Michelle Wang
Prof. Raman
ECON 255

Data Brief

Introduction

This report investigates the relationship between household spending on food and various variables using OLS regression. The dataset that we are working with is from the latest fielding of the Census's Household Pulse Survey – started in April 2020 – which is designed to collect data on the social and economic impacts of COVID-19 on American households in all 50 US states and the District of Columbia. The survey records data on the following topics:

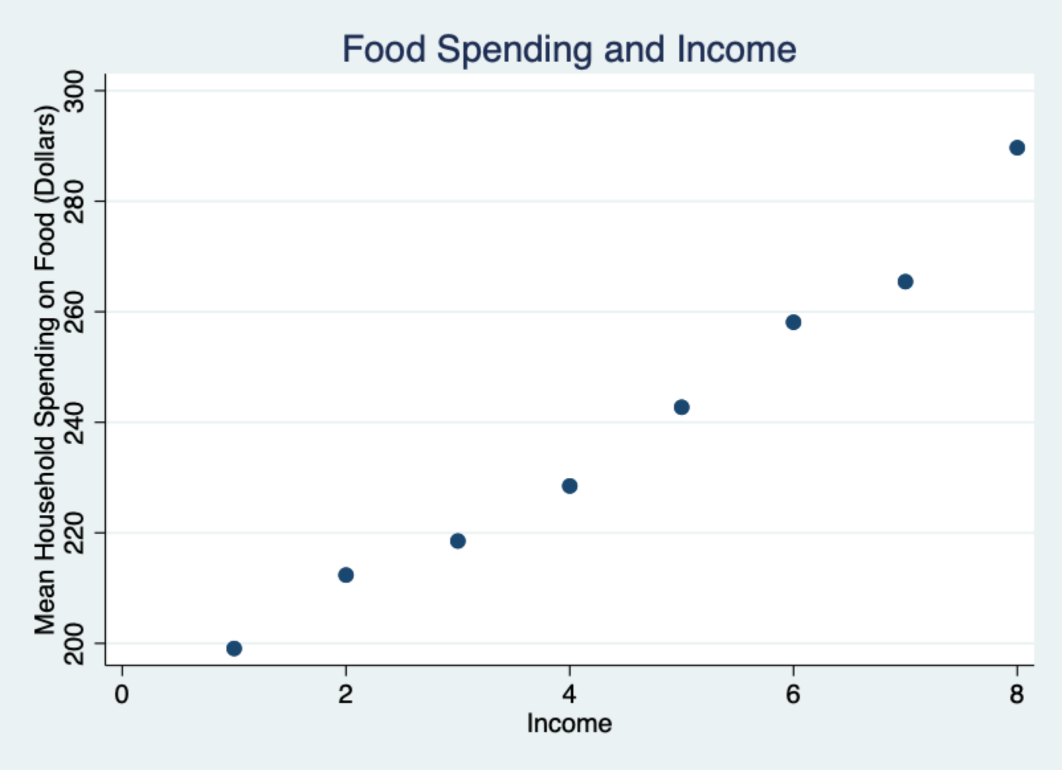
- Employment status
- Food security
- Housing security
- Physical and mental health
- Access to healthcare
- Education disruptions
- Household spending
- Various other measures of government interaction

The primary relationship of interest is the relationship between household spending on food and groceries (tspndfood) and income. This relationship is important because we want to know if income is related to how much people spend on food, and indirectly, their food security. We will also be exploring how additional variables such as the total number of people in the household, household spending on prepared meals, and difficulty paying household expenses are associated with household spending on food. The total number of people in the household is relevant because the more people there are, the more food they need to be fed and hence, more necessary spending on food. The household spending on prepared meals is relevant because if people spend a lot on prepared meals, they likely also spend a lot on food and groceries. Difficulty paying for household expenses is relevant because people's ability to afford expenses affects how they use their money, which impacts their spending on food.

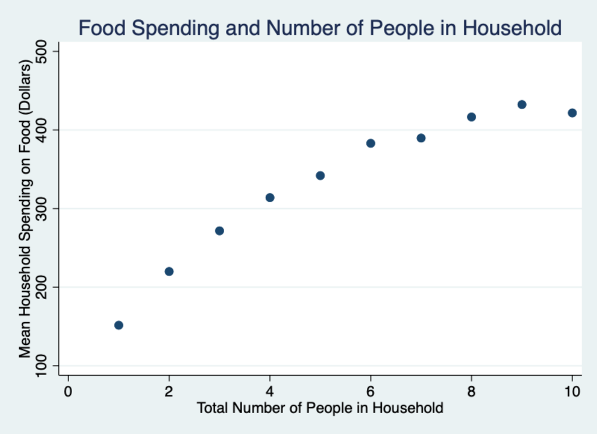
These relationships are important because they allow us to understand how household spending, financial well-being, and food security are related. This dataset is appropriate for investigating this relationship because it covers a large geographical area and contains data from all over the country about household spending on food, income, number of people in a household, etc.

