Semiparametric estimation of the relationship between recessions and health

Marta Boczoń*[1]

[1]*Department of Economics, University of Pittsburgh, Pittsburgh, United States*

July 25, 2020

Abstract

I identify and address several shortcomings of the existing and widely applied approach in studying the relationship between business cycles and the well-being of economic agents. In contrast to the existing literature, I allow for the relationship in question to be nonlinear by relying upon semiparametric estimation techniques. Moreover, I proxy for the state of the economy by analyzing both average economic conditions as well as the observed variability in growth cycles. While my initial results complement rather than contradict those in the literature, they provide a novel and much needed reconsideration of how to correctly analyze the relationship between economic recessions and/or expansions and health.

Key words: BMI, recession, business cycle, health economics
JEL classification: C14, E32, I12

1 Introduction

In this paper, I examine the relationship between the Body Mass Index (BMI), which is a continuous measure of weight standardized for height, and the unemployment rate. This research topic is not novel and thus, there already exists a comprehensive literature on this subject in both economics and medical science. However, in this work, I am the first one to conduct an in-depth analysis on the relationship in question. My research does not contradict existing results, but supplements it to the extent that it could initiate a new discussion on *how* and *why* the economy affects people's health-related outcomes.

*Marta Boczoń. Address: 230 South Bouquet Street, 4918A W. W. Posvar Hall, Pittsburgh PA, 15213; Tel.: +1(412)-577-8949, E-mail address: mjb249@pitt.edu.

In the existing literature on economic conditions and health-related outcomes, the mean unemployment rate is used as the main, and usually the only, proxy for economic conditions. An important question, however, is that of whether the mean unemployment rate captures the essence of varying economic conditions. In Figure 1, I plot four lines (black, blue, red, and green) that each represents a different hypothetical trajectory of the mean unemployment rate. Importantly, all unemployment profiles have the same mean (3.5%) computed over the period of eighteen months. First, observe that whereas the black line has the range (defined as the difference between maximum and minimum values) equal to zero, the red line has the range equal to 3.5%. Thus, despite having the same mean, the two unemployment profiles are fundamentally different. The black line indicates a stable economy, whereas the red line an economy subjected to macroeconomic fluctuations. Second, recognize important differences between unemployment profiles represented by the red and blue lines. Despite having the same mean and range, the two lines portray economies in opposite phases of the business cycle. The red line illustrates economic recession, whereas the blue line economic expansion. Finally, notice unemployment profiles depicted by the black and green lines. The two lines have the same mean and the same first difference (defined as the difference between the unemployment rate at time t and $t-18$). Yet, one corresponds to a stable economy, whereas the other one to an economy entering a recession. This discrepancy is well captured by the range variable which is substantially different for the two unemployment profiles. Now, let's consider a real life example using information on two respondents from the Behavioral Risk Factor Surveillance System (BRFSS) data set, which I use in my estimation. The first individual, say A, was a resident of Kentucky as of January 1987. The second individual, say B, was a resident of Alabama as of January 2010. At the time of the interviews, A and B were exposed to almost the same mean unemployment rates computed over the period of past eighteen months (9.43% and 9.69%, respectively). On the other hand, the range of the unemployment rate computed over the same period was dramatically different for the two respondents. A faced a range of state unemployment rate of 0.5, while that faced by B was 6.2, which is twelve times larger than that for the other individual. Consequently, A experienced a period of relative economic stability, while B faced a period of rapid and severe economic changes. Therefore, in this research, I proxy for economic conditions using not only the mean unemployment rate, but also its range. Moreover, I distinguish between periods of economic expansion and economic contraction. Thus, I allow the effect of the unemployment rate on the BMI to differ depending on whether an individual was subjected to increasing or decreasing unemployment over a given period of time.

The second important contribution of this paper is that I consider a possibility of a nonlinear relationship between the BMI and either the mean or the range of the unemployment rate. Specifically, I estimate a semiparametric partially linear model that allows for any functional form of the underlying relationship. Since I estimate a semiparametric (in contrast to a fully nonparametric) model I can still control for a broad range of variables such as demographic and socioeconomic characteristics, state fixed effects, and a liner time trend. Moreover, in my estimation, I account for sparse regions in the density of the unemployment rate by using a k-nearest neighbor estimator (as opposed to a kernel estimator).

My results of a fully parametric linear regression model indicate that the relationship between the BMI and mean unemployment rate is always negative and statistically significant. Specifically, a one percentage point *increase* in the mean unemployment rate, *decreases* BMI by 0.03, whereas in periods of declining unemployment, a one percentage point *decrease* in the mean unemployment rate, *increases* BMI by 0.01. Because the order of magnitude of these estimates as well as their sign and significance are in line with existing literature, a novelty of this finding lies in statistically significant differences between the two estimated coefficients. Thus, I find that the mean unemployment rate influences people's weight differently depending on a phase of the business cycle. This is a conjecture that so far has not been addressed in the related literature. My results of a semiparametric partially linear model indicate that, in periods of declining unemployment, the relationship between the range of the unemployment rate and BMI is highly nonlinear and consequently, can be severely underestimated by a linear regression model. Moreover, I find that economic agents largely ignore minor changes in unemployment when the economy is relatively tame and only react to economic fluctuations that exceed some threshold. Then, the range of the unemployment rate has an affect on the BMI that is at least twice as strong as that of the mean.

This research builds on a seminal paper by Ruhm (2000) "Are recessions good for your health?" In his study, Ruhm examines the relationship between a range of health outcomes (such as the BMI, tobacco use, alcohol consumption, physical activity) and the mean unemployment rate. He estimates a fully parametric linear regression model with individual-level controls as well as state and year fixed effects using the BRFSS data from 1987 to 1995. He finds that a one percentage point increase in the mean unemployment rate decreases BMI by 0.02 and concludes that recessions have beneficial effect of people's weight. In his other research, "Good times make you sick", Ruhm (2003) uses micro-level data from the 1972-1981 National Health Interview Surveys to examine the relationship between health status (such as number of sick days) and macroeconomic fluctuations. Using the same econometric specification and the mean unemployment rate as a proxy for economic conditions, he finds that most measures of health worsen when the economy strengthens. In neither of the two papers, however, he allows for the effect of the mean unemployment rate to vary with the business cycle. Moreover, he does not consider the possibility that the dispersion of the unemployment rate can have a more pronounced effect of health than the mean.

2 BMI

BMI is a proxy for body fat that has a broad range of advantages and disadvantages. Despite the controversy, it remains the most widely used measure for body fat. The data on BMI is readily available across time, regions, and population subgroups. It facilitates large-scale research on obesity and weight-related health risks. In particular, it allows me to study the relationship between economic conditions and body weight using data on more than 4.7 million individuals over the course of 27 years.

BMI, defined as the ratio of the body mass to the squared body height,

$$\mathrm{BMI} = \frac{\mathrm{mass}_{\mathrm{kg}}}{\mathrm{height}_{\mathrm{m}}^2} = \frac{\mathrm{mass}_{\mathrm{lb}}}{\mathrm{height}_{\mathrm{in}}^2} \times 703, \tag{1}$$

is simple to interpret and calculate. It is a straightforward function of only two variables, weight and height, that both can be routinely measured in an inexpensive and noninvasive way. Moreover, weight and height measurements do not require a trained personnel or a clinical setting and thus, can be self-reported. In particular, in my study, I use a self-reported data on weight and height collected in a phone survey.

One of the main disadvantages of BMI is that it does not differentiate between lean body tissue (such as organs, bones, muscles, tendons) and fat mass. Moreover, it does not discriminate between subcutaneous (mostly harmless fat located under the skin) and visceral (mostly harmful fat surrounding internal organs) types of fat. These two deficiencies may result in underestimating health risks of a normal-weight individual with excessive body fat. Second, BMI does not factor in underlying differences in body fat by age, sex, and race. On average, for the same BMI, older people have a higher percentage of body fat than younger people. Likewise, for an equivalent BMI, women tend to have more body fat than men and Asians more body fat than whites. Finally, BMI may not apply to athletes as well as pregnant and breastfeeding women. BMI is moderately correlated with other indirect measures of body fat such as waist circumference, underwater weighing, and bioelectrical impedance analysis. The latter two procedures require expensive equipments and highly-qualified medical personnel and thus, are usually performed in designated research facilities. Underwater weighing relies on Archimedes' principle, which states that an object immersed in a fluid is buoyed up by a force equal to the weight of the fluid displaced by the object. In particular, it exploits the fact that fat mass is less dense than water whereas lean tissue is denser. The other technique, bioelectrical impedance, relies on the differences in tissues with respect to resistance to flow of electrical current. In particular, it exploits the fact that tissues containing large amounts of fluid and electrolytes are highly conductive, whereas fat and bone impedes the electrical current.

