The Impact of COVID-Related Health Concerns on US Consumer Demand

Tianxiao (Michael) Xu*

NYU Abu Dhabi, Class of 2023
tx542@nyu.edu

Abstract

With the onset of the COVID-19 pandemic and resulting nation-state policies, economies on a global scale have experienced an atypical recession, where sectors' susceptibility to shocks and the shape of their recoveries depend on their reliance on in-person contact, as opposed to sensitivity to real business cycles or income elasticities. On a state level in the United States, the substitution effect for aggregate demand becomes prominent, while any income losses become compensated. With the construction of a probabilistic event accounting for such substitution away from in-person services and towards remote services, a parsimonious linear probability model (LPM) empirically establishes the association between US consumers' COVID-related health concerns and the probability of such a substitution event occurring. Further principal component analysis and random forest classification findings corroborate the structural validity of the LPM specification, lending support to the central hypothesis.

Keywords: *COVID-19, public health, consumer demand, consumer behavior, substitution effect*

I. Introduction

Following the onset and spread of the COVID-19 pandemic and the corresponding measures taken by nation-state-level sociopolitical institutions, various large economies experienced reductions in economic activity as measured through GDP or its growth rate. For instance, offline consumption in China reduced by over 1.2 percent of its 2019 GDP, during a 12-week period following the initial COVID outbreak in Wuhan (Chen et al. 2021, pp. 308). In the EU, there was an initial contraction in value-added terms (index 2015=100) by more than 11 percent in the second quarter of 2020 (Canton et al. 2021, pp. 4). In the US—the venue of focus for this paper due to the availability of publicly accessible data and abundance of existing literature—COVID-induced shocks started in the first quarter of 2020 and reached -6.6 percent in terms of real GDP growth rate; the shocks then worsened in magnitude to -34.3 percent in the subsequent quarter (Bekaert et al. 2020, pp. 13-14).

As opposed to a typical recession where various sectors respond differently based on their sensitivity to business cycles, the pandemic has heterogeneous effects upon different sectors, contingent upon their dependence on in-person contacts. Contact-intensive sectors have suffered disproportionately in both the initial shock and recovery, whereas the effect can be mitigated in sectors where contactless interactions, such as teleworking, are ready alternatives (Canton et al. 2021, pp. 4). In the case of the US, a working paper with the NBER by Chetty et al. (2022, pp. 10-11) constructed a publicly available, granular-level dataset that includes aggregated, anonymized data on credit and debit card transactions of total US card spending through Affinity Solutions Inc. The paper highlights that health concerns towards oneself and others, rather than a loss of current or expected disposable income, drove US consumers' spending reductions—57 percent of which comes from reduced expenditure on economic activities that necessitate in-person contact (Ibid., pp. 18-19). Rather than being predictable by the income elasticities of demand, the changes in US consumer expenditure across various categories of goods and services align with their implied health risks. As such, the pandemic has altered both the level and composition of US consumer consumption, and the research question that naturally follows this change is: To what extent could consumers' COVID-related health concerns contribute to structural changes in their demand, substituting away from in-person services towards remote services, on a state level, in the US?

While the existing consumer-level data and macroeconomic indicators both empirically corroborate the

*I would like to thank my advisor, Professor Bedoor AlShebli, for insightful and productive weekly meetings, constant feedback and suggestions that challenge me to explore new directions, and help and support throughout the entire year. I would also like to thank my convener, Professor Samreen Malik, for structured guidelines throughout the entire progress, as well as for encouraging me to push further in my model construction and analyses.

existence of a recession driven by a decrease in US consumer demand, identifying the driving factors behind this change has important implications for various economic agents in the economy. For firms, a better understanding of the consumer's decision problem throughout various stages of the consumption journey can help them better predict consumer behavior, thereby obtaining comparative advantages (Safara 2022, pp. 1526). On the other hand, for government and sociopolitical institutions, identifying consumers' confidence or expectations as an essential source contributing to changes in aggregate demand could help with constructing more appropriate policies, which are necessary for post-COVID economic recovery and lasting changes such as the green and digital transitions (Erik et al. 2021, pp. 6-8).

Addressing the research question, the central hypothesis is that due to COVID-related health concerns, consumers substitute away from in-person services towards remote services. Such a hypothesized substitution effect is likely manifested through a decrease in consumer spending in in-person service categories and a corresponding increase in remote ones. As an initial step, separate multiple linear regression (MLR) ordinary least squares (OLS) models are used to explain variations in either in-person or remote expenditure, using a set of regressors that include survey-based, state-level weighted aggregates gauging consumers' health concerns, entity- and time-fixed effects, as well as other controls following literature. Then, a dummy variable accounting for the event of the aforementioned substitution effect occurring is constructed to associate both remote and in-person expenditures (%Δ) in the same regression. A linear probability model (LPM), using the same set of regressors as the MLR OLS models, is implemented to explain such a binary outcome variable. Whereas neither MLR OLS model is robust to the inclusion of a more comprehensive set of survey-based independent (X) variables, the estimated coefficient on the X variable avoid_contact in the LPM remains robust to the exclusion of less strong explanatory X variables. Further principal component analysis (PCA) and random forest classification (RFC) findings corroborate the feature importance of survey-based X variables and justify the inclusion of other controls, thereby lending support to the plausibility of the LPM specification (structural validity) and the central hypothesis.

The rest of this paper is structured in the following way. Section II summarizes the motivations drawn from existing literature, as well as the study's key contributions to it. Section III describes the main dependent (Y) variables, independent (X) variables, and other controls. It then outlines both the initial MLR OLS models and the subsequent construction of a binary dependent variable and the corresponding explanatory LPM. Section IV presents the results, discussions, and robustness checks for the aforementioned model specifications. Section V lists potential further research directions. Section VI concludes the paper.

II. Literature Review

The first strand of literature establishes the consumer demand side of the economy as a self-contained system, the outcome of which initiates subsequent impacts on macroeconomic activities. Starting from general considerations of consumer theories, when facing exogenous shocks (such as the pandemic) that exacerbate uncertainty, consumers would first substitute away from immediate consumption in the current period and move toward saving to buffer against unknown future periods. This is empirically corroborated by a study that uses US household-level Chase bank account data and establishes that aggregate consumer consumption expenditure fell by over 35 percent in March 2020, following the declaration of a national emergency in the US (Cox et al. 2020, pp. 37). Such reductions in consumer spending would then have downstream effects on other agents and institutions in the economy. In a study that explores nonparametric, non-Gaussian features of macroeconomic forecast revisions to identify the sources of the pandemic's shock on inflation and real GDP growth, Bekaert et al. (2020, pp. 13) attributed about two-thirds [1] of the initial pandemic-induced decline in US economic activities to negative aggregate demand shocks.

A secondary order of negative shocks may also arguably emerge, as recessions are typically associated with increased unemployment and subsequent decrease in consumers' disposable income. However, in the case of the US, on a state level, this downward-spiral mechanism was effectively mitigated by the government's fiscal responses shortly after the onset of the pandemic. With liquidity injection to households and firms, arrangement of automatic stabilizers such as unemployment benefits, and fiscal stimuli from the Treasury, some might even argue that relatively low-income households' disposable income has actually increased (Chetty et al. 2022, pp. 41). As the consumers' purchasing power remains de facto nearly unchanged, the US case constitutes an empirically observed Slutsky compensated price change. This helps attenuate any second-order negative effect on expenditure reduction from

[1] -4.3 percent out of a total of -6.6 percent.

loss of current or expected income, and theoretically resolve endogeneity concerns from other macroeconomic entities beyond the scope of this study. Together, the two strands of literature cited above thereby justify the study's unique focus on the microeconomic side of the US economy throughout the pandemic.

The last strand of literature suggests that in some cases, consumers' health concerns affect (changes in) their demand through a different channel than that of actual, realized COVID incidence. On an aggregate, macroeconomic level, consumers' anticipation and confidence play an important role in economic recovery—such that optimistic estimates may provide underlying conditions conducive to a more rapid recovery than average (Bekaert et al. 2020, pp. 15). Other organizations such as the OECD have also highlighted factors such as consumer confidence, perceived health risks, government containment measures and policy support, and the adaptability of firms in readjusting and meeting demands, in determining the pace of recovery for various economies (OECD, 2020). In the case of China, using comprehensive consumer offline transactions captured by UnionPay POS machines and QR scanners around the lockdown of Wuhan in 2020, Chen et al. (2021, pp. 308) empirically establish that consumers' willingness to consume is "independent of the effect of supply disruptions or negative income shock". My own study thus seeks to explore similar relationships in the context of the US, on the state level.

The paper contributes to the existing body of literature in the following ways. Firstly, to the best of the author's knowledge, the study will be the first of its kind to incorporate a broad spectrum of publicly available datasets, merge or aggregate (using sample weights provided) such data on the US state level, as the baseline for correlation inferences between US consumers' health concerns and changes in their expenditure patterns. Many of the existing studies cited above make use of privileged data with very limited accessibility and replicability. The granularity of the data used here, publicly accessible through datasets hosted via independent institutions such as GitHub, Google, and the US CDC, helps ensure the robustness of statistical inferences made in this study and the ease of replication or extension studies. The choice of using such data further calls upon relevant international entities to make anonymized, aggregate-level data more readily available and easily accessible, for the purpose of student research or big data practices.

The paper also organically combines various methods across the existing body of literature—notably applying the coexistence of respondent-level surveys and aggregate-level transaction data currently used to estimate marginal propensities to consume (MPCs) in the context of substitution effects in consumer demand (Baker et al. 2020, pp. 3-4). While consumer health-related anticipations can be proxied through nationally representative surveys (Ibid.), transaction data can also capture the in-flow funds of financial institutions. Therefore, though the focal scope of this study is microeconomic in nature, it still indirectly accounts for underlying macroeconomic institutions and structures.