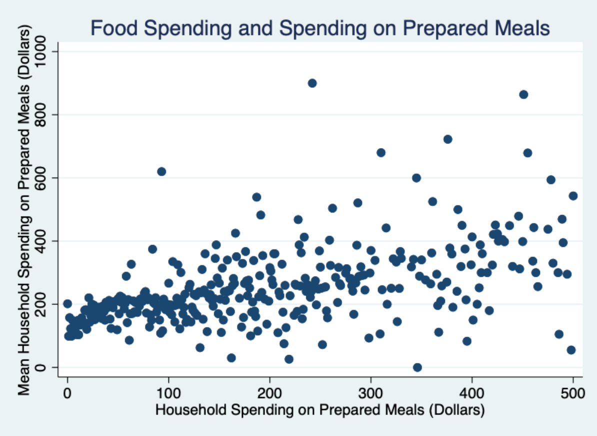
Some data hazards that this dataset may be associated with are Reinforces existing bias and Difficult to understand. It may "reinforce existing bias" such as the stereotype that low-income households are food insecure. The dataset may also be “difficult to understand” because many of the variables don’t have representative names, such as the household spending on food we use for analysis in this report, which is represented as tspndfood. Although there is documentation to help identify variables, this still makes the dataset confusing and difficult to work with.

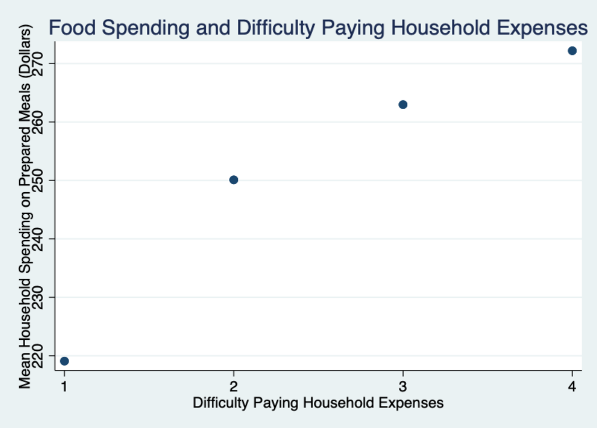
Descriptive statistics



This is a scatterplot of the mean household spending on food, measured in dollars, and income, measured in 8 groups:
1) Less than $25,000
2) $25,000 - $34,999
3) $35,000 - $49,999
4) $50,000 - $74,999
5) $75,000 - $99,999
6) $100,000 - $149,999
7) $150,000 - $199,999
8) $200,000 and above
It shows a clear positive linear relationship between household spending on food and income, which means that our conditional expectation of household spending on food increases as income increases.