Despite the moderate correlations of BMI with other proxies for body fat, BMI is strongly correlated with various health outcomes. First, a study by Renehan et al. (2008) on BMI and incidence of cancer published in Lancet, one of the oldest and most prestigious medical journals, shows that for men, a 5 unit increase in BMI is strongly associated with colon, kidney, thyroid, and esophageal cancers. Similarly, for women, a 5 unit increase in BMI is highly correlated with uterus, gallbladder, esophageal, and kidney cancers. Second, another Lancet publication by Prospective Studies Collaboration (2008) on BMI and cause-specific mortality indicates that a 5 unit increase in BMI is on average associated with 30% higher overall mortality, 40% for vascular mortality, 60-120% for diabetic, kidney- and liver-related mortality, 10% for cancer-related mortality and 20% for respiratory and all other mortality. Finally, Mokdad et al. (2003) in their research published in Journal of the American Medical Association show that, compared to normal-weight individuals, adults with BMI exceeding 40 have an odds ratio of 7.37 for diagnosed diabetes, 6.38 for high blood pressure, 1.88 for high cholesterol levels, 2.72 for asthma, 4.41 for arthritis, and 4.19 for fair or poor health.

3 BRFSS

In my study, I use micro-level data from the BRFSS. The BRFSS, which is the largest health-related survey system in the world, measures behavioral risk factors of the non-institutionalized adult population aged 18 or older residing in the U.S. In particular, the system collects uniform state-level data on tobacco and alcohol consumption, health care coverage, HIV prevention, chronic health conditions, and physical activity. The BRFSS is administrated by the Center for Disease Control and Prevention (CDC) and is supported by participating U.S. states and territories. The state health departments are involved in developing and updating the survey and are responsible for all field operations including conducting phone interviews. On the other hand, the CDC's Office of Surveillance, Epidemiology, and Laboratory Services weights and analyses the survey data. The BRFSS questionnaire consists of three parts, a compulsory core component containing demographic and basic health-related questions, optional modules dedicated to specific health conditions (such as cardiovascular disease, asthma, diabetes), and additional questions developed by participating states and hence, neither monitored nor processed by CDC.

The BRFSS survey was initiated in 1984 as a landline phone survey. Over the years, as the percentage of households without landline phones increased, the population coverage started to vary across states, geographic regions, and subpopulations. In particular, the percentage of cell phone only households exceeded 30% in the first half of 2011. Moreover, there was a substantial increase in the use of wireless communication for households using both cell and landline phone services. Therefore, in 2011, the BRFSS introduced a cellular phone survey aimed at alleviating the non-telephone coverage. In 2011, the target population for cellular phone samples consisted of households without landline phone but with a working cellular phone. By 2015, the target population encompassed all private residences and college housing with a working cell phone.

The data collected in the BRFSS landline phone survey are obtained through a complex sample design. Since 2001, all participating states use the Disproportionate Stratified Sampling (DSS) design. Phone numbers are divided into medium and high density strata, which are sampled separately. The high-density stratum contains more household phone numbers and thus, is sampled at a higher rate than the medium density stratum. A phone number belongs to the high or the medium density stratum depending on the number of listed household phone numbers in its one hundred blocks. Before 2001, states did not follow the same sampling technique. For instance, in 1996, fourteen states used the DSS design with low and high density strata, thirty one states used the Mitofsky-Waksburg sampling methodology, six states used other probability sampling methods and four used non-probability sampling designs that did not conform to the BRFSS standards. The most commonly used sampling method was Mitofsky-Waksburg three-stage cluster sampling procedure. In the first stage, phone numbers were grouped into one hundred blocks. Then, these primary sampling units (clusters) were randomly sampled, and from each sampled block, one phone number was drawn at random in the second stage. If a selected phone number belonged to a household, the entire cluster was accepted and sampled from until a target number of completed interviews was obtained. Finally, in the third step, one adult at least eighteen year old was randomly selected from a household to participate in an interview. Until 1995, all participating states constituted a single geographic stratum. In 1996, Rhode Island, Wisconsin, and Utah introduced geographic stratification to provide accurate sample sizes for smaller geographically defined populations. Over time, the number of states with more than one geographic stratum as well as the number of geographic strata within a state increased. For instance, in 2010, all but six participating states were geographically stratified resulting in a total of five hundred sixty three geographic strata across the U.S.

4 Data

I use the repeated cross-section BRFSS data from 1984 through 2010. Since I cannot directly compare the survey data before and after 2011 (due to the addition of cellular phones and to changes in the construction of sample weights discussed below), my analysis does not cover the period from 2011 to 2016. Moreover, in this study, I focus solely on the U.S. states and the District of Columbia and disregard the BRFSS data on the U.S. Territories (Guam, Puerto Rico, and Virgin Island).

In 1984, when the survey was initiated, only 15 states participated in the BRFSS (see Table 1: Column 1). Over time, the number of participating states increased. It doubled by 1987, tripled by 1990, and since 1996, all the U.S. states and the District of Columbia have been conducting the survey[1]. The number of respondents have been gradually increasing over the course of years (see Table 1: Column 2). In 1984, the BRFSS surveyed the total of 12,300 individuals. The number of participants exceeded 100,000 by 1993, 200,000 by 2001, and 300,000 by 2005. Since 2007, there are, on average, 425,000 respondents surveyed each year with a record high number of participants (445,000) interviewed in 2010. Since 1996, when for the first time all states participated in the program, the average number of respondents per state increased from 3,500 to 8,300 in 2010 (see Table 1: Column 3). In this study, I restrict my sample to individuals with non-missing values of weight and height and with BMI greater or equal to 12. It results in excluding on average 3.87% of all observations. In particular, I leave out the smallest fraction of observations (2.71%) in 1990 and the largest fraction (4.94%) in 2001 (see Table 1: Column 4). As a result, the total number of observations decreases by 4.33% from 4,970,945 to 4,755,645 (see Table 1: Column 5).

Questions on weight and height are standard queries that belong to the core part of the BRFSS questionnaire. Consequently, they are asked to respondents from all states across all years. In particular, weight and height are assessed by asking "About how much do you weigh without shoes?" and "About how tall are you without shoes?", respectively. For respondents interviewed prior to 1987, I calculate BMI based on reported weight (in pounds) and height (in feet and inches) using formula (1). For respondents interviewed between 1987 and 2010, I rely on BMI calculated and provided by BRFSS as an addition to the main data file. In my data set, BMI is rounded to one decimal point and ranges from 12 to 100. According to a CDC categorization, a

[1] In 2004, Hawaii did not participate in the BRFSS.

BMI below 18.5 indicates underweight, from 18.5 to 25.0 normal weight, from 25.0 to 30.0 overweight, and above 30.0 obesity. Moreover, a BMI above 30.0 can be further divided into obesity of class I (30.0-35.0), severe obesity of class II (35.0-40.0) and morbid obesity of class III (40.0 and more).

Nowadays, the United States faces a major obesity problem with one of the highest obesity rates worldwide and the highest among all 35 OECD member countries. Since the beginning of the 1980s, there has been a dramatic increase in the mean BMI of the adult U.S. population. Specifically, between 1984 and 2010, the mean BMI of adult U.S. residents increased by 18% from 23.4 to 27.6 (see Figure 2: Panel 1). Moreover, the mean BMI increased within all population subgroups defined by sex, age, race, ethnicity, marital status, employment status, and education level(see Figure 2: Panels 2-8). In particular, females, blacks, Hispanics, unmarried, as well as people aged 18-34, with some college, and employed for wages experienced the most pronounced increase in the mean BMI. From 1984 to 2010 the number of obese U.S. residents increased from 9% to 27% (see Figure 3). Moreover, whereas the number of overweight adults rose by almost ten percentage points, the number of people with normal weight declined from 56% to 32%. Consequently, as of 2010, there were more overweight people than those with normal weight and 60% of the adult U.S. population were either overweight or obese. On the other hand, there was a decline in the number of underweight adults from 3.5% to 1.5%. Note that unlike previous outcomes, a declining percentage of people who are underweight is desirable. Underweight, which is typically caused by illness, malnutrition, and eating disorder, poses significant health risks such as increased risk for complications from surgery and decreased function of immune system.