Last but not least, to the best of the author's knowledge, this study is also the first of its kind to empirically establish that US consumers' COVID-related health concerns operate on a unique channel than other confounding explanatory factors (see section III.I.3 below). It also newly proposes a probabilistic event of the theorized substitution effect occurring and finds its statistically significant association with measures of consumer's health concerns. Distinguishing health concerns from actual (realized) COVID disease incidence is essential, as those disparate explanatory factors could entail very different policy implications.

III. Data and Methodology

I. Data

I.1 Main Explained (Y) Variables

The dataset for the explained (Y) variables originates from the main paper by Chetty et al. (2022, pp. 10-11), who constructed publicly available panel data series, down to the ZIP code level, on a daily basis from 2020-01-13 to 2021-12-10 (698 days). The specific dataset containing aggregated, anonymized US consumer spending data was collected by the company Affinity Solutions Inc., and is representative of total US consumer card spending (Ibid.). Due to the susceptibility of high-frequency transactional data to noises and cyclic fluctuations, Chetty et al. (2022, pp. 9) have already cleaned the data using standard measures such as imposing continuity where large or discrete jumps are found, smoothing extrema fluctuations through the implementation of 7-day moving averages, as well as deseasonalizing the series by normalization. Additionally, with considerations of confidentiality, the units are reported as percentage changes (%Δ) relative to the variables' mean values in January 2020 as opposed to the actual levels (Ibid.). The aforementioned public dataset is published on GitHub by Opportunity Insights at `https://github.com/OpportunityInsights/EconomicTracker`.

The data for US consumer spending (%Δ) is available at the city, county, and state levels. However, for the city level, only the total aggregate spending is reported,

and breakdowns of spending by categories (types of goods and services) are not available. For the county-level data, the number of distinct observations over time is only 53, which is nearly the same as that for the state-level [2] data. With further consideration that the most disaggregated level for many of the control variables would be available only on the state level, the entity index, i, is set on the state level, for all $i \in$ {50 states & Washington D.C.}.

Within the state-level dataset of credit/debit card transaction data collected by Affinity Solutions Inc., the two variables of interest are naturally spend_remoteservices, which measures consumer spending in remote services[3], and spend_inperson, which measures consumer spending in in-person services[4]. These two broad categories constitute an intermediate level of aggregation, as determined by the company Affinity Solutions Inc., and are also justified in the main paper's descriptive statistics and visualizations of the changes in US consumer card expenditures across sectors (Chetty et al. 2022, pp. 18).

As the outcome variables of interest for the initial separate regression models are spend_remoteservices and spend_inperson, it would be instructive to first obtain a visual representation of the density distribution for both of them, overlaid against each other, as illustrated in Figure 1 below:

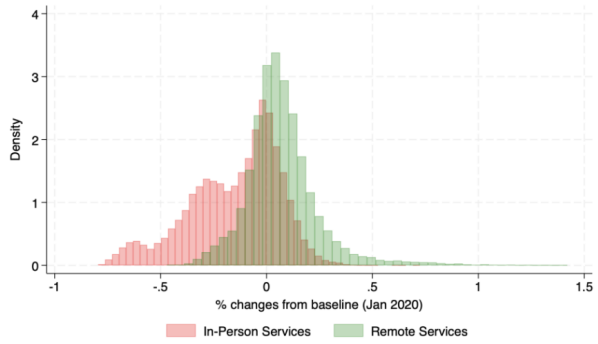


Figure 1: Density Plot for In-Person vs. Remote Services, % Changes from Jan 2020

In Figure 1 above, descriptively, one can identify a shift in distribution from spend_inperson to spend_remoteservices, on a state level, in the US, such that the mean of the former (colored in red) is negative and that of the latter (colored in green) is positive, and the distribution of the former appears more spread-out (with a larger standard deviation) than the latter. These visual observations can be corroborated by the summary statistics for both aforementioned variables as reported in Table 1 below:

Variable Name	count	mean	std	min	25%	50%	75%	max
spend_inperson	34900	-0.1603	0.1973	-0.787	-0.286	-0.129	-0.0093	0.284
spend_remoteservices	34900	0.0823	0.1367	-0.465	-0.0076	0.0743	0.17	0.934

Table 1: Summary Statistics for In-Person and Remote Services, % Changes from Jan 2020

I.2 Main Explanatory (X) Variables and Processing

The dataset for the key explanatory (X) variables of interest is also publicly available at `https://github.com/YouGov-Data/covid-19-tracker` (Jones 2020), and comes from nationally representative surveys of various countries (including the US) about COVID-related symptoms, COVID testing, self-isolation, social distancing, and behaviors. The surveys were conducted by Imperial College London, in conjunction with YouGov; the data are anonymized but available on the respondent level—each row of the raw dataset accounts for one individual who responded to the survey at a given point in time. A sample weight variable is also included in the dataset (under the variable label *weight*), based on the age, gender, and region of each respondent. This approach of merging results from nationally representative surveys that gauge consumers' anticipations and concerns to explain outcomes in the format of transaction data follows the precedence of Baker et al. (2020, pp. 5). This paper extends the original study's method from the context of estimating consumers' marginal propensity to consume (MPC) to inferences regarding the impacts of their COVID-related health concerns on expenditure.

Based on the discussion of relevant literature in section II above, Table 2 below lists several key survey question variables of interest, where those under the section "Starting X-vars of interest" constitute the initial set of survey-based X variables of focus, to be denoted as $S^1_{it} := \{\text{wear\_mask}, \text{avoid\_work}, \text{avoid\_shop}, \text{eat\_sep}\}$ as the main explanatory variables of the initial MLR OLS models. Those under the other section, "Additionally possible X-vars", characterize a more comprehensive additional set of survey-based consumer behavioral variables that measure individual health concerns to-

[2] A total of 51 entities that include the 50 US states and Washington D.C.

[3] This includes the following merchant category codes (MCCs): ADM administrative and support and waste management and remediation services; EDU education; FIN finance and insurance; INF information; PST professional, scientific, and technical; PUB public administration; and UCM utilities, construction, and manufacturing.

[4] This includes the following MCCs: ACF accommodation and food services; HCS healthcare and social assistance; AER arts, entertainment, and recreation; TWS transportation and warehousing; REN rental and leasing; REP repair and maintenance; and PLS personal and laundry services.

ward self and others, to be denoted as

$$S^2_{it} := \{\text{will\_isolate}, \text{sanitizer}, \text{avoid\_contact}, \text{avoid\_out}, \text{avoid\_guest}, \text{avoid\_small}, \text{avoid\_mid}, \text{avoid\_large}\}$$

so that in the subsequent robustness checks for the initial model estimations, the set of all main explanatory variables is $S_{it} := S^1_{it} \cup S^2_{it}$.

Explanatory Variables

Survey-based behavior variables as measures of individual consumers' health concerns toward self and others.

Starting X-vars of interest:

- `i12_health_1` (renamed `wear_mask`): Worn a face mask outside your home (e.g. when on public transport, going to a supermarket, going to a main road)
- `i12_health_9` (renamed `avoid_work`): Avoided working outside your home
- `i12_health_16` (renamed `avoid_shop`): Avoided going to shops
- `i12_health_18` (renamed `eat_sep`): Eaten separately at home, when normally you would eat a meal with others

Additionaly possible X-vars (that gauge health *concerns* - vs. realized health/symptoms or mobility changes):

- `i9_health` (renamed `will_isolate`): Thinking about the next 7 days... would you isolate yourself after feeling unwell or having any of the following new symptoms: a dry cough, fever, loss of sense of smell, loss of sense of taste, shortness of breath or difficulty breathing?
  - NOTE: response ∈ { 1 - Yes, 2 - No, 99 - Not sure }
- `i12_health_3` (renamed `sanitizer`): Used hand sanitiser
- `i12_health_5` (renamed `avoid_contact`): Avoided contact with people who have symptoms or you think may have been exposed to the coronavirus
- `i12_health_6` (renamed `avoid_out`): Avoided going out in general
- `i12_health_11` (renamed `avoid_guest`): Avoided having guests to your home
- `i12_health_12` (renamed `avoid_small`): Avoided small social gatherings (not more than 2 people)
- `i12_health_13` (renamed `avoid_mid`): Avoided medium-sized social gatherings (between 3 and 10 people)
- `i12_health_14` (renamed `avoid_large`): Avoided large-sized social gatherings (more than 10 people)

Table 2: Sets of Survey-Based X Variables of Interest

Note that the first labels with numbers are the variable labels as coded by the original survey and dataset, and the labels inside the brackets afterward are new, mnemonic names to facilitate interpretation. Note also that all variables listed in Table 2 above are categorical variables that take on a finite set of string values with their corresponding numeric values: for all $response_{it} \in S_{it} \setminus \{\text{will\_isolate}\}$, $response_{it} \in \{$"Always", "Frequently", "Sometimes", "Rarely", "Not at all"$\}$, on a corresponding Likert scale of 1-5 by survey construction. For the variable will_isolate, $response_{it} \in \{$"Yes", "No", "Not sure"$\}$, with the corresponding numeric values of 1, 2, and 99. Nonetheless, as the responses associated with will_isolate correspond to a binary variable, the mapping is further linearly transformed for this study, such that the corresponding value for "Yes" is 1, that for "No" is 0, and that for "Not sure" is NumPy.NaN (in Python), for ease of operation and interpretation. See Table 3 below for the summary statistics of all aforementioned survey-based X variables, after each $response_{it}$ has been converted to its corresponding numeric value as detailed above. The first four rows describe variables in the set S^1_{it}, the subsequent eight rows describe those in the set S^2_{it}, and the last row summarizes the aforementioned sample weight variable weight.

