The Health Effect of Commuting: Evidence from Metro Opening in China*

Chen Fang†

January 10, 2023

Abstract

Extreme commuting in mega-cities is becoming a huge problem for both government officers and large amount of workers. The social cost, especially the health effect of commuting needs more attention by economists. This paper introduces the background and summarize related literature about the commuting, urban development and its social effects (externality). Then I use the CFPS micro survey data to analyze the casual effect between commuting and health state. Using the traditional OLS model, I find that commuting will be detrimental for individual's health. Then I employ a staggered DID model with the opening of metro system as an exogenous policy shock which can reduce workers' commuting time. The result shows the new approach of commuting by urban metro system has no significant effect on commutes' physical health state while depress them more mentally. I provide robustness check (ordered logit model, parallel and placebo tests), heterogeneity and mechanism analysis to make the regression results reliable. At last, I refer to the classical AMM model and establish a modified version of model including the negative utility of commuting for individual.

Keywords: Commuting, Health Effect, Urbanization, Social Cost

JEL Codes: I18, J61, R13, R41

*I would like to thank Professor Xuezheng Qin for helpful comments and suggestions. I am responsible for all remaining errors.

†College of Environmental Sciences and Engineering, Peking University. Email: fang@pku.edu.cn

1 Introduction

With rapid urbanization and industrialization, the cities in China are experiencing an extraordinary process of sprawl in their coverage area and an increase in population. Emerging mega-cities gradually start to play an important role in China's economy. However, the unavoidable consequence of "over-development" is many public infrastructures and services cannot satisfy the demand of citizens in an economical and cost-effective way. Nowadays, a typical phenomenon is, for plenty of residents, the distance between their living and working places is so long that they have to bear long-time and high-cost daily commuting. Recently, a report [1] has gone viral on the Internet, which is focused on the special part of the "Daily Commuting Group" (commuters or *Shangbanzu* in Chinese), whose commuting pattern is called "Extreme Commuting" (one-way time of more than 60 minutes). This report offers a cross-section perspective on their difficulties and tiredness in daily commuting back and forth from the suburb (living) to the city center (working), evoking resonate and thinking in people about commuting.

According to the newest statistics[2], the proportion of commuters with a one-way time of more than 60 minutes in 44 major domestic cities is 13% in 2021, the population suffering extreme commuting has exceeded 14 million. These trends reveal the truth that the social pressure from commuting is an ignorant problem of urban planning. The root of this phenomenon is "Home-Work Imbalance" (or "Jobs-Housing Imbalance"), which spurs the agglomeration of huge-scale real estate in some of the suburban districts or towns (for example, *Tiantongyuan* Community and *Huilongguan* Community are two famous real estates in Changping District of Beijing), these towns or districts are often jokingly called "Commuter Towns" (*Wo Cheng* or *Shui Cheng* in Chinese)[3].

As a typical symptom of "urban disease", commuting is considered as one of the least enjoyable activities (Kahneman et al., 2004). Large-scale long-time commuting leads to many

[1]每日人物｜极端通勤，1400 万人的集体内耗
[2]中规智库｜《2022 年度中国主要城市通勤监测报告》
[3]Other examples include Kunshan and Yanjiao.

social costs, such as traffic congestion in rush hours of morning and evening peak, which can cause pollutants emissions and energy consumption, detrimental to the maintenance of clean air and the efforts to deal with climate change. In addition, commuting takes up so much time that workers have to cut off their sleep and relaxation time, they also have to be concentrated enough to bear the jam of people and traffic during the exhausting journey for a long time every day. This is a huge threat to their health and many commuters say they are exhausted and weary during commuting especially after a long day working, they often don't have enough time for breakfast or have to run and walk quickly to chase the coming bus or subway, which is also a heavy burden and disturbance for their digestion system. The "sub-health" state from extreme commuting is valuable for deeper analysis and discussion, which will be meaningful for better urban planning and the achievement and implementation of the "Healthy China Initiative".