In my analysis, I control for a wide range of individual-level characteristics such as sex (male, female), age (18-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-69, 70-74, 75-79, 80 and more years old), race (white, black, other), ethnicity (Hispanic, non-

Hispanic), marital status (married, unmarried), education level (high-school dropout or less, high school graduate, some college, college graduate or more), and employment status (employed for wages, self-employed, either out of work or unable to work, homemaker, student, or retired). Along with queries about weight and height, demographic questions also belong to the core part of the BRFSS questionnaire and thus, are asked to all respondents across all years. Answers to all demographic questions, except for sex, have missing entries. On average, across all years, there are 0.3% missing observations on age, 0.6% on race, 0.3% on ethnicity, 0.2% on marital status, 0.2% on education level, and 0.3% on employment status (see Figure 4: Panels 1-7). Specifically, there are 92,500 individuals with a non-missing BMI, but at least one missing covariate. In order not to exclude these observations from my analysis, I introduce a range of dummy variables, which are equal to one if an information is missing and zero otherwise.

In addition to information on respondent's weight, height, demographic, and socioeconomic characteristics, the BRFSS provides information on respondent's state of residence as well as a year, month, and day of the BRFSS interview. This information is available for *all* interviewees participating in the BRFSS and there is no missing data in any of the aforementioned variables. In my analysis, I use this information together with monthly state-level unemployment rate data to proxy for macroeconomic conditions a respondent was subjected to prior to the BRFSS interview. The unemployment rate data comes from the U.S. Bureau of Labor Statistics. The data is seasonally adjusted, covers all 50 states and the District of Columbia, starts in January 1976, eight years before the initiation of the BRFSS, and is rounded to one decimal point. To proxy for respondent-specific macroeconomic conditions I construct three measures, which are functions of monthly unemployment rates in respondent's state of residence from up to P months prior to an interview. The first measure, $M_{ist}^{AV,P}$, which I refer to as *the mean*, is a simple non-weighted average of unemployment rates in P months preceding an interview,

$$M_{ist}^{AV,P} = P^{-1} \sum_{p=1}^{P} U_{s,t-p}, \tag{2}$$

where (ist) denotes a respondent i from state s interviewed in month-year t and $U_{s,t-p}$ is the unemployment rate in state s at time $t-p$.

The second measure, $M_{ist}^{RG,P}$, which I refer to as *the range*, is the difference between the largest and smallest unemployment rate values in past P months as of the time of an interview,

$$M_{ist}^{RG,P} = \max\{U_{s,t-p}\}_{p=1}^{P} - \min\{U_{s,t-p}\}_{p=1}^{P}. \tag{3}$$

Finally, the third measure, $M_{ist}^{FD,P}$, which I refer to as *the difference*, is the first difference between unemployment rate one and P months prior to an interview,

$$M_{ist}^{FD,P} = U_{s,t-1} - U_{s,t-P}. \tag{4}$$

Next, in order to distinguish between periods of economic expansion and contraction, I use a dummy variable, D_{ist}^{P} indicating if an individual was exposed to a decrease or increase in the unemployment rate in the months preceding his/ her interview. Specifically, $D_{ist}^{P} = \mathbf{1}_{[U_{ist}^{FD,P} \geq 0]}$ and thus, D_{ist}^{P} is equal to one if the unemployment rate increased (economic conditions worsened), and zero otherwise. Lastly, I interact D_{ist}^{P} with both *the mean* and *the range* to allow for the economic conditions to influence the BMI differently in recessions and economic expansions. In particular, $M_{ist}^{AV_R,P} = M_{ist}^{AV,P} \cdot D_{ist}$ and $M_{ist}^{RG_R,P} = M_{ist}^{RG,P} \cdot D_{ist}$ are *the mean* and *the range*, respectively, in periods of rising unemployment rate, and $M_{ist}^{AV_R,P} = M_{ist}^{AV,P} \cdot (1 - D_{ist})$ and $M_{ist}^{RG_R,P} = M_{ist}^{RG,P} \cdot (1 - D_{ist})$ are *the mean* and *the range*, respectively, in periods of declining unemployment rate.

In my estimation, I weight the observations by the BRFSS post-stratification weights. It allows me to account for varying probabilities of selection as well as nonresponse and noncoverage among different segments of the population. Consequently, I obtain estimates that are nationally representative. The BRFSS post-stratification weights are products of so-called design and adjustment weights. The design weight is a product of the inverse of the sampling fraction of a respondent's stratum, the number of adults in a respondent's household, and the inverse of the number of residential phone numbers in a respondent's household. The adjustment weight is computed as the number of people in a respondent's age-by-sex or age-by-race-by-sex category in the population of a region or a state divided by the sum of design weights of all respondents from the same category. Over years, as the number of respondents increased, both the mean and variance of post-stratification weights steadily decrease (see Table 1: Columns 6-7). Specifically, whereas in 1984, one respondent represented, on average, 4,850 individuals, by 2010, this number dropped to 527 [2]. In 2011, the BRFSS replaced post-stratification by a weighting methodology called iterative proportional fitting (*ranking*). In contrast to post-stratification, ranking allows sample and population distributions of main demographic characteristics to match more accurately and on more than three margins. One of the consequences of replacing post-stratification with ranking is that, without further adjustments, the BRFSS data before and after 2011 are not comparable, i.e., cannot be used for across-time analysis. Hence, at the moment, I restrict my sample to years prior to 2011. In the near future, however, I am planning to adjust the data and then, re-estimate the model using additional 2.7 million observations from years 2011-2016. To achieve that I will adapt a correction technique described in a paper by Dwyer-Lindgre et al. (2015) on drinking patterns in the United States published in the American Journal of Public Health.

[2]The final post-stratification weights provided by the BRFSS in 1994 are of a different order of magnitude than weights from other years. In particular, the mean, variance, and other moments are 100 times larger. I assume it to be caused by a computational error and to account for that, I divide all respondents' post-stratification weights by 100.

5 Estimation

In this research, I estimate the relationship between macroeconomic conditions and the BMI using two approaches that are a fully parametric linear model and semiparametric partially linear model.

The estimating equation of a fully parametric linear model is given by

$$Y_{ist} = \beta_0 + X_{ist}'\beta + Z_{ist}\gamma + u_{ist} \quad i = 1, ..., N, \ s = 1, ...S, \ t = 1, ...T, \tag{5}$$

where $X_{ist} = \left[X_{ist}^{C'}, X_{ist}^{S'}, X_{ist}^{T'}\right]'$ is a K_1x1 vector of control variables, Z_{ist} is a K_2x1 vector of macroeconomic regressors, and u_{ist} is an error term. The parameters of the model β_0, β, and γ are of dimensions 1x1, K_1x1, and K_2x1, respectively. As in Section 4, (ist) denotes an individual i from state s interviewed at time t.

The dependent variable is the BMI. The vector of control variables, X_{ist}, contains individual-level controls, X_{ist}^C, i.e., dummy variables for sex, age, race, ethnicity, marital status, educational attainment, and employment status (see Section 4 for more details), state dummies, X_{ist}^S, and a linear time trend, X_{ist}^T. State dummies capture all time-invariant state characteristics that may influence the BMI such as food culture and public awareness, whereas a linear time trend captures the trend component of the BMI. In my first specification, the vector Z_{ist} is three-dimensional and describes average economic conditions a respondent was exposed to prior to the BRFSS interview. Specifically, $Z_{ist} = \left[M_{ist}^{AV_R,P}, M_{ist}^{AV_E,P}, D_{ist}^P\right]'$. In the second specification, the vector Z_{ist} is five-dimensional. It not only describes average economic conditions, but also accounts for the magnitude of economic fluctuations. Specifically, $Z_{ist} = \left[M_{ist}^{AV_R,P}, M_{ist}^{AV_E,P}, M_{ist}^{RG_R,P}, M_{ist}^{RG_E,P}, D_{ist}^P\right]'$, where $M_{ist}^{AV_R,P}, M_{ist}^{AV_E,P}, M_{ist}^{RG_R,P}, M_{ist}^{RG_E,P}$ and D_{ist}^P were defined in Section 4.