Variable Name	count	mean	std	min	25%	50%	75%	max
wear_mask	33940	2.1972	1.5443	1	1	1	3	5
avoid_work	15320	3.1199	1.6905	1	1	3	5	5
avoid_shop	33940	2.6902	1.4458	1	1	2	4	5
eat_sep	16982	3.4855	1.6749	1	2	4	5	5
will_isolate	28957	0.7311	0.4434	0	0	1	1	1
sanitizer	33940	2.1702	1.3168	1	1	2	3	5
avoid_contact	33940	1.8704	1.3886	1	1	1	2	5
avoid_out	33940	2.7435	1.4251	1	2	2	4	5
avoid_guest	33940	2.3815	1.5145	1	1	2	3	5
avoid_small	33940	2.5689	1.5494	1	1	2	4	5
avoid_mid	33940	2.3503	1.5224	1	1	2	3	5
avoid_large	33940	2.0799	1.4923	1	1	1	3	5
weight	33940	1	0.5592	0.0694	0.8037	0.9393	1.0836	17.4288

Table 3: Summary Statistics for All Survey-Based X Variables (in S_{it}) and *weight*

Among the variables on the 1-5 Likert scale (the set $S_{it} \setminus \{\text{will\_isloate}\}$), eat_sep has both the highest mean and median (50% percentile) values, whereas avoid_contact has the lowest mean, and both wear_mask and avoid_large share the lowest median along with avoid_contact. For the single variable will_isolate that is mapped into a binary variable, the mean value is 0.7311, which implies that 73.11% of all respondents in the survey would isolate themselves in the next seven days, after feeling unwell or experiencing the new symptoms as listed in the corresponding survey question.

As the area of the study is US states, it naturally follows that these variables can be further aggregated into state-level variables by: 1) mapping each response from string format into integer numeric format, so that the variables are now ordinal categorical variables; 2) aggregating by summing the integer-valued responses of individuals who answered at a given time t, in a given state i, weighted by the aforementioned sample weight variable under the label weight (so that the state-level aggregates are not biased). This way, each survey question listed in Table 2 and the sets defined above are bijectively mapped to one state-level aggregate ordinal categorical variable. Let us overload the notation S_{it} so that it also denotes the vector(s) consisting of all such state-level aggregates.

Nonetheless, due to the use of the sample weight variable, weight, to adjust for population representativeness, each outcome state-level aggregate is no longer on the same Likert scale as previously in the individual responses to the survey questions. Rather, the numeric values, on the adjusted scale, have no intrinsic meaning, and any final interpretation must re-adjust and account for those weighted scales. Hence, after arriving at a final model that is robust and ready for conclusive interpretations, we will need to normalize all explanatory variables so that the coefficients can be readily inter-

[5]The most up-to-date data tracker can be accessed at `https://covid.cdc.gov/COVID-data-tracker/#cases_newcaserateper100k`

preted in terms of changes in standard deviations. The summary statistics for all survey-based state-level aggregate ordinal categorical variables are reported in the respective columns in Table A.1 of Appendix A.

I.3 Other Control Variables and Merging

The first set of additional control variables are COVID incidence data reported by the US CDC[5], which include the numbers of total cases (tot_cases), new cases (new_cases), total deaths (tot_deaths), and new deaths (new_deaths), readily available on the US state level (CDC, 2023). The incorporation of COVID incidence controls into regression models follows the approach of Chen et al. (2021) and Carvalho et al. (2021) and accounts for the part of variations in the outcome expenditure variables uniquely explained by the actual realization of COVID cases (% Δ). Therefore the source of confounding effects upon the estimated relationship between the outcome and consumers' COVID-related concerns is mitigated. For the purpose of this study, the publicly available historical records dataset is used[6]

The second set of additional control variables comes from the Google COVID-19 Community Mobility Reports[7], which document changes in activity trends over time, is aggregated at various geopolitical levels, and is also categorized under different types of locations (Google LLC, 2023). Accounting for any substitution effects due to changes only in consumers' physical activity routines, the dataset reports percentage changes in US state-level aggregate mobility trends from the baseline[8], for locations under the categories of retail & recreation (retail_recreation), grocery & pharmacy (grocery_pharmacy), parks (parks), transit stations (transit_stations), workplaces (workplaces), and residential (residential) (Ibid.). Existing literature uses the Mobility Reports as a proxy for the amount of time that consumers spend outdoors (Chetty et al. 2022, pp. 19), and found significance in the estimated coefficient of (private) bank transaction data on workplaces and transit_stations in the case of Spain (Carvalho et al. 2021, pp. 10).

The third and final set of additional controls includes dummy variables constructed as indicators for the implementation of state-level COVID-related policies. Following Carvalho et al. (2021, pp. 2-4), policies that involve opening and closing establishments are incorporated as the policy controls of particular focus. The cross-sectional dataset for such policies comes from the COVID-19 US State Policy (CUSP) database hosted by Boston University (Raifman et al. 2020). The US state-level data under the "Closures & Reopening" tab is used, and corresponding binary variables are constructed for each policy. For each state, the row observations with dates before the implementation of the policy in that corresponding state are assigned a value of 0. Those with dates on and after are assigned a value of 1. If there is no available information regarding the date of policy implementation for that state, then *NumPy.NaN* (in Python) is used as an indicator for missing data.

The summary statistics for all aforementioned additional control variables are reported respectively throughout Tables A.2-A.4 in Appendix A. Finally, the main data frame, containing both main Y (detailed in section III.I.1 above) and X (detailed in section III.I.2 above) variables, is left-merged with all these datasets of additional control variables, using the double indices of t (*datestamp*) and *i* (*state*) that together characterize the panel data structure.

II. Initial MLR OLS Models

As an initial step, two separate multiple linear regression (MLR) ordinary least squares (OLS) models are used to explain the two outcome variables $Y_{it} \in \{spend\_remoteservices_{it}, spend\_inperson_{it}\}$:

$$Y_{it} = \alpha_0 + \beta \cdot S_{it} + \gamma \cdot \text{COVID}_{it} + \lambda \cdot \text{Policies}_{it}$$

$$+\mu \cdot \text{Mobility}_{it} + X_i + T_t + \epsilon_{it}$$

Where i (entity index) $\in$ {50 states & Washington D.C.}, t (time index) is (on a daily basis) from 2020-01-13 to 2021-12-10; $Y_{it} \in \{\text{spend\_remoteservices}_{it}, \text{spend\_inperson}_{it}\}$ as defined in section III.I.1; S_{it} is the vector of survey-based, state-level weighted aggregates quantifying consumers' health concerns as defined in section III.I.2; COVID_{it} is the vector of COVID incidence controls, Policies_{it} is the vector of (state-level) COVID-related policy controls, Mobility_{it} is the vector of mobility controls from Google's COVID-19 Community Mobility Reports, all three of which are defined in section III.I.3; X_i and T_t are the entity (state) and time (day) fixed effects, respectively.

More specifically, the starting set of two initial separate MLR OLS models includes:

$$\text{spend\_remoteservices}_{it} = \alpha_0 + \beta \cdot S^1_{it} + \gamma \cdot \text{COVID}_{it} \quad (1.1a)$$

$$+\lambda \cdot \text{Policies}_{it} + \mu \cdot \text{Mobility}_{it} + X_i + T_t + \epsilon_{it} \quad (1.1b)$$

[6]The historical records dataset, on a weekly frequency, can be accessed and downloaded at `https://data.cdc.gov/Case-Surveillance/Weekly-United-States-COVID-19-Cases-and-Deaths-by-/pwn4-m3yp`

[7]Accessible at `https://www.google.com/covid19/mobility/`.

[8]The baseline is the median value, for the corresponding day of the week, during the 5-week period Jan 3–Feb 6, 2020.

And

$$\text{spend\_inperson}_{it} = \alpha_0 + \beta \cdot S^1_{it} + \gamma \cdot \text{COVID}_{it} \quad (1.2a)$$

$$+\lambda \cdot \text{Policies}_{it} + \mu \cdot \text{Mobility}_{it} + X_i + T_t + \epsilon_{it} \quad (1.2b)$$

Whereas the set of two separate MLR OLS models for robustness checking includes:

$$\text{spend\_remoteservices}_{it} = \alpha_0 + \beta \cdot S_{it} \quad (2.1a)$$

$$+\gamma \cdot \text{COVID}_{it} \quad (2.1b)$$

$$+\lambda \cdot \text{Policies}_{it} + \mu \cdot \text{Mobility}_{it} + X_i + T_t + \epsilon_{it} \quad (2.1c)$$

And

$$\text{spend\_inperson}_{it} = \alpha_0 + \beta \cdot S_{it} + \gamma \cdot \text{COVID}_{it} \quad (2.2a)$$

$$+\lambda \cdot \text{Policies}_{it} + \mu \cdot \text{Mobility}_{it} + X_i + T_t + \epsilon_{it} \quad (2.2b)$$

Where, as already defined in section III.I.2 above, S^1_{it} is the set of the main state-level aggregate explanatory X variables for the initial MLR OLS models, and $S_{it} = S^1_{it} \cup S^2_{it}$ is the set of all (state-level aggregate) explanatory variables for robustness checks of those two separate initial MLR OLS models.