Another point is the development of metro system, which is important for many commuters. China have built so many new urban rail transit systems [4] in past 20 years. Before the year 2000, only Beijing, Tianjin, Shanghai and Guangzhou had metro systems[5]. But by 2020, China has put 233 urban rail transit lines into operation in 44 cities with length more than 7,500 km, which is a huge progress[6]. The metro building boom in China is such a miracle in line with the rapid urbanization, the degree of urbanization increases from 36.09% (2000) to 63.89% (2020)[7]. The characteristics of metro systems include on-time rate above 99%, high speed, low commute cost, all of these make metro a convenient and reliable approach for commuting. The operation of metro system extends the radius and distance from residential spot to working place (commuters can choose further region from the center of city to live with lower rent cost and house price).

I make two main contributions to present literature. First, I add the latest CFPS (China Family Panel Studies) micro data in my research, which contains five waves (2012, 2014,

[4] In fact, urban rail transit is a broader concept, it includes subway, light rail, tram and maglev.
[5] Why Has China Built So Many Metro Systems? (May 25, 2022).
[6] China builds more urban rail transit lines in 2020 (Jan 10, 2021).
[7] Data Resource: National Bureau Statistics.

2016, 2018 and 2020). The full combined panel data contains the observations of at least 10000 households in five periods, which is a complete nationally-representative longitudinal data and worthwhile to be applied in empirical research. Second, to deal with the endogenous problems, I use the opening of metro system in a city as a key exogenous shock. The cities that have experienced metro opening during the survey periods belong to the treatment group, contrast to the controlled cities having no metro or always having metro from 2010 to 2020. The staggered DID (difference-in-differences) strategy is used to eliminate endogeneity.

The goal of this research includes three parts: (1) Specify and quantify the causality between the commuting intensity and people's health state; (2) Explore the transmission channel of the effect the commuting on people's health; (3) Offer some implications about the relationship among commuting pattern, urban projection and the health state of modern people for policymakers.

The rest of this paper is organized as follows. Section 2 lists three pillars of related literature on this topic and their main conclusions; Section 3 describes data sources and basic variables of raw data; Section 4 introduces the main empirical strategies I applied and presents the regression results; Section 5 complements robustness check and heterogeneous analysis based on the Section 4, also provides some possible mechanism and tries to modify traditional model according to the results. Section 6 concludes and gives some implications.

2 Literature Review

Actually, commuting is a troubling problem across the world, nearly every city has a stage of imbalances between limited resources and an overloaded population, and commuting pressure is common in foreign countries. There is a lot of literature researching on it, mainly including three pillars.

The first pillar is the social effects of commuting. Except for the real and opportunity cost of commuting from the individual perspective, there are different sources of the social cost of commuting, such as air pollutants emissions, energy consumption (commuting will intensify traffic congestion and prolong the time of vehicles on the road) and obstacles for the employment of vulnerable groups (the low-income group often has difficulty to seek high-return jobs which in demand of high-cost and inconvenient commuting) (Kain, 1968; Wang et al., 2018; Zheng et al., 2014). In addition, the tiredness caused by commuting will reduce a person's working productivity and worsen their working performance. Hence, when extrapolating to the whole of society, the productivity and competitiveness of a firm will be impeded by extreme commuting. Monte et al. (2018) estimate the welfare increase from the reduction in commuting costs. Lu et al. (2019) use a natural experiment (the opening of a nearby subway station) to identify how an improvement in commuting affects employees' working performance of two firms. Sun and He (2022) discuss the negative impact of "Home-Work Separation" and the induced time squeeze on the enterprises' production efficiency. However, commuting is not a problem if the time can be controlled within a reasonable range, it seems that there are many benefits for the individual's well-being and cities' development inversely. Wang and Wei (2018) provide evidence of this, they find that when leisure and commuting were complementary, commuting time will force individuals to improve long-term performance and long-term income levels. The separation of working and living places is also beneficial for enterprises to enjoy the advantages of an agglomeration economy more effectively and improve labor productivity (Lucas and Rossi-Hansberg, 2002).