I estimate equation (5) by weighted least squares. Because the BRFSS collects the data using a complex sampling procedure (see section 3), I follow the CDC's recommendation and incorporate the BRFSS sampling design into my estimation. Thus, my regression estimates are adjusted for the BRFSS strata, clusters, and sampling weights.

In addition to estimating a fully parametric linear model, I estimate its semiparametric partially linear version. An estimating equation is given by

$$Y_i = W_i'\delta + g(V_i) + v_i \quad i = 1, ..., N, \tag{6}$$

where the subscripts (st) are omitted for the ease of notation, g denotes an unknown smooth function, W_i is a K_3x1 vector of control variables, $V_i = [V_{1i}, V_{2i}]$ is a 2x1 vector of proxies for macroeconomic conditions, and v_i is an error term. The parameter vector of the model δ is of dimension K_3x1.

In the first specification $W_i = X_i$ and $V_i = \left[M_i^{AV,P}, D_i^P\right]$. In the second specification, $W_i = \left[X_i', M_i^{AV,P}, M_i^{AV,P}\right]'$ and $V_i = \left[M_i^{RG,P}, D_i^P\right]$. Note that g is a function of two arguments in both specifications. The first argument, V_{1i}, is a continuous variable equal to either $M_i^{AV,P}$ or $M_i^{RG,P}$. The second argument, V_{2i}, is discrete and takes on only two values, which are zero and one. Also, neither W_i nor V_i contain a constant term. It allows me to identify the parameter vector δ, while leaving the nonparametric component g unconstrained.

To estimate the model, I apply the three-step estimation procedure proposed by Robinson (1988) and moreover, weight the data by the BRFSS post-stratification weights. In the first step, I estimate $E\left(Y_i|V_i\right)$ and $E\left(W_i|V_i\right)$ by nonparametric regression as suggested by Racine & Li (2004). The corresponding estimates $\hat{E}\left(Y_i|V_i\right)$

and $\hat{E}(W_i|V_i)$ are given by

$$\hat{Y}_{iv} := \hat{E}(Y_i|V_{1i}, V_{2i} = v) = \frac{N^{-1}\sum_{j=1}^{n} Y_j K_h(V_{1j} - V_{1i}) \cdot \mathbf{1}_{[V_{2i}=v]}}{\hat{f}(V_{1i}, V_{2i} = v)}, \tag{7}$$

$$\hat{W}_{iv}^{w} := \hat{E}(W_i^w|V_{1i}, V_{2i} = v) = \frac{N^{-1}\sum_{j=1}^{N} W_j^w K_h(V_{1j} - V_{1i}) \cdot \mathbf{1}_{[V_{2i}=v]}}{\hat{f}(V_{1i}, V_{2i} = v)}, \tag{8}$$

where $i = 1, ..., N$, $k = 1, ...K$, and $v = 0, 1,$, and where W_i^w denotes the w^{th} component of W_i, $\hat{f}(V_{1i}, V_{2i} = v) = N^{-1}\sum_{j=1}^{N} K_h(V_{1j} - V_{1i}) \mathbf{1}_{[V_{2i}=v]}$ is an estimated density function of V (see Racine & Li (2003)), and moreover

$$K_h(V_{1j} - V_{1i}) \begin{cases} h^{-1}\kappa((V_{1j} - V_{1i})/h), & \text{kernel estimator} \\ r^{-1}\kappa((V_{1j} - V_{1i})/r), & k\text{-nearest neighbors estimator,} \end{cases} \tag{9}$$

where h denotes the bandwidth of a kernel estimator and r is the Euclidean distance between V_{1i} and the k^{th} nearest neighbor of V_{1i} among the V_{1j}'s in the k-nearest neighbors estimator as described by Liu & Lu (1997).

In the second step, I estimate the parametric component of the model by weighted least squares regression of $\left(Y_i - \hat{Y}_i\right)$ on $\left(W_i - \hat{W}_i\right)$, where $\hat{Y}_i = \left[\hat{Y}_{i0}, \hat{Y}_{i1}\right]$ and $\hat{W}_i = \left[\hat{W}_{i0}, \hat{W}_{i1}\right]$. I obtain the feasible $\sqrt{n}$-consistent Li (1996) estimator of δ given by

$$\hat{\delta} = \left(\sum_{i=1}^{N}\left(W_i - \hat{W}_i\right)\left(W_i - \hat{W}_i\right)' \hat{f}_i^2\right)^{-1} \sum_{i=1}^{N}\left(W_i - \hat{W}_i\right)\left(Y_i - \hat{Y}_i\right)\hat{f}_i^2 \tag{10}$$

Following Li (1996) all observations are weighted by their estimated densities $\hat{f}_i = \left[\hat{f}(V_{1i}, V_{2i} = 0), \hat{f}(V_{1i}, V_{2i} = 1)\right]$. This method differs from the trimming technique proposed by Robinson (1988), in which only observations with *large enough* densities remain in the sample and thus, contribute to the estimation of $\hat{\delta}$. Since the dependence of an estimator on an unknown trimming parameter can be problematic, I follow Li (1996) approach avoiding an arbitrary selection of that parameter.

Finally, in the third step, I consistently estimate the nonparamteric component g by

$$\hat{g}\left(V_{1i}, V_{2i}=v\right)=\frac{N^{-1} \sum_{j=1}^{N}\left(Y_{j}-W_{j}^{\prime} \hat{\delta}\right) K_{h}\left(V_{1j}-V_{1i}\right) \cdot \mathbf{1}_{\left[V_{2i}=v\right]}}{\hat{f}\left(V_{1i}\right)}, \quad v=0,1. \quad (11)$$

6 Results

The related literature uses the default value $P = 12$ (months) for the time window in equations 2 to 4 but offers no justifications for that particular choice. In order to analyze the potential sensitivity of my results with respect to the choice of P, I will instead consider several values ranging from two to eighteen. This will enable me to analyze how BMI is affected by economic conditions over periods of time up to one year and a half before interview time. To investigate the dependence on P, I estimate the fully parametric linear model (see equation 5) for $P \in \{2, ..., 18\}$. I run seventeen regressions for each of the two specifications (one for *the mean* and one for *the mean* and *the difference*) discussed in Section 5. I find that when regressing the BMI on the mean unemployment rate, coefficients associated with the *mean* are barely affected by the choice of P (see Figure 5). However, when regressing the BMI on both the mean unemployment rate and its range, coefficients associated with the *range* clearly depend on the choice of P. They initially monotonically decrease as P increases until reaching stable levels for $P \geq 7$. This makes sense as one would expect an adjustment period in the response of BMI to changes in unemployment. Since, furthermore, confidence intervals, shrink as P increases, I decided to use $P = 18$ for the parametric and semiparametric results discussed below.

Comparing the 95% confidence intervals for state dummies, I find that, with the exception of two states, the remaining forty nine states can be classified into two subsets within which fixed effects are not significantly different from one another (see Figure 6). The two singletons are Colorado with the estimated coefficient of -0.6 and the District of Columbia with the coefficient of -0.5. The two subsets, to which I refer to as *the 3rd set* and *the 4th set*, contain nineteen and thirty states, respectively. The average value of the coefficients is equal to -1.0 in *the 3rd set* and to 0.5 in *the 4th set*. Interestingly, these two subsets are geographically separated. Whereas states belonging to *the 3rd set* are concentrated in Northeast and West, states belonging to *the 4th set* are clustered in Midwest and South. I use this classification to re-estimate a fully parametric linear model. Specifically, I replace the initial state fixed effect parameters with only three parameters corresponding to the above classification. The advantage of this approach is that I can substantially reduce the number of parameters in the estimation of equation 5 at the cost of a negligible loss of information.

The regression of the BMI on the mean unemployment rate indicates that the relationship between the BMI and mean unemployment rate is always negative and statistically significant (see Table 2). Specifically, in periods of rising unemployment, a one percentage point *increase* in the mean unemployment rate, *decreases* BMI by 0.03, whereas in periods of declining unemployment, a one percentage point *decrease* in the mean unemployment rate, *increases* BMI by 0.01. Moreover, these two coefficients of interest are statistically different, as indicated by the adjusted Wald test. Consequently, I conclude that average economic conditions affect the BMI differently depending on the phase of the business cycle. In particular, the relationship intensifies in recessions and abates during economic expansions.