The choice of the MLR OLS model follows Carvalho et al. (2021, pp. 6), where an outcome variable in the form of a first-order difference[9] can be explained using COVID-related explanatory variables such as lockdown policy dummy variables and daily COVID incidence. The key independent variables consist of nationally representative, survey-based consumer behavioral variables that measure individual health concerns toward self and others. This follows from Baker et al. (2020, pp. 3-4), which proxies consumer expectations[10] through nationally representative surveys and combines such data with transaction records to derive more precise estimates of the marginal propensity to consume (MPC). The inclusion of fixed effects follows from Cox et al. (2020, pp. 51), where entity- (each state, thus accounting for the location) and time-fixed effects both incorporate the unobservable or unmeasurable variables and account for heterogeneity[11] by income or location.

III. Dummy Y Variable and LPM

An important limitation to the two separate MLR OLS models outlined in section II above—either the set of initial models (1.1) and (1.2) or the subsequent sets of robustness checks (2.1) and (2.2)—is that the two models are separate and not directly comparable or related. This is a significant shortcoming that needs to be addressed, as the two separate models may only each address one of two separate and independent hypotheses:

H1a: a decrease in in-person expenditures (%Δ)

H1b: an increase in remote expenditures (%Δ)

And cannot answer the (one) central research hypothesis regarding the substitution effect:

H1: a decrease in in-person expenditures (%Δ) and a corresponding increase in remote expenditures (%Δ)

⇔ a substitution away from in-person and towards remote expenditure (%Δ)

It is hence instructive to construct a variable that can relate both remote and in-person expenditures (%Δ) in the same regression. Let us denote Y1:=Δ spend_remoteservices, and Y2:=Δ spend_inperson. The central hypothesis theorizes a substitution away from in-person (Y2) and towards remote (Y1) expenditure (%Δ) , which implies a higher Δ spend_remoteservices (Y1) than Δ spend_inperson (Y2). And on relative terms, this exactly corresponds to the probability event (Y1 > Y2) ⇔ (Y1 - Y2 > 0).

Thus, on the basis of the existing two Y variables (*spend_remoteservices* and *spend_inperson*), we construct a new binary dependent Y variable — a new column variable with the same t and i indices and thus the same panel data structure — $\textit{Remote\_g\_InPerson}_{it}$, that is equal to 1 if $Y1 > Y2$ (i.e., Δspend_remoteservices > Δspend_inperson) and 0 otherwise. Please refer to Appendix B for a more detailed proof of why this construction uniquely corresponds to the Research Question (RQ) and central hypothesis. As all other components of the complete data frame remain unchanged, this new Y variable $\textit{Remote\_g\_InPerson}_{it}$, is explained using the same sets of X variables (III.I.2) and additional controls (III.I.3), and can be characterized by:

$$\text{Remote\_g\_InPerson}_{it} = \alpha_0 + \beta \cdot S_{it} + \gamma \cdot \text{COVID}_{it} + \lambda \cdot$$

$$\text{Policies}_{it} + \mu \cdot \text{Mobility}_{it} + X_i + T_t + \epsilon_{it} \quad (3)$$

Since the outcome variable, $\textit{Remote\_g\_InPerson}_{it}$, is a binary variable, the updated regression model as outlined in equation (3) above is effectively a linear probability model (LPM), where the estimated coefficients on the X and control variables account for marginal changes in the probability of the substitution event $\textit{Remote\_g\_InPerson}_{it}$ occurring.

[9] Y-o-Y growth rate in the original paper; percentage changes from baseline in the context of this study.

[10] Regarding unemployment, salary cuts, tax increases, benefit cuts, stock market performance, and the duration of the pandemic (Baker et al. 2020, pp. 3-4).

[11] Which are likely fixed for each state (entity) for the duration of this study.

IV. Results and Discussion

I. Separate MLR OLS Models and Robustness

We start with discussions on estimates of the two separate MLR OLS models as characterized in section III.II above. For the first round of initial results, only the restricted subset of survey-based variables, S_{it}^1, is used. As illustrated in the left panel[12] of Table 4 below, even after controlling for all vectors of confounders as suggested by existing literature, the estimated coefficient of $\texttt{spend\_remoteservices}_{\texttt{it}}$ on $\texttt{wear\_mask}_{\texttt{it}}$ remains statistically significant at the 10% level ($p = 0.057$). On the other hand, for the other separate model explaining $\texttt{spend\_inperson}_{\texttt{it}}$, after accounting for $\texttt{COVID}_{\texttt{it}}$, $\texttt{Policies}_{\texttt{it}}$, and $\texttt{Mobility}_{\texttt{it}}$ (as well as i and t fixed effects), the consumer health concern-related variables are no longer significantly explanatory of variations in the outcome, as illustrated in the right panel[13] of Table 4 below. The R^2_{adj} is expectedly quite high for both model estimates ($R^2_{adj} = 0.847$ for (1.1) and $R^2_{adj} = 0.966$ for (1.2)) due to the inclusion of both time- and entity-fixed effects.

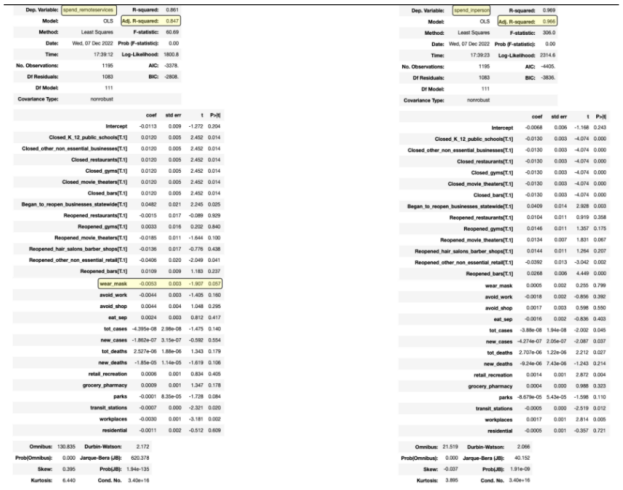


Table 4: Baseline Regression Results for Separate MLR OLS Models
(All Controls Added; *i* and *t* F.E. Included)

Nonetheless, for the model explaining *spend_inperson*$_{it}$ (1.2), the estimated coefficients on other control variables corroborate the findings from existing literature cited above. The estimated coefficients on all closing-related policy dummy variables are negative and significant; those on opening dummies are positive, but few are significant—except for other *non-essential* retail with a negative and significant coefficient. The estimated coefficients on COVID cases and rates are negative and significant, but interestingly, that of the total number of deaths is positive. Lastly, for the Google Mobility Reports, the estimated coefficients on *retail & recreation* and *workplaces* are positive and significant; those on *parks* and *transit* stations are negative and significant. The significance of the estimated coefficient on *workplaces* and *transit* stations previously established in the case of Spain by Carvalho et al. (2021, pp. 10) thus also applies in this context.

The empirical findings above then lend support to the plausibility of the model specifications, and we are ready to proceed by first examining the robustness of the estimated coefficient of *spend_remoteservices*$_{it}$ on *wear_mask*$_{it}$. Theoretically, prominent sources of endogeneity are already mitigated to the best extent possible and reasonable. Second-order negative effects on expenditure reduction from the loss of current or expected income are already mitigated by the US fiscal responses to the pandemic (Chetty et al. 2022, pp. 41; Chen et al. 2021, pp. 308). Potential omitted variable bias (OVB) by income quartile and populations, a source of concern identified by studies like Cox et al. (2020, pp. 51), is not applicable in the context of this study. As the timespan is relatively short (less than two years), and the level is on US states and not counties or ZIP codes—such sources of OVB or heterogeneity are likely fixed for each state for the duration of the study, and thus absorbed by the state fixed effects X_i. Thus, the focus would mainly be on expanding the set of survey-based X variables, from S_{it}^1 to S_{it}, to ensure its representativeness as a measure of state-level, aggregate consumer concerns.

However, as illustrated in Table 5 in the appendix[14], upon the inclusion of the entire set of all possible survey-based variables of interest—characterized by the vector S_{it} as defined in the last paragraph of section III.I.3—the estimated coefficients of either Y variable on the survey-based variables become not significant at any commonly used level. In other words, neither initial MLR OLS model is robust to the inclusion of additional survey-based X variables. Based on such empirical findings, the LPM characterized in section III.II above then becomes necessary to address the RQ and central hypothesis.

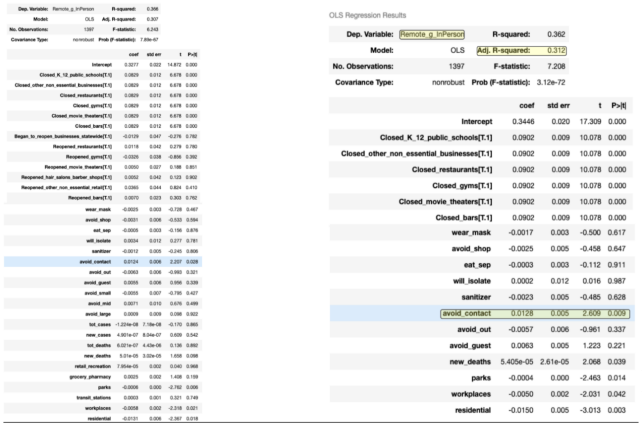
[12] Estimates for the model in equation (1.1).

[13] Estimates for the model in equation (1.2).

[14] The left panel corresponds to the estimates for the model in equation (2.1); the right panel corresponds to the estimates for the model in equation (2.2).