The second pillar is the effect of urban economic development on citizens' physical and mental health. The effect includes two aspects. On the one hand, urbanization has made a significant contribution to the decline in mortality in high-income regions because of the introduction of effective medicines, antibiotics, and vaccinations (Davis, 1956). On the other hand, during the last decades, rapid economic growth and urbanization triggered the deterioration of the environment, especially in China. Water and air pollution threaten the health and disturb people's normal life significantly, premature death is really a tough problem (Cohen et al., 2004). COVID-19 brings a larger health challenge to citizens, where they have a higher probability to be infected and bear the lock-down management. The result is COVID-19 has crowded out non-COVID-19-related healthcare demands, while distress and anxiety become more common (Brodeur et al., 2021). In addition, citizens' lifestyles and health expectations have changed a lot with urbanization, citizens tend to consume more fat and smoke more (Van de Poel et al., 2012). In China, "Health Human Capital" is extensively applied in economics research (Liu et al., 2004; Wang et al., 2008), and the "Healthy China" (2019-2030) goal gradually becomes a core objective for many cities, which shows an opportunity to balance urban economic development and citizens' health state.

The third pillar is focused on the specific relationship between commuting and health. The conclusion is obvious and clear, commuting will indeed influence citizens' basic health state, both physically and mentally. Both Nie and Sousa-Poza (2018) and Sun et al. (2019) use the CFPS micro data to estimate the effect of commuting duration on public health, they conclude that the negative effect is significant and sleep time is a key mediation variable. Mental health is also very important when considering commuting. Commuting means the trade-off between the efforts for commuting and the return provided by working, according to the hypothesis of rational man, one person is trying to maximize his or her utility by making choices, but the fact is the extension of the time for commuting cannot be offset by the better return of working. This phenomenon is called "Commuting Paradox", which can weaken the happiness and well-being of people (Wu, 2017). Roberts et al. (2011) take advantage of the British Household Panel Survey to identify the effect of commuting on the psychological health of men and women and the reasons for gender differences. Their result shows that commuting has an important detrimental effect on the psychological health of women, but not on men. Women's greater sensitivity to commuting time seems to be a result of their larger responsibility for day-to-day household tasks, including childcare and housework. Therefore, many scholars put forward the concept of the "Happiness Commute" (which can be roughly equal to commuting less than 5 kilometers) nowadays.

3 Data

3.1 Data Source

My analysis is based on data from the China Family Panel Studies (CFPS) implemented by the Institute of Social Science Surveys (ISSS) of Peking University, which currently consists of six waves (2010 to 2020, this survey is implemented every two years). This survey, administered to a nationally representative sample from across 25 provinces/municipalities/autonomous regions containing 95% of the Chinese population, is designed to capture socioeconomic development and economic/non-economic well-being in Chinese households, encompassing multiple dimensions such as educational attainment, family relationships and migration, physical and mental health, and economic activities (Xie and Lu, 2015).

CFPS contains a series of questions about the life habits, health state and working pattern of respondents. A series of questions including the frequency (smoking, drinking alcohol and exercising) and duration (noon break, night sleeping and house keeping), collect the respondents' life habits. Another series of questions collect the basic health information, including the chronic diseases history and the frequency of hospitalization. In the questionnaire, there are also questions about the working, such as one-way commute time for workers (labelled as "qg3011") and the monthly take-home pay ("qg11").

In particular, there are two key dependent variables in my research. The first one is the self-evaluated health state (SEHS, labelled as "qp201") which includes five levels from 1-5 (higher index means worse health state). The second one is the mental health score (MHS, labelled as "cesd20sc"). It is the sum of the scores a person evaluates for a series of descriptions[8] related to feeling and behaviors. The scale is designed by the Center for Epidemiologic Studies Depression Scale (CES-D), it is an authoritative tool to measure the extent of individual's depression. In CFPS, 2016, 2018 and 2020 wave add the CESD scale test in the questionnaire, the higher score represents the worse and more negative mental

[8]Some descriptions are "I was bothered by things that usually don't bother me.", "I did not feel like eating; my appetite was poor", respondents are needed to give the frequency of each description.

health state.