More interestingly, the regression of the BMI on both the mean and range of the unemployment rate indicates that not only average economic conditions, but also the amplitude of economic fluctuations both have a strong and statistically significant effect on the BMI (see Table 3). Specifically, in periods of rising unemployment, a one percentage point *increase* in the range of the unemployment rate, *decreases* BMI by 0.01, whereas in periods of declining unemployment, a one percentage point *increase* in the range of the unemployment rate, *increases* BMI by 0.04. This suggests that rapid economic recoveries have an effect on BMI that is four times stronger than that of recessions. Moreover, it shows that in periods of declining unemployment, the amplitude of economic fluctuations has an effect on BMI that is twice as strong than that of the *the mean*. In summary, these results suggests that dispersion of the unemployment rate plays an important, but so far obscure, role in understanding the relationship between economic conditions and the BMI. Moreover, I want to emphasize the importance of distinguishing between different phases of the business cycle in modeling that relationship.

As discussed in Section 5, I also estimate a semiparametric partially linear model using two different estimators in the nonparametric regressions, 7, 8), and 11. In both estimators, I use the Epanechnikov function, which is the most efficient in minimizing the mean integrated squared error. In the kernel estimator, I estimate the bandwidth using the Silverman's rule of thumb estimator. In the k-nearest neighbor estimator, I experiment with a range of values for k and find that my results are robust relative to the choice of k. Moreover, I find that in regions where the density of the unemployment rate is sparse, the k-nearest neighbor estimator outperforms the kernel estimator. Therefore, I discuss next my main results obtained using the k-nearest neighbor estimator.

I find that the relationship between the mean unemployment rate and BMI can be reasonably approximated by a fully parametric linear model. The same holds for the range of the unemployment rate in periods of rising unemployment. However, in periods of declining unemployment, the relationship between *the range* and BMI is highly nonlinear and thus, can be severely underestimated by a linear regression model. Specifically, for small enough values of dispersion, the BMI is unaffected by changes to the amplitude of economic fluctuations. However, as the difference between maximum and minimum unemployment rate values gets larger and in particular, exceeds 2%, the BMI starts to logarithmically increase with *the range*. This suggests that economic agents largely ignore minor changes in unemployment when the economy is relatively tame and only react to economic fluctuations that exceed some threshold. Note that these results are robust relative to controlling for month fixed effects, cohort fixed effects as well as to restricting the range of BMI to $[12, 56)$.

Acknowledgment

I am grateful to Jean-François Richard, Stefania Albanesi, and Arie Beresteanu for their assistance and suggestions. Remaining errors are my sole responsibility.

References

Dwyer-Lindgre, L., Flaxma, A. D. & Ng, M. (2015), 'Drinking patterns in us counties from 2002 to 2012', American Journal of Public Health **105**, 1120–27.

Li, Q. (1996), 'On the root-n-consistent semiparametrc estimation of partially linear models', Economics Letters **51**(3), 277–285.

Liu, Z. & Lu, X. (1997), 'Root-n-consistent semiparametric estimation of partially linear models based on k-nn method', Econometric Reviews **16**(4), 411–420.

Mokdad, A. H., Ford, E. S. & Bowman, B. A. (2003), 'Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001', Journal of the Medical American Association **289**, 76–79.

Prospective Studies Collaboration (2008), 'Body mass index and cause-speci
c mortality in 900 000 adults: collaborative analyses of 57 prospective studies', Lancet **373**, 1083–96.

Racine, J. & Li, Q. (2003), 'Nonparametric estimation of distributions with categorical and continuous data', Journal of Multivariate Analysis **86**, 266–292.

Racine, J. & Li, Q. (2004), 'Nonparametric estimation of regression functions with both categorical and continuous data', Journal of Econometrics **119**, 99–130.

Renehan, A., Tyson, M., Egger, M., Heller, R. & Zwahlen, M. (2008), 'Body mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies', Lancet **371**, 69–78.

Robinson, P. (1988), 'Root- n-consistent semiparametric regression', Econometrica **56**(4), 931–954.

Ruhm, C. (2000), 'Are recessions good for your health?', Quarterly Journal of Economics **115**(2), 617–650.

Ruhm, C. (2003), 'Good times makes you sick', Journal of Health Economics **22**(4), 637–658.

The Behavioral Risk Factor Surveillance System (1984-2015c), 'Data files', Center for Disease Control and Prevention .

The Behavioral Risk Factor Surveillance System (1984-2015e), 'Questionnaire', Center for Disease Control and Prevention .

The Behavioral Risk Factor Surveillance System (1996-2015a), 'Codebook', Center for Disease Control and Prevention .

The Behavioral Risk Factor Surveillance System (1996-2015b), 'Comparability of data', Center for Disease Control and Prevention .

The Behavioral Risk Factor Surveillance System (1996-2015d), 'Overview', Center for Disease Control and Prevention .

Appendix A

Table 1: Column 1: Number of states participating in the BRFSS by year. Columns 2-3: Total number of observations and average number of observations per state (including observations with missing BMI) by year. Column 4: Percentage of observations with missing BMI by year. Columns 5-7: Total number of observations, mean of post-stratification weights and standard deviation of post-stratification weights by year (excluding observations with missing BMI).

Year	(1) Number of states	(2) Number of observations	(3) Average number of observations per state	(4) Percentage of missing values on BMI	(5) Number of observations	(6) Mean of post-stratification weights	(7) Std. dev. of post-stratification weights
1984	15	12,258	817	3.30%	11,853	4850.61	6113.42
1985	22	25,221	1,146	2.86%	24,500	3539.35	4427.95
1986	26	34,395	1,323	2.86%	33,413	3000.56	3857.33
1987	33	50,081	1,518	2.87%	48,645	2322.11	3431.34
1988	37	56,448	1,526	2.83%	54,849	2535.02	3311.28
1989	40	66,867	1,672	4.44%	63,901	2398.65	2967.23
1990	45	81,557	1,812	2.71%	79,348	2090.67	2652.84
1991	48	87,846	1,830	2.89%	85,308	2055.03	2450.50
1992	49	96,213	1,964	3.11%	93,220	1903.70	2058.12
1993	50	102,263	2,045	2.98%	99,219	1848.01	2130.75
1994	50	105,853	2,117	3.36%	102,292	1780.64	2182.72
1995	50	113,934	2,279	3.23%	110,255	1698.22	2378.75
1996	51	122,268	2,397	3.94%	117,446	1612.80	1905.14
1997	51	133,321	2,614	3.60%	128,518	1490.79	1801.77
1998	51	146,992	2,882	3.94%	141,196	1365.22	1748.80
1999	51	156,937	3,077	3.84%	150,910	1291.82	1773.92
2000	51	180,244	3,534	4.48%	172,173	1138.84	1614.40
2001	51	205,140	4,022	4.94%	195,012	1035.64	1471.93
2002	51	240,735	4,720	4.74%	229,327	887.07	1375.03
2003	51	257,659	5,052	4.85%	245,169	844.44	1365.37
2004	50	296,971	5,939	4.70%	283,003	736.16	1284.03
2005	51	349,901	6,861	4.62%	333,744	632.25	1136.53
2006	51	347,790	6,819	4.91%	330,726	646.06	1322.51
2007	51	423,783	8,309	4.58%	404,363	536.62	1084.30
2008	51	406,749	7,975	4.51%	388,423	565.67	1083.81
2009	51	424,592	8,325	4.58%	405,135	546.34	1053.97
2010	51	444,927	8,724	4.77%	423,697	527.15	959.39
Total		4,970,945			4,755,645		