II. Parsimonious, Normalized LPM

As illustrated in the left panel of Table 6 below, the estimates of $Remote\_g\_InPerson_{it}$ on the full set of regressors—corresponding to the model outlined in equation (3)—appear to already lend support to the hypothesis, as the estimated coefficient on *avoid_contact* is positive and significant at the 5% level. We then proceed to derive a parsimonious LPM model to explain $Remote\_g\_InPerson_{it}$, starting with the extended set of survey-based independent (X) variables, S_{it}. Less strong explanatory survey-based X variables, as characterized firstly by the p-value and then by the magnitude of the estimated coefficient, are excluded from the model. In the end, we derive a parsimonious LPM, where the estimated coefficient on *avoid_contact* remains essentially unchanged (from 0.0124 in the full model to 0.0128 in the parsimonious LPM) and still significant at the 5% significance level.



OLS Regression Results

Dep. Variable:	Remote_g_InPerson	R-squared:	0.362
Model:	OLS	Adj. R-squared:	0.312
No. Observations:	1397	F-statistic:	7.208
Covariance Type:	nonrobust	Prob (F-statistic):	3.12e-72

	coef	std err	t	P>|t|
Intercept	0.3446	0.020	17.309	0.000
Closed_K_12_public_schools[T.1]	0.0902	0.009	10.078	0.000
Closed_other_non_essential_businesses[T.1]	0.0902	0.009	10.078	0.000
Closed_restaurants[T.1]	0.0902	0.009	10.078	0.000
Closed_gyms[T.1]	0.0902	0.009	10.078	0.000
Closed_movie_theaters[T.1]	0.0902	0.009	10.078	0.000
Closed_bars[T.1]	0.0902	0.009	10.078	0.000
wear_mask	-0.0017	0.003	-0.500	0.617
avoid_shop	-0.0025	0.005	-0.458	0.647
eat_sep	-0.0003	0.003	-0.112	0.911
will_isolate	0.0002	0.012	0.016	0.987
sanitizer	-0.0023	0.005	-0.485	0.628
avoid_contact	0.0128	0.005	2.609	0.009
avoid_out	-0.0057	0.006	-0.961	0.337
avoid_guest	0.0063	0.005	1.223	0.221
new_deaths	5.405e-05	2.61e-05	2.068	0.039
parks	-0.0004	0.000	-2.463	0.014
workplaces	-0.0050	0.002	-2.031	0.042
residential	-0.0150	0.005	-3.013	0.003

Table 6: LPM with Full Set of Regressors (left) and Parsimonious LPM (right)

The R^2_{adj} measure actually slightly increased following this process, as it penalizes for the overfitting of too many independent variables that are not contributing to the explanatory power of the model. This process of deriving the parsimonious LPM, as illustrated in the right panel of Table 6 above, characterizes the first step of the robustness checks for the estimated coefficient on *avoid_contact*—robust to the exclusion of less strongly explanatory independent variables.

Referring to previous discussions in section III.I.2 regarding re-adjusting the sample-weighted scale for the state-level aggregate X variables (as also illustrated in Table A.1 of Appendix A), and for more meaningful and straightforward interpretation, we further normalize all variables [15] involved in the LPM for statistical inferences in terms of standard deviation. The regression table for the normalized, parsimonious LPM is reported in Table 7 below; *ceteris paribus*, on average, on the US state level, 1 standard deviation increase in consumers' frequency of avoiding contact with people who have symptoms or whom they believe may have been exposed to the coronavirus, is associated with a 0.1371 standard deviation increase in the probability of the substitution event as theorized in the central hypothesis occurring.

Normalized LPM Model for Interpretation

Dep. Variable:	Remote_g_InPerson	R-squared:	0.362
Model:	OLS	Adj. R-squared:	0.312
No. Observations:	1397	F-statistic:	7.208
Covariance Type:	nonrobust	Prob (F-statistic):	3.12e-72

	coef	std err	t	P>|t|
Intercept	0.3578	0.037	9.567	0.000
Closed_K_12_public_schools[T.1]	0.0921	0.006	16.080	0.000
Closed_other_non_essential_businesses[T.1]	0.0921	0.006	16.080	0.000
Closed_restaurants[T.1]	0.0921	0.006	16.080	0.000
Closed_gyms[T.1]	0.0921	0.006	16.080	0.000
Closed_movie_theaters[T.1]	0.0921	0.006	16.080	0.000
Closed_bars[T.1]	0.0921	0.006	16.080	0.000
wear_mask	-0.0228	0.046	-0.500	0.617
avoid_shop	-0.0392	0.086	-0.458	0.647
eat_sep	-0.0052	0.046	-0.112	0.911
will_isolate	0.0003	0.020	0.016	0.987
sanitizer	-0.0293	0.061	-0.485	0.628
avoid_contact	0.1371	0.053	2.609	0.009
avoid_out	-0.0902	0.094	-0.961	0.337
avoid_guest	0.0881	0.072	1.223	0.221
new_deaths	0.0160	0.008	2.068	0.039
parks	-0.0311	0.013	-2.463	0.014
workplaces	-0.0690	0.034	-2.031	0.042
residential	-0.0837	0.028	-3.013	0.003

Table 7: Normalized Parsimonious LPM

III. Additional Robustness Checks for LPM

In the previous process of deriving the parsimonious LPM, the estimated coefficient of $Remote\_g\_InPerson_{it}$ on *avoid_contact* is both established as statistically significant for the comprehensive set of all pertinent state-level survey-based X variables gauging US consumer's health concerns and robust to the exclusion of less strongly explanatory X variables. For subsequent robustness checks, we will borrow concepts in machine learning to further corroborate the plausibility of the LPM model design, thereby inferring structural validity, following Lu and White (2014).

III.1 Principal Component Analysis (PCA)

The method of principal component analysis (PCA) is especially applicable in large datasets with many regressors, as it implements dimension reduction upon the dataset such that the resulting principal components (PC) correspond to coordinates of an orthogonal linear transformation of the original data (Jolliffe and Cadima

[15] Without normalizing dummy variables

2016, pp. 2). While this approach helps preserve the most meaningful and strong explanatory variations in the set of regressors, the process of projection implies that we cannot know exactly where each PC comes from, but can only infer based on how strongly each PC is correlated with each explanatory (X) variable.

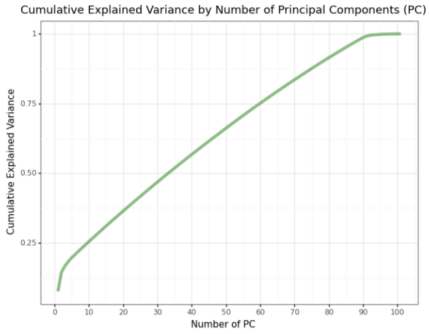


Figure 2: Cumulative Explained Variance by Number of Principal Components (PC)

In the normalized LPM (the estimates for which are reported in Table 7 above), there are a total of 767 explanatory variables, including time- and entity-fixed effects. As illustrated in Figure 2 above, dimension reduction is only meaningful until 101 PCs—then no reduction can be implemented and all variations are preserved; the cumulative explained variance increases most rapidly for the first 2 PCs, and then almost linearly in the number of PCs. For the purposes of this exercise, we will focus on the first 3 PCs.

As illustrated in the three respective panels in Table 8 below, the first PC (7.65% variance explained; leftmost panel) is very strongly correlated with survey-based variables which include *avoid_contact*; then it is also correlated with both time- and entity-fixed effects. The second PC (6.78% variance explained; middle panel) is very strongly correlated with time-fixed effects (especially 2020-03-05), Google Mobility Report controls, survey-based X variables, and entity-fixed effects. Lastly, the third PC (2.23% variance explained; rightmost panel) is correlated with controls from Google Mobility Report and COVID incidence, and both time- and entity-fixed effects. From the fourth PC onwards, the variance explained is less than 2% and linearly decreases in the number of PCs as aforementioned in Figure 2 above and thus is not reported in the table.

	X Variable	Correlation With PC1
0	avoid_out	0.964362
1	eat_sep	0.963971
2	sanitizer	0.959938
3	avoid_shop	0.957760
4	will_isolate	0.954199
5	avoid_guest	0.954064
6	avoid_contact	0.938058
7	wear_mask	0.922581
8	residential	0.520344
9	datestamp_2020-04-16 00:00:00	0.390467
10	datestamp_2020-08-20 00:00:00	0.208887
11	datestamp_2020-04-09 00:00:00	0.197184
12	datestamp_2020-04-30 00:00:00	0.192835
13	datestamp_2020-06-25 00:00:00	0.181248
14	datestamp_2020-07-23 00:00:00	0.172286
15	datestamp_2020-05-28 00:00:00	0.147747
16	state_Georgia	0.137188
17	state_Florida	0.133867
18	datestamp_2020-04-23 00:00:00	0.133071
19	datestamp_2020-06-04 00:00:00	0.124861

	X Variable	Correlation With PC2
0	datestamp_2020-03-05 00:00:00	0.992337
1	workplaces	0.539353
2	avoid_contact	0.130015
3	avoid_shop	0.111260
4	wear_mask	0.107331
5	avoid_out	0.105974
6	sanitizer	0.097366
7	will_isolate	0.089322
8	avoid_guest	0.089146
9	eat_sep	0.081652
10	state_New York	0.068511
11	state_Texas	0.044385
12	state_Illinois	0.035540
13	state_Arkansas	0.033007
14	state_Indiana	0.030626
15	state_Tennessee	0.030171
16	state_Michigan	0.026949
17	state_Oklahoma	0.025865
18	state_South Carolina	0.024689
19	state_Wisconsin	0.024181