I also collect the information about the year of metro opening in Chinese large and middle-sized cities from the Internet in order to generate a binary variable ("subway") to denote whether a specific city has a metro system or not in specific year (0 means no metro, 1 means with metro).

3.2 Descriptive Analysis

I restrict the study sample to people above 16 and under 65, containing 34,179 individuals and 91,568 observations covering China's 25 provinces, providing a rather large sample for relatively accurate estimations. In the empirical regression process, I further drop the observations with missing information on the key variables such as one-way commute time, self-evaluated health state, mental health score, gender and age.

Table 1 presents the key variables and their summary statistics. I include one-way commute time, self-evaluated health state and mental health state as key variables, with some demographic and socioeconomic factors for controlling, such as one's age in years, urban-rural status, gender, education level[9].

According to Table 1, about 55% of the observations are male, and the average age is near 42. Two groups in my research have different characteristics in commute time, respondents living in cities with metro report a longer time taken for commuting than the group living in cities without metro, while their basic physical and mental health state have no significant differences. Moreover, commuters living in cities with metro have higher education and income level, which is reasonable because people who have received more education tend to live in large cities and get higher salary.

As for the metro opening data, Table 2 presents the opening time of metro system in different cities from 2012-2020 (collected from Internet, have been verified). There are 31 cities in China opened metro systems during 2010-2020, on average 5∼6 more cities will

[9] I also use several index about commuters' life habits and basic health state in CFPS raw data, such as the frequency for exercising, doing chores and entertainment.

have metro systems in operation after each wave of CFPS. The distribution of these cities is not so balanced, mainly concentrated in eastern cities, most of them are capital cities and middle-sized cities in province with larger population and better economy performance.

4 Empirical Strategies

4.1 Baseline Regression

I employ a simple OLS (Ordinary Least Squares) baseline econometric specification strategy as follows:

$$HL_{it} = \beta_0 + \beta_1 \cdot CT_{it} + \delta_0 \cdot Z_{it} + \alpha_i + \tau_t + \epsilon_{it} \tag{1}$$

The dependent variable HL_{it} is the health state (including physical self-evaluation and mental depression level) of respondent i in year t. The key variable CT_{it} indicates the one-way commute time of the respondent. The vector Z_{it} is a set of demographic correlates, including gender, age, urban-rural status, year of education. α_i denotes individual fixed effects. τ_t indicates year fixed effects. ϵ_{it} is the error term. Since I use the panel data, two-way fixed effects model is applied to control the individual heterogeneity.

Because the one-way commute time variable CT_{it} is reported by the respondents themselves, it not obeys a normal distribution and there are a lot of outliers unavoidably. Meanwhile, there exists lots of zero-valued observations which are meaningful in economics, taking the logarithm of CT_{it} is not suitable ($ln(0)$ is undefined) as a result. I apply the Inverse Hyperbolic Sine (IHS) transformation to approximate a normal distribution. IHS is popular among applied econometricians these years because it can reduce the effect of outliers and allows retaining zero-valued observations at the same time[10].

The key transformation equation is (x means the one-way commute time here):

[10]For more information and mathematics deduction about the IHS transformation, please refer to the Bellemare and Wichman (2020).

$$\tilde{x} = \text{arcsinh}(x) = \ln\left(x + \sqrt{x^2+1}\right) \tag{2}$$

Using IHS transformation, I generate a new variable ihs_CT_{it}. The coefficient of ihs_CT_{it} should be interpreted as the elasticity estimates.

Table 3 reports the regression results of OLS model, Panel A and B presents the results with the dependent variables *Self-Evaluated Health State* and *Mental Health Score* respectively. For both panels, Column (1) and (2) only contains one-way commute time CT_{it}, while other Columns add some control variables including demographic controls (one's age, gender, urban-rural status and education years) and habits controls (smoking, alcohol and noon break). Column (1) doesn't apply fixed effect model, while other five Columns both control individual and year fixed effects to reduce the heterogeneity. In Column (5) and (6), I replace the CT_{it} variable with ihs_{CT}, the coefficients are the elasticity respectively.