Appendix B

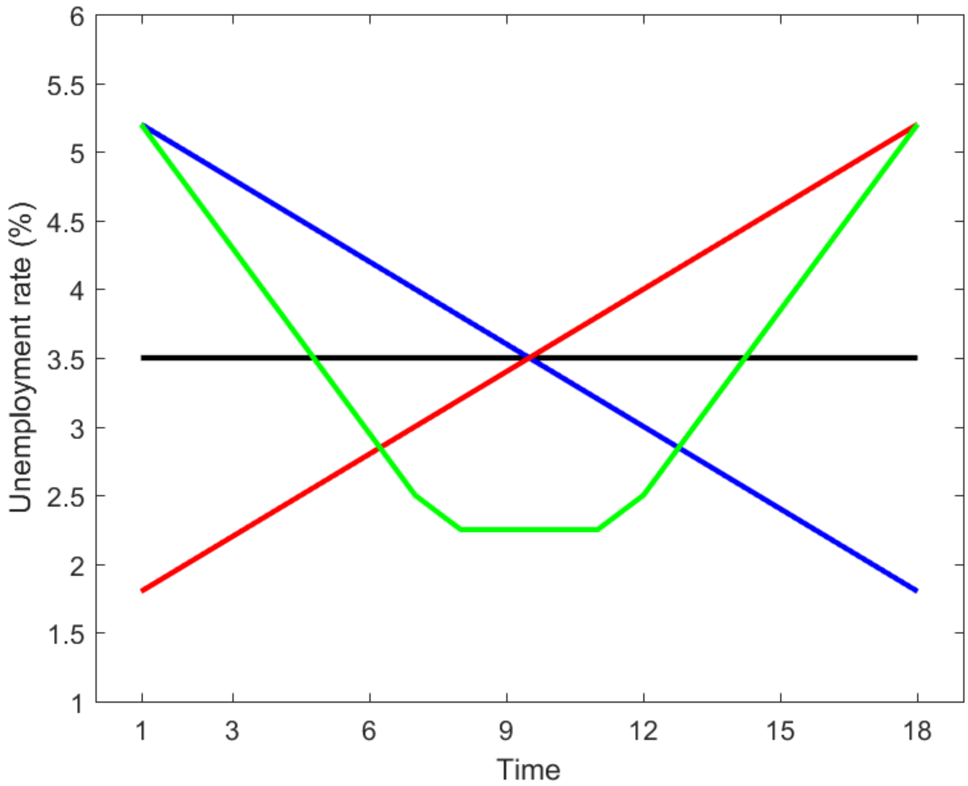


Figure 1: Four hypothetical unemployment rate profiles with the same mean (3.5%) computed over the period of 18 months.

Table 2: Linear regression results with the mean unemployment rate computed over eighteen months and with state fixed effects replaced by state group dummies. I use BRFSS data from 1984 to 2010. The dependent variable is the BMI. A vector of control variables contains individual-level characteristics and a linear time trend. The data is weighted by the BRFSS post-stratification weights and adjusted for the BRFSS sampling design.

Variable	Coefficient	Stdandard error
The mean in periods of increasing unemployment	-0.030	0.003
The mean in periods of decreasing unemployment	-0.011	0.004
The difference dummy	0.099	0.030
Linear trend	0.131	0.001
Age 18-24	-0.585	0.027
Age 25-29	0.794	0.026
Age 30-34	1.369	0.026
Age 35-39	1.726	0.026
Age 40-44	2.022	0.026
Age 45-49	2.323	0.027
Age 50-54	2.513	0.026
Age 55-59	2.622	0.026
Age 60-64	2.470	0.024
Age 66-69	2.153	0.023
Age 70-74	1.702	0.023
Age 75-79	1.135	0.023
Female	-1.128	0.009
White	0.489	0.021
Black	2.050	0.026
Hispanic	0.823	0.020
Married	0.073	0.010
High school dropout or less	1.497	0.017
High school graduate	1.065	0.011
Some college	0.949	0.011
Employed for wages	-0.006	0.017
Self-employed	-0.384	0.020
Out of work or unable to work	0.851	0.025
Homemaker	-0.197	0.022
Student	-0.682	0.031
Age NaN	0.397	0.058
Race NaN	0.480	0.067
Ethnicity NaN	0.149	0.101
Education NaN	0.464	0.109
Marital status NaN	-0.548	0.094
Employment status NaN	-0.121	0.097
District of Columbia	0.150	0.036
The 3rd set of states	0.554	0.021
The 4th set of states	0.990	0.021
Constant	21.025	0.043
Observations	4755645	
R-squared	0.110	

Table 3: Linear regression results with the mean and range on the unemployment rate computed over eighteen months and with state fixed effects replaced by state group dummies. I use BRFSS data from 1984 to 2010. The dependent variable is the BMI. A vector of control variables contains individual-level characteristics and a linear time trend. The data is weighted by the BRFSS post-stratification weights and adjusted for the BRFSS sampling design.

Variable	Coefficient	Stdandard error
The range in periods of increasing unemployment	-0.014	0.006
The range in periods of decreasing unemployment	0.040	0.011
The mean in periods of increasing unemployment	-0.026	0.004
The mean in periods of decreasing unemployment	-0.018	0.004
The difference dummy	0.092	0.030
Linear trend	0.131	0.001
Age 18-24	-0.586	0.027
Age 25-29	0.794	0.026
Age 30-34	1.370	0.026
Age 35-39	1.726	0.026
Age 40-44	2.022	0.026
Age 45-49	2.323	0.027
Age 50-54	2.513	0.026
Age 55-59	2.622	0.026
Age 60-64	2.470	0.024
Age 65-69	2.154	0.023
Age 70-74	1.702	0.023
Age 75-79	1.135	0.023
Female	-1.128	0.009
White	0.490	0.021
Black	2.049	0.026
Hispanic	0.824	0.020
Married	0.073	0.010
High school dropout or less	1.497	0.017
High school graduate	1.064	0.011
Some college	0.949	0.011
Employed for wages	-0.006	0.017
Self-employed	-0.384	0.020
Out of work or unable to work	0.852	0.025
Homemaker	-0.197	0.022
Student	-0.682	0.031
Age NaN	0.397	0.058
Race NaN	0.480	0.067
Ethnicity NaN	0.148	0.101
Education NaN	0.464	0.109
Marital status NaN	-0.548	0.094
Employment status NaN	-0.123	0.097
District of Columbia	0.147	0.036
The 3rd set of states	0.558	0.021
The 4th set of states	0.992	0.021
Constant	21.015	0.043
Observations	4755645	
R-squared	0.110	

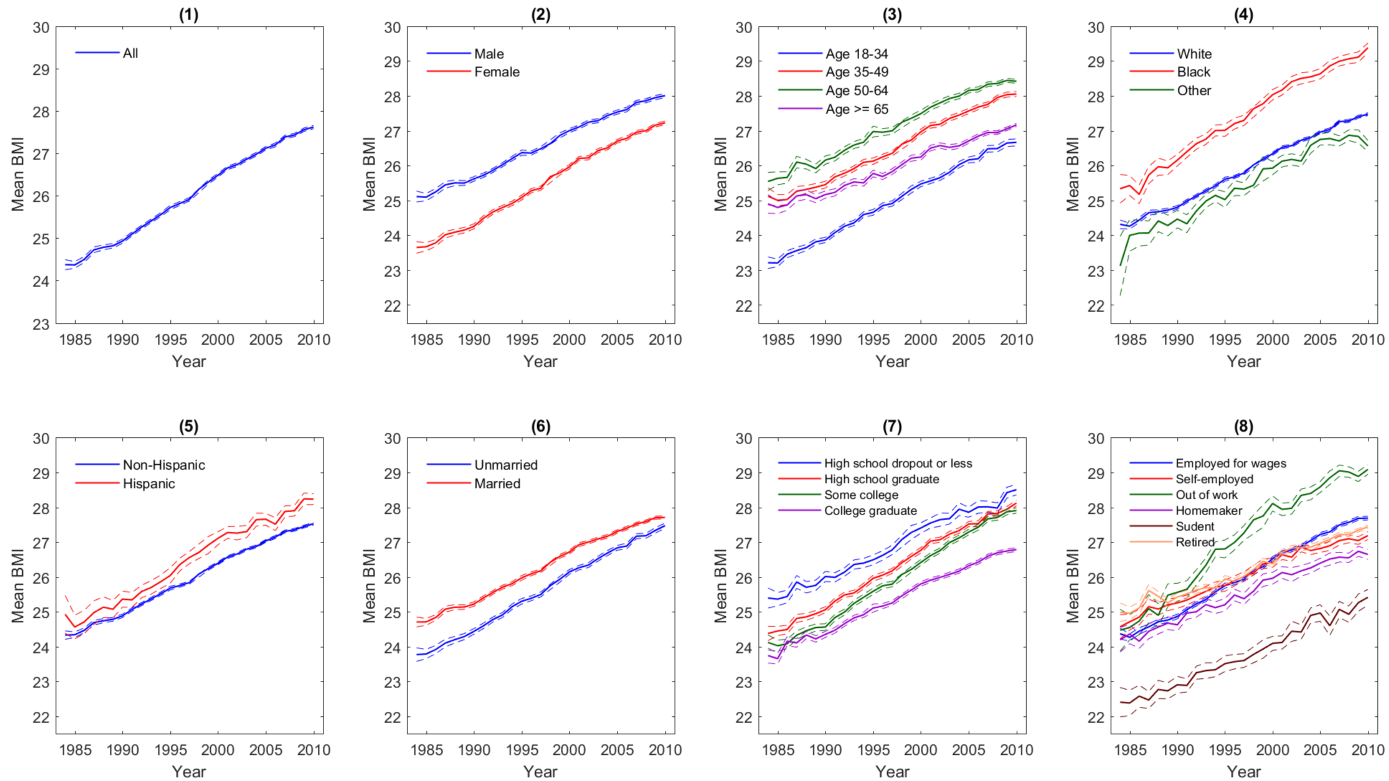


Figure 2: Mean BMI of adult U.S. residents aged 18 and over by between 1984 and 2010 (1). Mean BMI of adult U.S. residents aged 18 and over by (2) sex, (3) age, (4) race, (5) ethnicity, (6) marital status, (7) education level, and (8) employment status between 1984 and 2010. All observations are weighted by the BRFSS post-stratification weights. Dashed lines indicate 95% confidence interval.