	X Variable	Correlation With PC3
0	residential	0.641671
1	new_deaths	0.441626
2	datestamp_2020-04-23 00:00:00	0.266528
3	datestamp_2020-05-07 00:00:00	0.245880
4	datestamp_2020-04-30 00:00:00	0.222668
5	state_Hawaii	0.208223
6	datestamp_2020-04-09 00:00:00	0.176726
7	datestamp_2021-02-04 00:00:00	0.176611
8	state_Massachusetts	0.161651
9	state_New Jersey	0.152674
10	state_Nevada	0.133558
11	state_Texas	0.127975
12	datestamp_2020-05-14 00:00:00	0.115869
13	state_California	0.113411
14	state_Florida	0.111012
15	state_Connecticut	0.106211
16	datestamp_2021-03-04 00:00:00	0.098500
17	state_Maryland	0.091816
18	datestamp_2020-03-05 00:00:00	0.089718
19	state_Washington	0.088299

Table 8: The Correlation Between Each PC and Various X Variables

III.2 Feature Importance Metric Using Tree-Based Classifier

Considering that the explained variable $Remote\_g\_InPerson_{it}$ is a binary variable, an alternative method to directly rank the importance of the current set of explanatory variables in the LPM is using a tree-based classifier such as the random forest classifier in Python's scikit-learn library (Pedregosa et al. 2011). To mitigate the issue of overfitting inherent in decision trees, a train-test split is first implemented, with the test size equal to 25% of the data, and a random state of 0 (for replicability purposes). As n_estimators, the parameter for the number of decision trees in the random forest classifier is a hyperparameter, we first determine its optimal value using cross-validation with 5 groups. Subsequently, we instantiate the optimal classifier with the derived optimal hyperparameter, fit the train data, and rank the feature importance score for each X variable through the *feature_importances*_property that reports the Gini importance (Ibid.).

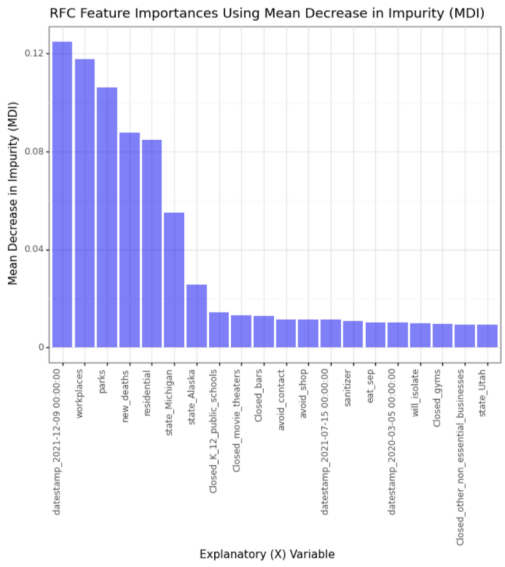


Figure 3: RFC Feature Importances Using Mean Decrease in Impurity (MDI)

As illustrated in Figure 3 above, in this specific instantiation of the random forest classifier (with optimal hyperparameter *n_estimators* = 137), and with the specific aforementioned random state and test size of the train-test split, the additional control variables—firstly those from the Google Mobility Report, followed by the CDC COVID incidence data and finally closing-related policies—as well as the time- and entity-fixed effects have the strongest feature (Gini) importance. Nonetheless, *avoid_contact* still ranks as the most important explanatory variable out of all survey-based X variables.

IV. Discussions and Limitations

With the construction of the binary variable $Remote\_g\_InPerson_{it}$ and the corresponding LPM that directly address the single central research hypothesis, we start from a comprehensive set of state-level aggregate survey-based explanatory variables that gauge US consumers' concerns and arrive at a parsimonious LPM, normalized for better interpretability (the estimates for which are reported in Table 7 above). The estimates suggest that (*ceteris paribus*, on average, on the US state level), a greater extent of consumer health concerns as measured through *avoid_contact* is associated with a greater probability of the substitution event occurring as theorized in the central hypothesis. The normalized LPM is robust to the exclusion of less strongly explanatory X variables, and further PCA and RFC analyses corroborate the importance of survey-based X variables and justify the inclusion of other controls.

An important note on the robustness checks undertaken above is that such exercises are not intended to establish internal validity (a causal relationship between consumers' health concerns and substitution/changes in their expenditure). The most prominent hindrance along the way would be that neither the dates related to the onset or spread of the COVID-19 pandemic, nor the relevant nation-state-level policies, have sources of variations that are as good as random and resemble those from the gold standard of causal inferences as in randomized controlled trials (RCT). The inherent susceptibility of survey-based data to additional unobservable endogeneity issues such as reporting bias—especially when the pertinent questions are related to the daily life and health of individuals—further weakens the grounds for causal inferences. Nonetheless, following Lu and White (2014), the additional robustness checks in section IV.III above help establish the plausibility of the model specifications in this study and lend support to its structural validity.

Another important source of limitation is that the intermediate level of spending category aggregation is exogenously determined by Affinity Solutions Inc. As the author does not have access to the actual expenditure *level* data, it is impractical to construct alternatives, as the sum of first-order differences is not permutable with the first-order difference of alternative aggregates which must first be adjusted for representativeness and noise reduction, thus necessitating the original level data for any alternative formulations. Similarly, though the Revealed Preferences approach in microeconomics (WARP and SARP) may help with testing for the possible duration of the theorized substitution event through non-parametric testing, such tests would require matrixes of expenditure level data and price levels, which are infeasible to the scope of this capstone study.

Lastly, the current survey data published by the main NBER paper through Opportunity Insights on GitHub covers a relatively short period of time—less than two years, immediately before and during the pandemic. To the best of the author's knowledge, though there exist other nationally representative surveys that gauge US consumers' health-related concerns—such as USC's Understanding America Study[16]—such other surveys started only after the onset of the pandemic or the declaration of the national emergency and cannot constitute potential alternative measures of S^1_{it} or S_{it}, as estimates or inferences regarding the substitution event require data points both before and after the pandemic. More post-pandemic data points would also be necessary for the generalizability of the association inferences based on the parsimonious LPM in the end.

V. Further Research Directions

A first strand of possible research extension would be contingent upon increased data availability as time passes post-pandemic, such that the different points in time for inferences of substitution effects could be pre-, during, or post-pandemic. Further extension studies could be conducted as more post-pandemic data points become available, for generalizability and inferences regarding whether the estimated relationship between consumer health concerns and the substitution event persists over time.

Additionally, in the medium-to-long term, the pandemic may interact with ongoing trends such as digitalization—effects such as a stronger uptake/demand of digital technologies, adaptation to remote working methods, and stronger network effects associated with digital technologies are likely to stay. A different interesting area of study could be regarding such poten-

[16] "Understanding Coronavirus in America tracking survey" at `https://covid19pulse.usc.edu/`.

tial relationships between the accelerated digitalization trends and consumers' adoption of online formats of final purchases.

Lastly, access to expenditure level data may allow non-parametric testing using WARP and SARP, which can gauge the potential duration of the substitution effect. Such access–either with better accessibility to the spending level data or in cases of privileged non-disclosure uses—could also provide ground for the construction of alternative aggregation of spending categories, as further robustness checks for inferences in this study.

VI. Conclusion

Following the *atypical* recession brought about by the COVID-19 pandemic, this paper investigates the extent to which consumers' COVID-related health concerns help contribute to substitution effects in their demand away from in-person services and towards remote services, on a state level, in the US. Distinguishing health concerns from alternative potential contributing factors, such as actual COVID disease incidence, changes in physical mobility trends, and state-level COVID policies is essential, as those disparate explanatory factors could entail very different policy implications for governments, or insights into consumers' decision journey for firms that seek to establish comparative advantage.

Whereas neither separate initial MLR OLS model is robust to the inclusion of a more comprehensive set of survey-based independent (X) variables, the estimated coefficient of a newly constructed binary variable, $Remote\_g\_InPerson_{it}$, on the X variable *avoid_contact* in the LPM remains robust to the exclusion of less strongly explanatory X variables. Further principal component analysis (PCA) and random forest classification (RFC) analyses corroborate the feature importance of survey-based X variables and justify the inclusion of other controls, thereby ascertaining the plausibility of the LPM specification (structural validity). The eventual parsimonious LPM then provides evidence supporting the central hypothesis that (*ceteris paribus*, on average, on the US state level) a greater extent of consumer health concerns, as manifested through more cautious daily activities such as *avoid_contact*, is associated with a greater likelihood of the substitution event (away from in-person and towards remote expenditures (%Δ)) occurring.

This result highlights the unique role of consumers' behavioral traits in determining (changes in) their demand. It suggests that policies supporting post-COVID recovery and development should address and alleviate consumers' health-related concerns. Firms may also need to beware of any changes in the composition of consumer demand due to the substitution event occurring and adjust their strategies accordingly for profit maximization. More generally, this paper emphasizes the importance of incorporating multiple sources of granular, publicly available datasets to corroborate statistical evidence for identifying contributing factors and micro-level mechanisms behind macro-level socioeconomic phenomena.

References

Baker, Scott and Farrokhnia, R.A., (2020), "Income, Liquidity, and the Consumption Response to the 2020 Economic Stimulus Payments" No 2020-55, Working Papers *Becker Friedman Institute for Research In Economics* https://EconPapers.repec.org/RePEc:bfi:wpaper:2020-55.

