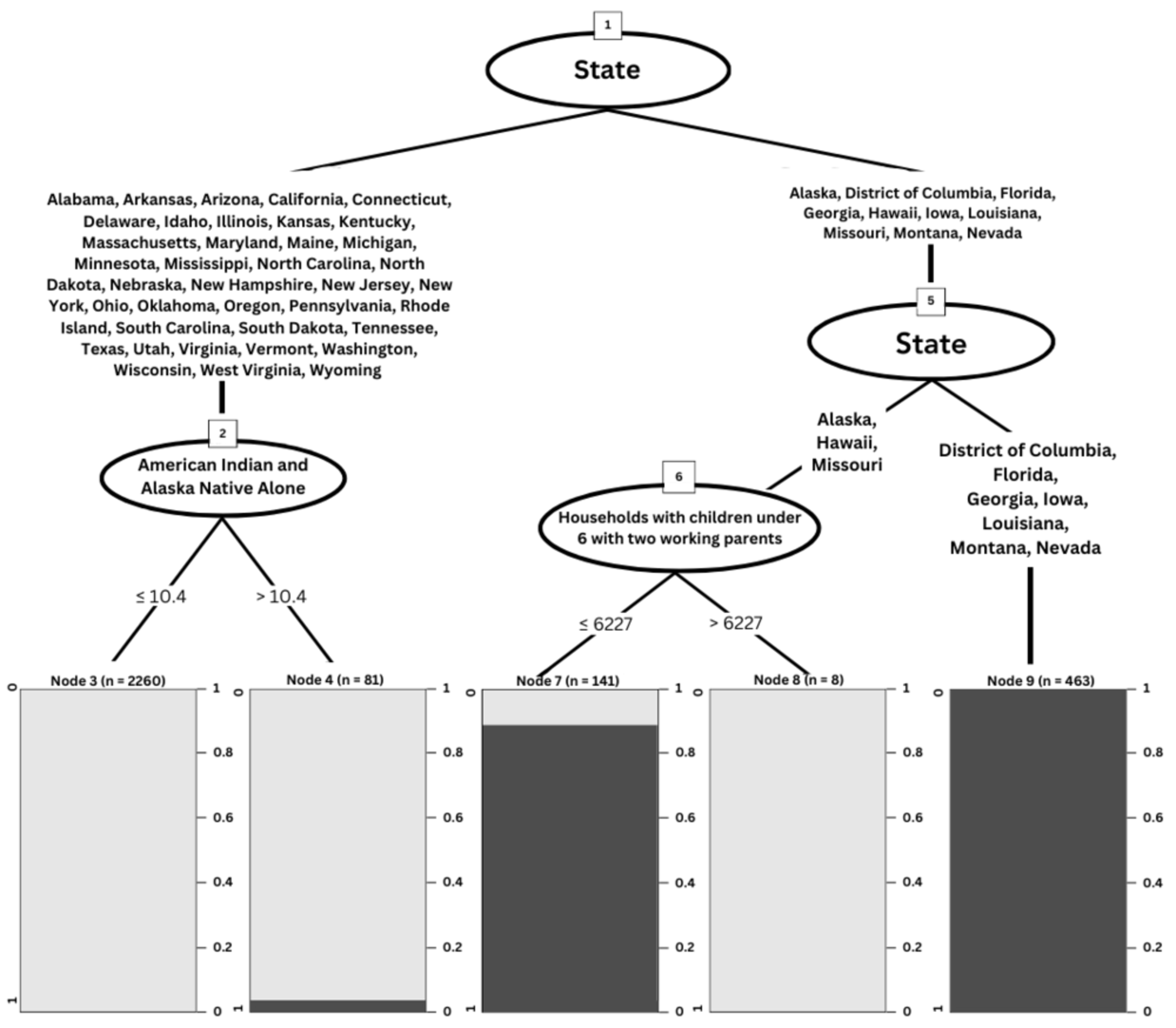
**References**

Landivar L.C., Graf N. L., Rayo G. A. (2023). Childcare Prices in Local Areas: Initial Findings from the National Database of Childcare Prices. *Women’s Bureau U.S. Department of Labor Issue Brief*, 1-11.