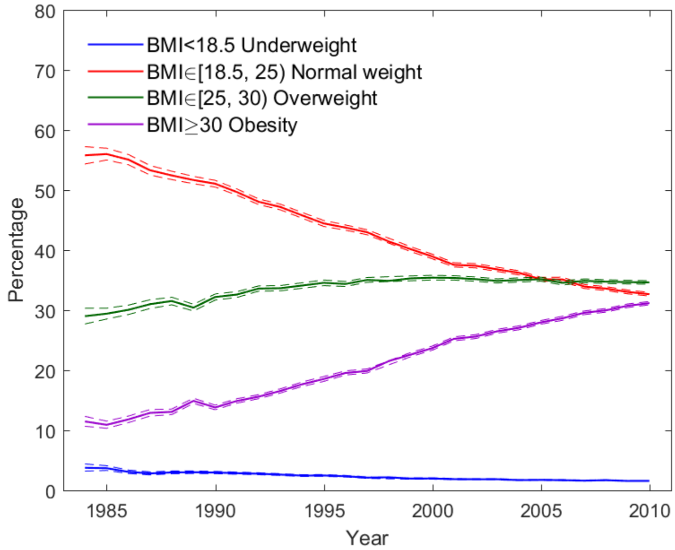


Figure 3: Percentage of adult U.S. residents aged 18 and over with underweight, normal weight, overweight, and obesity between 1984 and 2010. All observations are weighted by the BRFSS post-stratification weights. Dashed lines indicate 95% confidence interval.

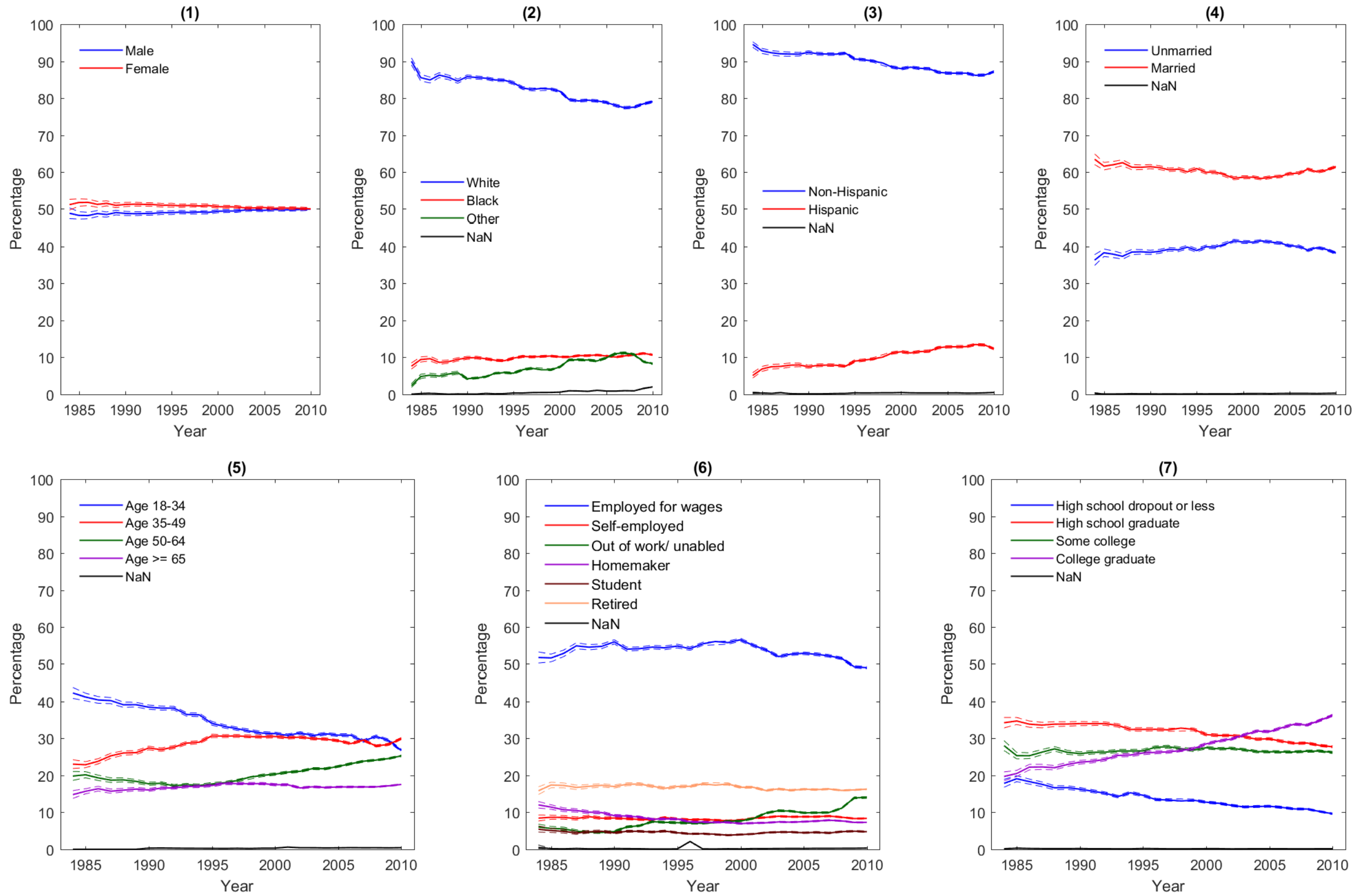


Figure 4: Summary statistics. Percentage of adult U.S. residents aged 18 and over by (1) sex, (2) race, (3) ethnicity, (4) marital status, (5) age, (6) employment status, and (7) education level between 1984 and 2010. All observations are weighted by the BRFSS post-stratification weights. Dashed lines indicate 95% confidence interval.