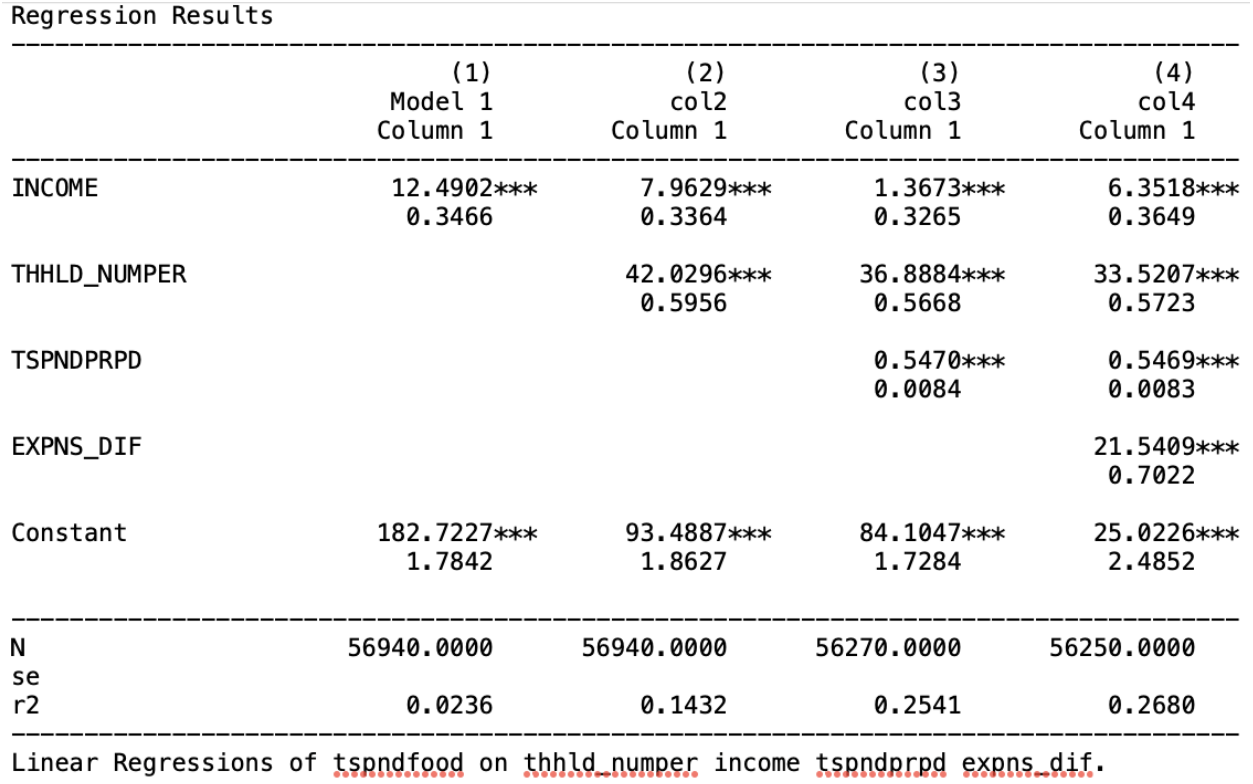

Food Spending and Number of People in Household


Food Spending and Spending on Prepared Meals


Food Spending and Difficulty Paying Household Expenses

These are scatterplots of mean household spending on food and our 3 additional variables: the total number of people in the household (thhld_numper), household spending on prepared meals in dollars (tspndprpd), and difficulty paying household expenses (expns_dif), which is measured in 4 levels: 1) Not at all difficult, 2) A little difficult, 3) Somewhat difficult, and 4) Very difficult. These scatterplots all show a positive linear relationship between household spending on food and the additional variable, which demonstrates that they are correlated with householding spending on food and can explain variability in our conditional expectations of household spending on food that our primary explanatory variable, income, cannot.

Econometric Analysis



```
Regression Results
-------------------------------------------------------------------------------------
                              (1)             (2)             (3)             (4)
                          Model 1            col2            col3            col4
                         Column 1        Column 1        Column 1        Column 1
-------------------------------------------------------------------------------------
INCOME                    12.4902***       7.9629***       1.3673***       6.3518***
                           0.3466          0.3364          0.3265          0.3649

THHLD_NUMPER                              42.0296***      36.8884***      33.5207***
                                           0.5956          0.5668          0.5723

TSPNDPRPD                                                  0.5470***       0.5469***
                                                           0.0084          0.0083

EXPNS_DIF                                                                 21.5409***
                                                                           0.7022

Constant                 182.7227***      93.4887***      84.1047***      25.0226***
                           1.7842          1.8627          1.7284          2.4852

-------------------------------------------------------------------------------------
N                      56940.0000      56940.0000      56270.0000      56250.0000
se
r2                         0.0236          0.1432          0.2541          0.2680
-------------------------------------------------------------------------------------
Linear Regressions of tspndfood on thhld_numper income tspndprpd expns_dif.
```

Discussion and Limitations

Regression 1 shows us that, on average, we predict that those in a higher income group will have $12.49 more household spending on food than those in the income group below them. The p-value (0.000) is less than a significance level of $\alpha = 0.05$, which means that we have sufficient evidence to reject the null hypothesis that there is no relationship between household spending on food and income.

Regression 2 shows us that, on average, for every additional person in the household, we predict about a $42.03 increase in the household spending on food, holding income constant. The p-value (0.000) is less than a significance level of $\alpha = 0.05$, which means that we have sufficient evidence to reject the null hypothesis that there is no relationship between household spending on food and the total number of people in the household. Additionally, on average, we predict that those in a higher income group will have $7.96 more household spending on food than those in the income group below them, holding the total number of people in the household constant. The p-value (0.000) is less than a significance level of =0.05, which means that we have sufficient evidence to reject the null hypothesis that there is no relationship between household spending on food and income.