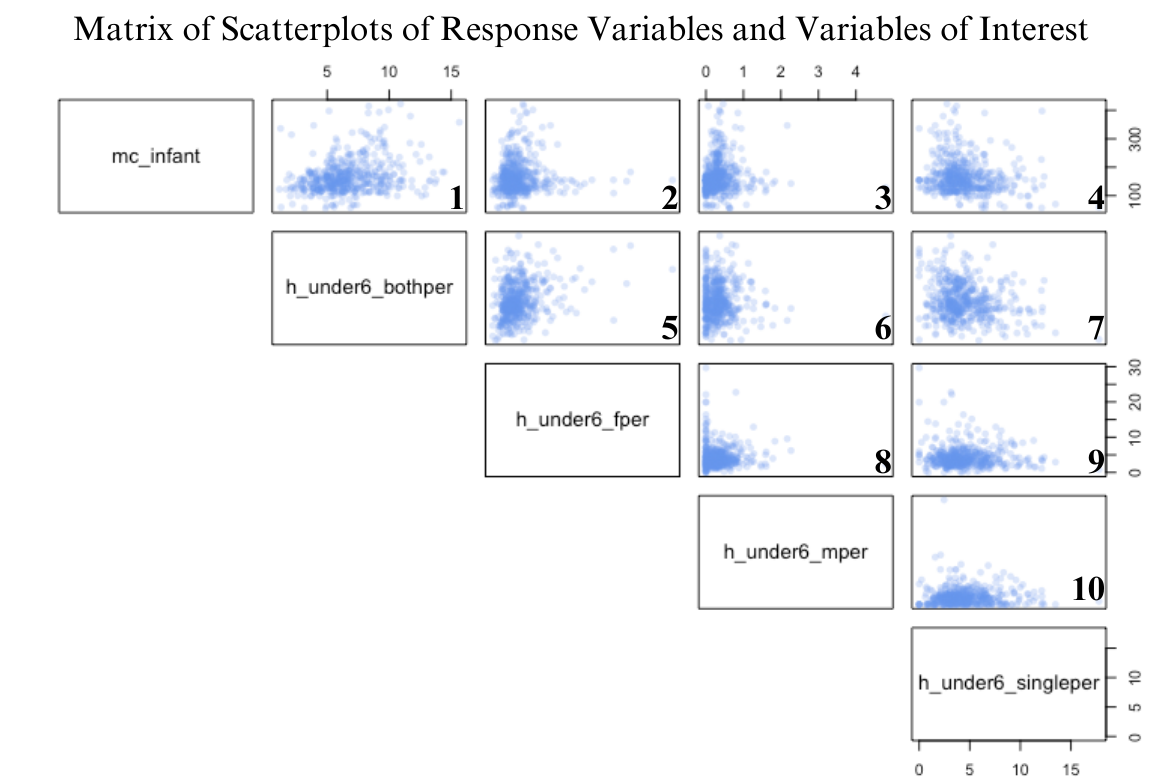
ICF. (2020). U. S. Department of Labor, Women’s Bureau National Database of Childcare Prices: Final Report. [Guidebook]. Available from https://www.dol.gov/agencies/wb/topics/featured-childcare.

Our data source: <https://github.com/rfordatascience/tidytuesday/blob/master/data/2023/2023-05-09/readme.md>

**Appendix**



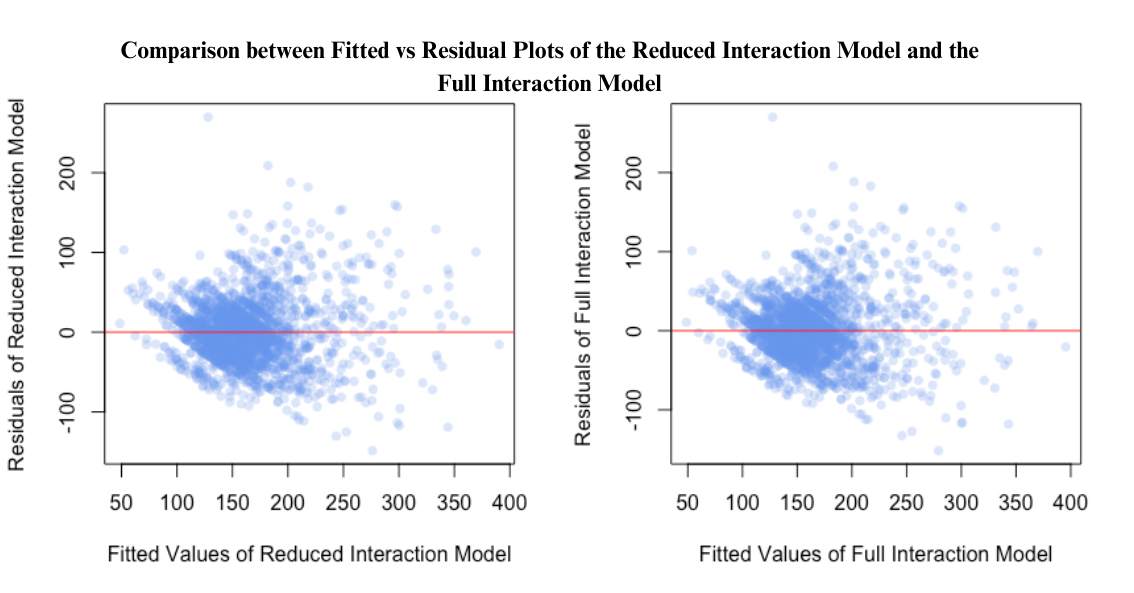
*Figure 1.* Tree predicting missingness of center-based childcare costs in counties