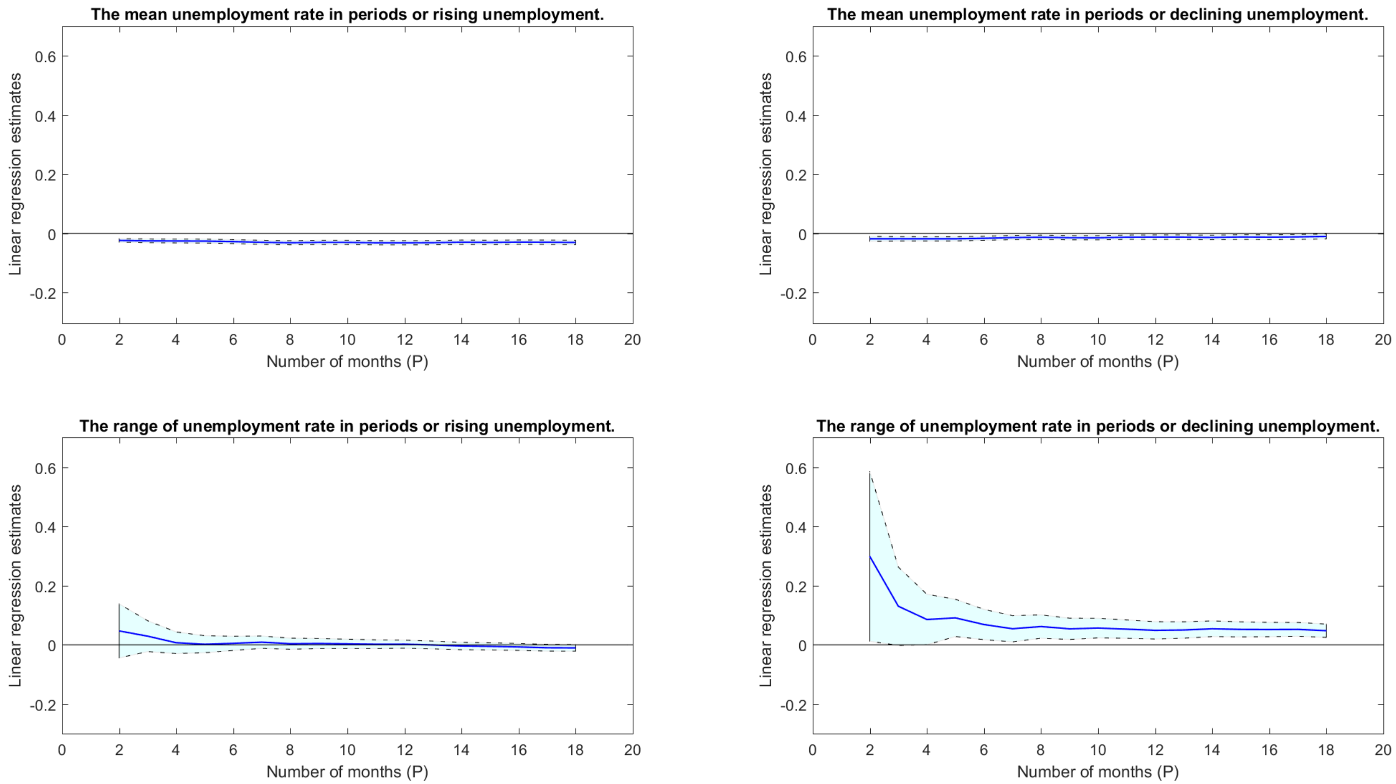
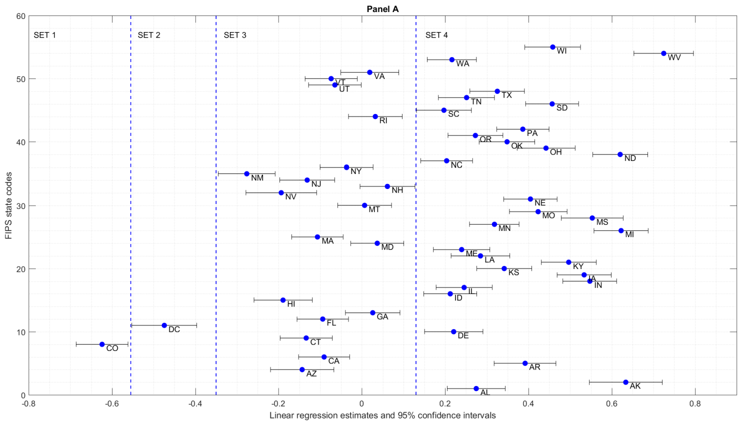


Figure 5: Linear regression estimates of the mean (top panel) and range (bottom panel) of the unemployment rate over P months in periods of rising (left) and declining (right) unemployment. The solid line indicates point estimates and the shaded area 95% confidence intervals. For estimation, I use BRFSS data from 1984 to 2010. The dependent variable is the BMI. A vector of control variables contains individual-level characteristics, state fixed effects, and a linear time trend. The data is weighted by the BRFSS post-stratification weights and adjusted for the BRFSS sampling design.



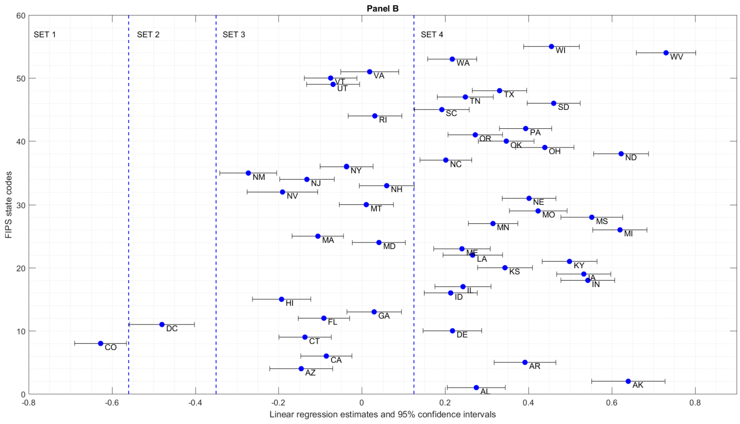


Figure 6: Linear regression estimates of the state fixed effects with 95% confidence intervals. Panel A corresponds to a regression of the BMI on the mean unemployment rate. Panel B corresponds to a regression of the BMI on the mean unemployment rate and its range. For estimation, I use BRFSS data from 1984 to 2010. In addition to state fixed effects, I control for individual-level characteristics and a linear time trend. The data is weighted by the BRFSS post-stratification weights and adjusted for the BRFSS sampling design. The two-letter annotations indicate the American National Standards Institute (ANSI) alphabetic codes for each state. Based on the estimated coefficients and 95% confidence intervals I classify the states in four no-overlapping groups as indicated in the upper left part of the figure.

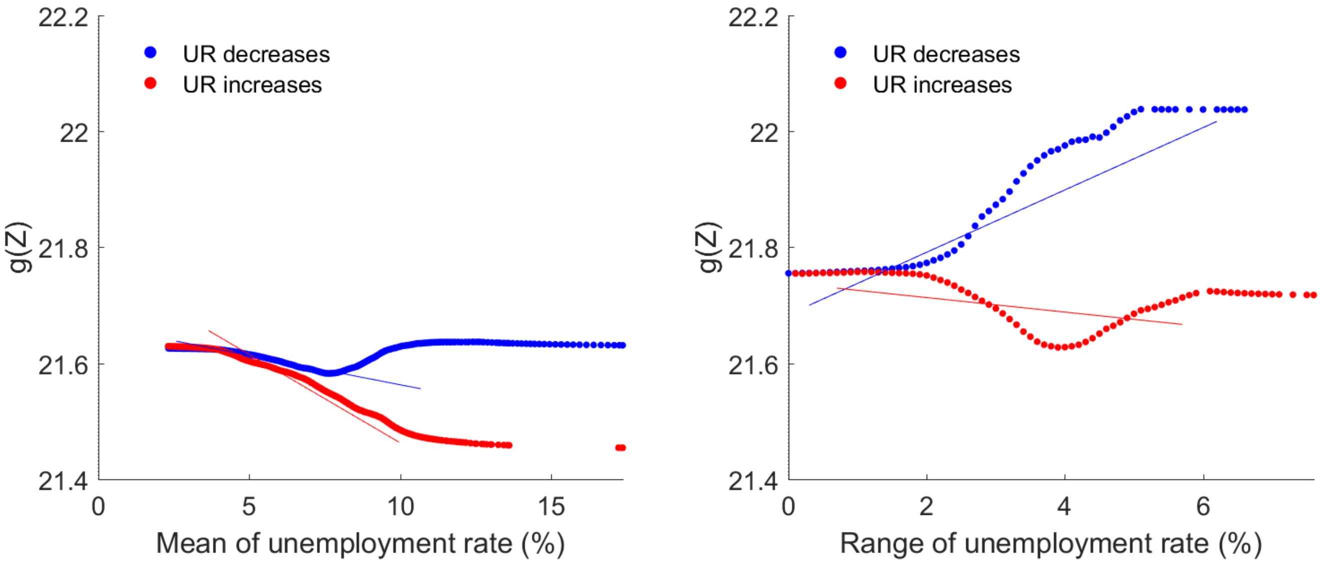


Figure 7: Semiparametric estimates of the functional relationship between the mean unemployment rate and the BMI (left panel) and the range of the unemployment rate and the BMI (right panel). Blue lines indicate periods of declining unemployment, whereas red lines refer to periods of increasing unemployment. For estimation, I use BRFSS data from 1984 to 2010. The dependent variable is the BMI. A vector of control variables contains individual-level characteristics, state fixed effects, and a linear time trend. The data is weighted by the BRFSS post-stratification weights. Left panel: number of variables entering the model parametrically: 82, number of unique values of the mean unemployment rate: 1,705 (two decimal points), total numbe of observations: 4,755,645. Right panel: number of variables entering the model parametrically: 83, number of unique values of the range of the unemployment rate: 136 (one decimal point), total number of observations: 4,755,645.