There is reason to believe that the total number of people in the household is relevant to household spending on food because the addition of this variable increased the R-squared value of the regression, meaning that this model explains more of the variability in our predicted values of household spending on food than Regression 1. Additionally, the standard error decreased, meaning that the addition of the total number of people in the household to the regression makes our model more accurate.

Regression 3 shows us that, on average, for every one additional dollar spent on prepared meals, we predict about a $0.55 increase in the household spending on food, holding income and total number of people in the household constant. The p-value (0.000) is less than a significance level of $\alpha = 0.05$, which means that we have sufficient evidence to reject the null hypothesis that there is no relationship between household spending on food and household spending on prepared meals. Additionally, for every additional person in the household, we predict about a $36.89 increase in the household spending on food, holding income and household spending on prepared meals constant. The p-value (0.000) is less than a significance level of $\alpha = 0.05$, which means that we have sufficient evidence to reject the null hypothesis that there is no relationship between household spending on food and the total number of people in the household. Furthermore, on average, we predict that those in a higher income group will have $1.37 more household spending on food than those in the income group below them, holding the total number of people in the household and household spending on prepared meals constant. The p-value (0.000) is less than a significance level of =0.05, which means that we have sufficient evidence to reject the null hypothesis that there is no relationship between household spending on food and income.

There is reason to believe that household spending on prepared meals is relevant to household spending on food because the addition of this variable increased the R-squared value of the regression, meaning that this model explains more of the variability in our predicted values of household spending on food than Regression 2. Additionally, the standard error decreased, meaning that the addition of the household spending on prepared meals to the regression makes our model more accurate.

Regression 4 shows us that, on average, we predict that those at a higher level of difficulty paying for household expenses will have $21.54 more household spending on food than those in the level below them, holding income, the total number of people in the household, and household spending on prepared meals constant. The p-value (0.000) is less than a significance level of $\alpha = 0.05$, which means that we have sufficient evidence to reject the null hypothesis that there is no relationship between household spending on food and difficulty paying for household expenses. Additionally, on average, for every one additional dollar spent on prepared meals, we predict about a $0.55 increase in the household spending on food, holding income, total number of people in the household, and difficulty paying for household expenses constant. The p-value (0.000) is less than a significance level of $\alpha = 0.05$, which means that we have sufficient evidence to reject the null hypothesis that there is no relationship between household spending on food and household spending on prepared meals. Additionally, for every additional person in the household, we predict about a $33.52 increase in the household spending on food, holding income, household spending on prepared meals, and difficulty paying for household expenses constant. The p-value (0.000) is less than a significance level of $\alpha = 0.05$, which means that we have sufficient evidence to reject the null hypothesis that there is no relationship between household spending on food and the total number of people in the household. Additionally, on average, we predict that those in a higher income group will have $6.35 more household spending on food than those in the income group below them, holding the total number of people in the household, household spending on prepared meals, and difficulty paying for household expenses constant. The p-value (0.000) is less than a significance level of =0.05, which means that we have sufficient evidence to reject the null hypothesis that there is no relationship between household spending on food and income.

There is reason to believe that difficulty paying for household expenses is relevant to household spending on food because the addition of this variable increased the R-squared value of the regression, meaning that this model explains more of the variability in our predicted values of household spending on food than Regression 3. Additionally, the standard error decreased, meaning that the addition of difficulty paying for household expenses to the regression makes our model more accurate.

However, our analysis is limited by the fact that we did not check to see if the variance of the residual term is constant because we only looked at scatterplots of the mean household spending on food. This is because it is easier to assess the linearity assumption for regression analysis by looking at scatterplots of the mean instead of the scatterplots of all values where there are so many data points that a linear relationship is not as clear. If this assumption for OLS is not satisfied, which we get a hint of in the heteroskedastic scatterplot between household spending on food and household spending on prepared meals, this negatively affects the reliability and validity of our linear regression estimates. Additionally, our data limits us from making causal claims about the relationship we've identified because it is an observational study and not an experimental one; there was no experiment or treatment administered to the subjects being studied, which would have allowed us to quantify how household spending on food would change if the individual received a different treatment. Had there been a treatment administered, we would then have been able to observe differences between a control group and the treatment group to make causal claims about household spending on food.