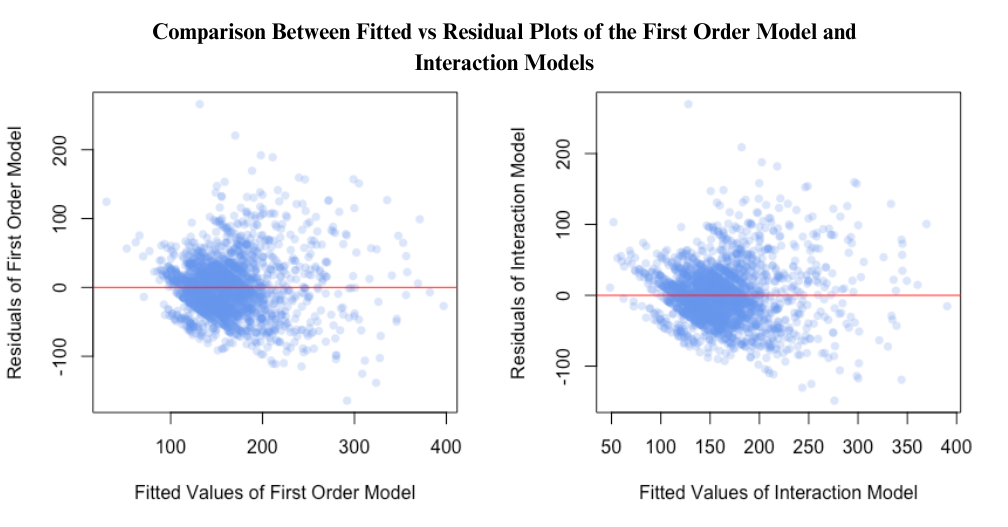
*Figure 2*. Matrix of Scatterplots of Response Variables and Variables of Interest

|  | First-Order Model | Reduced Interaction Model | Full Interaction Model |
| --- | --- | --- | --- |
| Intercept | 17.01\* | 33.92\* | 48.06\* |
| Both work | -3.992\* | -6.226\* | -8.551\* |
| Only mother works | -2.139\* | -44.42\* | -46.03\* |
| Only father works | 5.178\* | 3.532\* | 4.322\* |
| Only single mother works | 1.119\* | -4.629\* | -5.254\* |
| Median household income (mhi) | 0.003\* | 0.0032\* | 0.003\* |
| Both work:mother works |  |  | 1.667 |
| Both work:father works |  |  | -0.045 |
| Mother works:father works |  |  | 0.0046 |
| Both work:single mother works |  | 0.5312\* | 0.5627\* |
| Mother works:single mother works |  | 2.85\* | 2.902\* |
| Father works:single mother works |  | -0.26\* | -0.2755\* |
| Both work:mhi |  |  | 0.00004 |
| Mother works:mhi |  | 0.0008\* | 0.0006\* |
| Father works:mhi |  | -0.00008\* | -0.00009\* |
| Single mother works:mhi |  | 0.00006\* | 0.00007\* |

*Figure 3.* Fitted vs. Residual Plots of the Reduced and Full Interaction Models



*Figure 4.* Fitted vs. Residual Plots of the Reduced and Full Interaction Models



*Figure 5.* Fitted vs. Residual Plots of the First Order and Interaction Models

|  |  | Adjusted | BIC | AIC |
| --- | --- | --- | --- | --- |
| First-order model | 0.50 | 0.50 | 23297.60 | 23257.52 |
| Reduced interaction model | 0.52 | 0.52 | 23272.38 | 23197.94 |
| Full interaction model | 0.52 | 0.52 | 23298.18 | 23200.83 |

*Figure 6.* Table of model fitting criteria for each of the models

**

*Figure 7.* Scatterplots and regression lines for childcare costs vs. percentage of different numbers of working parents