Carvalho VM, Garcia JR, Hansen S, Ortiz Á, Rodrigo T, Rodríguez Mora JV, Ruiz P. (2021) "Tracking the COVID-19 crisis with high-resolution transaction data" *R. Soc. Open Sci. 8: 210218.*: https://doi.org/10.1098/rsos.210218

CDC (Centers for Disease Control and Prevention) " COVID Data Tracker. Atlanta, GA: US Department of Health and Human Services, CDC; 2023, May 01" https://covid.cdc.gov/covid-data-tracker

Chen, Haiqiang, Wenlan Qian, and Qiang Wen. 2021. "The Impact of the COVID-19 Pandemic on Consumption: Learning from High-Frequency Transaction Data" *BAEA Papers and Proceedings, 111: 307-11.*

COX, NATALIE, PETER GANONG, PASCAL NOEL, JOSEPH VAVRA, ARLENE WONG, DIANA FARRELL, FIONA GREIG, and ERICA DEADMAN. "Initial Impacts of the Pandemic on Consumer Behavior: Evidence from Linked Income, Spending, and Savings Data." *Brookings Papers on Economic Activity, 2020, 35–69.* https://www.jstor.org/stable/26996635.

"The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data" by Raj Chetty, John Friedman, Nathaniel Hendren, Michael Stepner, and the Opportunity Insights Team. November 2020. *Available at:* https://opportunityinsights.org/wp-content/uploads/2020/05/tracker$_p$*aper.pdf*

Erik Canton Federica Colasanti Jorge Durán Maria Garrone Alexandr Hobza Wouter Simons Anneleen Vandeplas, 2021. "The Sectoral Impact of the COVID-19

Crisis. An Unprecedented and Atypical Crisis" *European Economy - Economic Briefs 069*, Directorate General Economic and Financial Affairs (DG ECFIN), European Commission.

Geert Bekaert Eric C. Engstrom Andrey Ermolov, 2020. "Aggregate Demand and Aggregate Supply Effects of COVID-19: A Real-time Analysis" *Finance and Economics Discussion Series 2020-049*, Board of Governors of the Federal Reserve System (U.S.).

Google LLC "Google COVID-19 Community Mobility Reports "https://www.google.com/covid19/mobility/. Accessed: May 01, 2023.

Jolliffe Ian T. and Cadima Jorge. 2016 "Principal component analysis: a review and recent developments" *Phil. Trans. R. Soc. A*.374:20150202. https://doi.org/10.1098/rsta.2015.0202 Rutgers University.

Jones, Sarah P., Imperial College London Big Data Analytical Unit and YouGov Plc. 2020, Imperial College London YouGov Covid Data Hub, v1.0 *YouGov Plc*, April 2020. https://github.com/YouGov-Data/covid-19-tracker

OECD (2020). "Rebuilding tourism for the future: COVID-19 policy responses and recovery"

Raifman J, Nocka K, Jones D, Bor J, Lipson S, Jay J, and Chan P. (2020). "COVID-19 US state policy database." *Available at:* : www.tinyurl.com/statepolicies

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

Safara, F. A Computational Model to Predict Consumer Behaviour During COVID-19 Pandemic. *Comput Econ 59, 1525–1538 (2022)*: https://doi.org/10.1007/s10614-020-10069-3

Xun Lu, Halbert White," Robustness checks and robustness tests in applied economics" *Journal of Econometrics,* Volume 178, Part 1, 2014, Pages 194-206, ISSN 0304-4076, https://doi.org/10.1016/j.jeconom.2013.08.016.

V. Appendix A - Summary Statistics of Other Variables

Variable Name	count	mean	std	min	25%	50%	75%	max
wear_mask	14841	5.0023	13.5296	0	0	0	4.3044	286.8511
avoid_work	9444	0.6348	2.1629	0	0	0	0	52.165
avoid_shop	14841	6.1229	15.4427	0	0	0	5.4276	297.9017
eat_sep	11499	5.1078	14.6073	0	0	0	3.1457	271.6788
sanitizer	14841	4.9452	12.3682	0	0	0	4.4446	227.8711
avoid_contact	14841	4.2566	10.5633	0	0	0	3.9392	188.3273
avoid_out	14841	6.2516	15.6456	0	0	0	5.5745	284.8597
avoid_guest	14841	5.4272	13.8275	0	0	0	4.8579	264.4322
avoid_small	14841	5.847	14.707	0	0	0	5.203	266.327
avoid_mid	14841	5.3544	13.6458	0	0	0	4.7971	254.0655
avoid_large	14841	4.7406	12.2125	0	0	0	4.3218	227.9036
will_isolate	12206	0.668	1.6987	0	0	0	0.7513	27.439

Table A.1: Summary Statistics for All State-Level Aggregated Survey-Based X Variables (in S_{it})

Variable Name	count	mean	std	min	25%	50%	75%	max
tot_cases	8784	648861.8963	1174698.811	0	14381	206000.5	806024.25	11329293
new_cases	8784	11050.0179	28589.3391	-16945	574	3289	10894.25	790954
tot_deaths	8784	9171.5903	14540.4096	0	270.75	2979	12154	95810
new_deaths	8784	121.2071	265.2755	-3450	4	38	122	4915

Table A.2: Summary Statistics for US CDC COVID Incidence Control Variables ($COVID_{it}$)

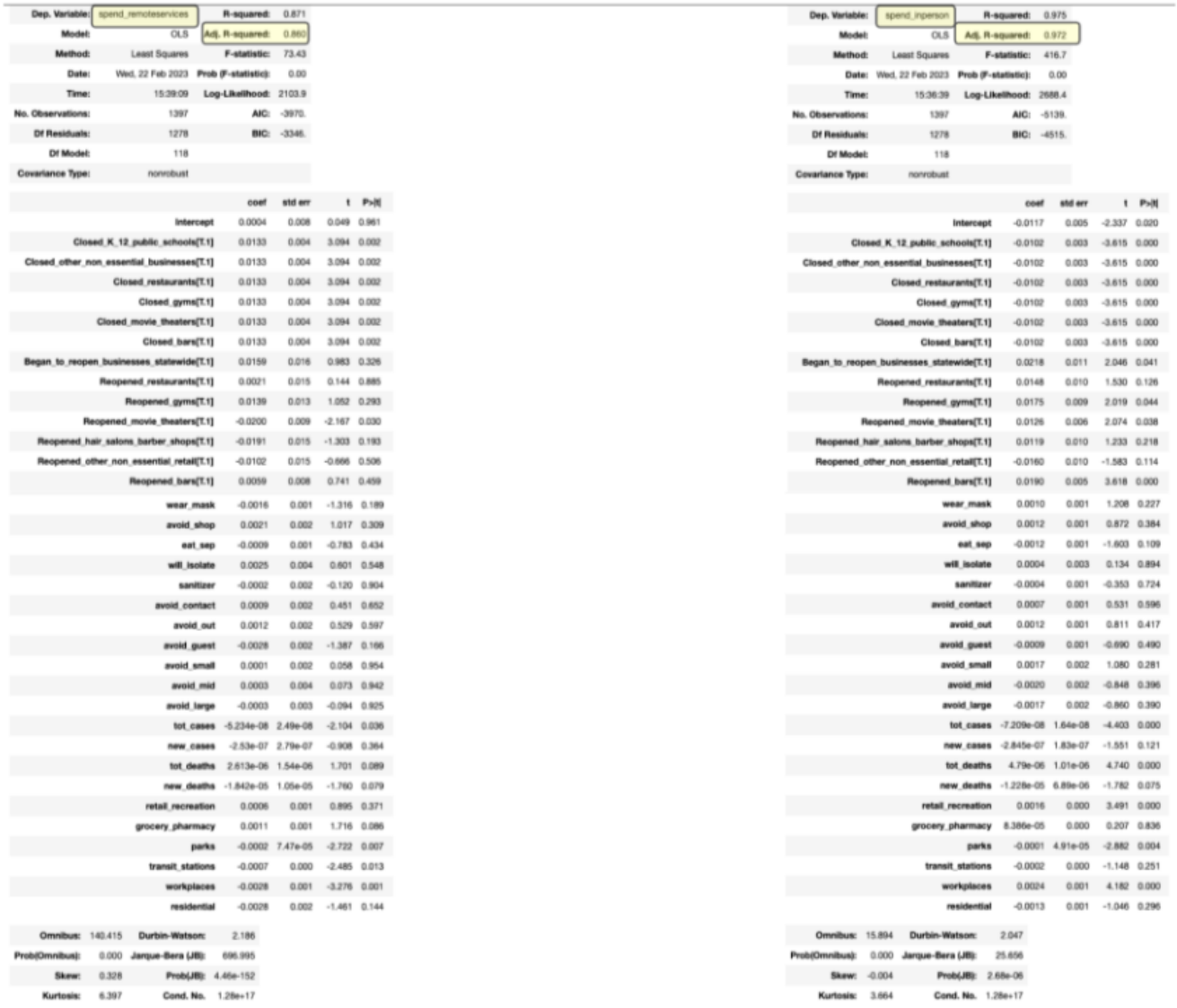
Variable Name	count	mean	std	min	25%	50%	75%	max

retail_recreation	34986	-10.3902	16.4132	-92	-18	-8	0	54
grocery_pharmacy	34986	0.2676	12.4822	-80	-6	0	7	73
parks	34484	50.1239	76.0034	-77	-4	28	84	636
transit_stations	34851	-15.5538	24.4127	-89	-33	-16	1	107
workplaces	34986	-25.9935	14.521	-88	-34	-26	-16	18
residential	34986	7.1189	5.6846	-8	3	6	10	36

Table A.3: Summary Statistics for Google Mobility Reports Control Variables ($Mobility_{it}$)

Variable Name	count	mean	std	min	25%	50%	75%	max
Closed_K_12_public_schools	34202	0.906	0.2918	0	1	1	1	1
Closed_other_non_essential_businesses	34202	0.8947	0.3069	0	1	1	1	1
Closed_restaurants	34202	0.9037	0.295	0	1	1	1	1
Closed_gyms	34202	0.9	0.3	0	1	1	1	1
Closed_movie_theaters	34202	0.8995	0.3006	0	1	1	1	1
Closed_bars	34202	0.9049	0.2934	0	1	1	1	1
Began_to_reopen_businesses_statewide	34202	0.8319	0.374	0	1	1	1	1
Reopened_restaurants	34202	0.8177	0.3861	0	1	1	1	1
Reopened_gyms	34202	0.7855	0.4105	0	1	1	1	1
Reopened_movie_theaters	34202	0.745	0.4359	0	0	1	1	1
Reopened_hair_salons_barber_shops	34202	0.8131	0.3898	0	1	1	1	1
Reopened_other_non_essential_retail	34202	0.8248	0.3802	0	1	1	1	1
Reopened_bars	34202	0.7314	0.4432	0	0	1	1	1

Table A.4: Summary Statistics for State-Level COVID Policy Control Variables ($Policies_{it}$)



Dep. Variable:	spend_remoteservices	R-squared:	0.871
Model:	OLS	Adj. R-squared:	0.860
Method:	Least Squares	F-statistic:	73.43
Date:	Wed, 22 Feb 2023	Prob (F-statistic):	0.00
Time:	15:39:09	Log-Likelihood:	2103.9
No. Observations:	1397	AIC:	-3970.
Df Residuals:	1278	BIC:	-3346.
Df Model:	118		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|
Intercept	0.0004	0.008	0.049	0.961
Closed_K_12_public_schools[T.1]	0.0133	0.004	3.094	0.002
Closed_other_non_essential_businesses[T.1]	0.0133	0.004	3.094	0.002
Closed_restaurants[T.1]	0.0133	0.004	3.094	0.002
Closed_gyms[T.1]	0.0133	0.004	3.094	0.002
Closed_movie_theaters[T.1]	0.0133	0.004	3.094	0.002
Closed_bars[T.1]	0.0133	0.004	3.094	0.002
Began_to_reopen_businesses_statewide[T.1]	0.0159	0.016	0.983	0.326
Reopened_restaurants[T.1]	0.0021	0.015	0.144	0.885
Reopened_gyms[T.1]	0.0139	0.013	1.052	0.293
Reopened_movie_theaters[T.1]	-0.0200	0.009	-2.167	0.030
Reopened_hair_salons_barber_shops[T.1]	-0.0191	0.015	-1.303	0.193
Reopened_other_non_essential_retail[T.1]	-0.0102	0.015	-0.666	0.506
Reopened_bars[T.1]	0.0059	0.008	0.741	0.459
wear_mask	-0.0016	0.001	-1.316	0.189
avoid_shop	0.0021	0.002	1.017	0.309
eat_sep	-0.0009	0.001	-0.783	0.434
will_isolate	0.0025	0.004	0.601	0.548
sanitizer	-0.0002	0.002	-0.120	0.904
avoid_contact	0.0009	0.002	0.451	0.652
avoid_out	0.0012	0.002	0.529	0.597
avoid_guest	-0.0028	0.002	-1.387	0.166
avoid_small	0.0001	0.002	0.058	0.954
avoid_mid	0.0003	0.004	0.073	0.942
avoid_large	-0.0003	0.003	-0.094	0.925
tot_cases	-5.234e-08	2.49e-08	-2.104	0.036
new_cases	-2.53e-07	2.79e-07	-0.908	0.364
tot_deaths	2.613e-06	1.54e-06	1.701	0.089
new_deaths	-1.842e-05	1.05e-05	-1.760	0.079
retail_recreation	0.0006	0.001	0.895	0.371
grocery_pharmacy	0.0011	0.001	1.716	0.086
parks	-0.0002	7.47e-05	-2.722	0.007
transit_stations	-0.0007	0.000	-2.485	0.013
workplaces	-0.0028	0.001	-3.276	0.001
residential	-0.0028	0.002	-1.461	0.144

Omnibus:	140.415	Durbin-Watson:	2.186
Prob(Omnibus):	0.000	Jarque-Bera (JB):	696.995
Skew:	0.328	Prob(JB):	4.46e-152
Kurtosis:	6.397	Cond. No.	1.28e+17

Dep. Variable:	spend_inperson	R-squared:	0.975
Model:	OLS	Adj. R-squared:	0.972
Method:	Least Squares	F-statistic:	416.7
Date:	Wed, 22 Feb 2023	Prob (F-statistic):	0.00
Time:	15:36:39	Log-Likelihood:	2688.4
No. Observations:	1397	AIC:	-5139.
Df Residuals:	1278	BIC:	-4515.
Df Model:	118		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|
Intercept	-0.0117	0.005	-2.337	0.020
Closed_K_12_public_schools[T.1]	-0.0102	0.003	-3.615	0.000
Closed_other_non_essential_businesses[T.1]	-0.0102	0.003	-3.615	0.000
Closed_restaurants[T.1]	-0.0102	0.003	-3.615	0.000
Closed_gyms[T.1]	-0.0102	0.003	-3.615	0.000
Closed_movie_theaters[T.1]	-0.0102	0.003	-3.615	0.000
Closed_bars[T.1]	-0.0102	0.003	-3.615	0.000
Began_to_reopen_businesses_statewide[T.1]	0.0218	0.011	2.046	0.041
Reopened_restaurants[T.1]	0.0148	0.010	1.530	0.126
Reopened_gyms[T.1]	0.0175	0.009	2.019	0.044
Reopened_movie_theaters[T.1]	0.0126	0.006	2.074	0.038
Reopened_hair_salons_barber_shops[T.1]	0.0119	0.010	1.233	0.218
Reopened_other_non_essential_retail[T.1]	-0.0160	0.010	-1.583	0.114
Reopened_bars[T.1]	0.0190	0.005	3.618	0.000
wear_mask	0.0010	0.001	1.208	0.227
avoid_shop	0.0012	0.001	0.872	0.384
eat_sep	-0.0012	0.001	-1.603	0.109
will_isolate	0.0004	0.003	0.134	0.894
sanitizer	-0.0004	0.001	-0.353	0.724
avoid_contact	0.0007	0.001	0.531	0.596
avoid_out	0.0012	0.001	0.811	0.417
avoid_guest	-0.0009	0.001	-0.690	0.490
avoid_small	0.0017	0.002	1.080	0.281
avoid_mid	-0.0020	0.002	-0.848	0.396
avoid_large	-0.0017	0.002	-0.860	0.390
tot_cases	-7.209e-08	1.64e-08	-4.403	0.000
new_cases	-2.845e-07	1.83e-07	-1.551	0.121
tot_deaths	4.79e-06	1.01e-06	4.740	0.000
new_deaths	-1.228e-05	6.89e-06	-1.782	0.075
retail_recreation	0.0016	0.000	3.491	0.000
grocery_pharmacy	8.386e-05	0.000	0.207	0.836
parks	-0.0001	4.91e-05	-2.882	0.004
transit_stations	-0.0002	0.000	-1.148	0.251
workplaces	0.0024	0.001	4.182	0.000
residential	-0.0013	0.001	-1.046	0.296

Omnibus:	15.894	Durbin-Watson:	2.047
Prob(Omnibus):	0.000	Jarque-Bera (JB):	25.656
Skew:	-0.004	Prob(JB):	2.68e-06
Kurtosis:	3.664	Cond. No.	1.28e+17

Table 5: Robustness Checks for Separate Baseline MLR OLS Models (All Controls Added; *i* and *t* F.E. Included)

VI. Appendix B - The Construction of $Remote\_g\_InPerson_{it}$

For simplicity of notation, let us define and denote $Y_1 :=$ Δspend_remoteservices and $Y_2 :=$ Δspend_inperson. The central research hypothesis theorizes a substitution away from in-person (Y_2) and towards remote (Y_1) expenditure (% changes), which would imply a higher Δspend_remoteservices(Y_1) than Δspend_inperson(Y_2). And on relative terms, this is equivalent to the event $Y_1 > Y_2 \Leftrightarrow Y_1 - Y_2 > 0$.

There may exist other possibilities/interpretations, but they do not apply as accurately in the context of this study. For instance, the event that $Y_1 > 0 \wedge Y_2 < 0$ is a subset of the event $Y_1 > Y_2$. But the former imposes additional restrictions that ignore, for instance, cases where both are positive/negative but still one is greater than the other.

Another possible formulation could take into account both the sign and magnitude of the percentage changes and corresponds to the dummy variable that is equal to 1 if:

$$Y_1 > Y_2 > 0$$

$$Y_1 > 0 > Y_2$$

$$0 > Y_2 > Y_1$$

For a concrete instance, this would mean that the dummy variable value is 1 if $Y_1 = -5\%$ and $Y_2 = -2\%$. But in that case, the change in Y_2 is less negative than that in Y_1—i.e., Y_1 suffers a greater extent of decrease than Y_2, which would be indicative of substitution away from Y_1 toward Y_2, a contradiction to the RQ focus and hypothesis.

The "greater than" operator in all cases is applicable, contributing to a simple $Y_1 > Y_2 \Leftrightarrow Y_1 - Y_2 > 0$ LPM model. The dummy dependent variable is named *Remote_g_InPerson*, which is equal to 1 if $Y_1 > Y_2$ (i.e., Δspend_remoteservices > Δspend_inperson) and 0 otherwise.