According to the regression results, the negative relationship between health state (both physical and mental) and one-way commute time estimated by OLS model is statistically significant ($p - value < 0.01$). Longer commuting indeed means worse physical health and higher probability to get depressed. One minute longer in commuting can increase the SEHS level reported by respondents about 0.001 and increase the MHS level about 0.01. Using the interpretation of IHS transformation, the conclusion is the elasticity between SEHS and one-way commute time is about 0.03 while the elasticity between MHS and one-way commute time is about 0.34. This can be interpreted that given a 1% increase in the commute time, the percent change in SEHS and MHS will increase 0.03% and 0.34% respectively.

4.2 Staggered DID Model

It is easy to find that the above baseline specification strategy assumes the health state of respondents are strictly exogenous. However, there are several endogenous problems hidden behind the baseline equation (1). i. Omitted Variables. Unobserved or unidentifiable factors can affect both person's health state and commute time, such as the health awareness of each observation. If a person cares more about his/her health level, then he/she perhaps prefers the job with less time for commuting and expenses more on medical services than others to improve the health state. ii. Measurement Error. The commuting time and the health state data (especially the physical self-evaluated health state) rely on the interviewee's recall and memory, which may be not so accurate and will unavoidably lead to measurement errors in raw data, but assuming the error purely random is reasonable here.

To reduce above possible biases, I consider the opening of metro system in a city as a natural experiment (policy treatment) to identify the casual effect between commute and health more clearly. I assume that the opening of metro system is exogenous enough to be a suitable policy shock. After the opening of metro system, commuters have more choices for commuting, the high speed and plenty of stops of metro system can make commuting more convenient, I can assume that the commute time will reduce to some extent accordingly. However, commuters may know the opening and operation information of metro systems in one city in advance, which can influence their choices about living and working places. But in my research, the application of panel data can control commuters who live in the same city, the migration behaviors will be excluded in the final results[11].

The staggered DID model is based on the DID model and it is an extension version of standard DID model. Standard DID assumes the natural experiment affects all individuals at the same time (two-period set-up), while in real world, many policies are implemented in a staggered fashion —they affect different individuals/areas and at different time periods (e.g. multi-phase policy programs), thus the "treatment status" varies by both i and t (panel data set-up). In this case, the staggered DID model should be applied to estimate the treatment effect —it is also called time-varying DID, multi-period DID, event study, etc.

[11] However, if a commuter chooses to migrate (moving the house) within a single city (with the expectation that new metro system will be put in operation in the near future), this bias can not be eliminated now because CFPS doesn't include detailed residential information and I can not identify the specific places those respondents have moved to and the change of their commute distance or time, but most of the respondents have not move according to the panel data.

Because the building and opening of metro systems has its effect to different cities in different year, which is a typical multi-period policy shock. I use the staggered DID model as follows to estimate the treatment effect:

$$HL_{it} = \beta_0 + \beta_1 \cdot subway_{it} + \delta_0 \cdot Z_{it} + \alpha_i + \tau_t + \epsilon_{it} \tag{3}$$

The dummy $subway_{it}$ allows individuals to be treated at different time periods (time varying treatment), if the city the respondent i living in has metro system in operation in time t, then $subway_{it}$ will be 1, otherwise the value of $subway_{it}$ will be 0. The meanings of other components in Equation (3) remains the same as Equation (1) (Z_{it}: control variables; α_i: individual FE; τ_t: time FE; ϵ_{it}: error term).

Table 4 reports the staggered DID regression results. I use the key explanatory variable $subway_{it}$ denoting whether the city where the person lives has its metro system in specific year. I also use two dependent variables (SEHS and MHS) to identify the casual effect after excluding possible biases, which is convenient for comparasion meanwhile. In Table 4, Column (1) to (3) represent the dependent variable $SEHS$ while (4) to (6) representing MHS. Column (2), (3), (5), (6) control demographic factors of individuals while Column (3) and (6) add life habits controls. All of six columns control two-way fixed effects (TWFE).

From the estimation results, the shortening of commute time (after the opening of metro system) has negative effect on self-evaluated physical health state, but the estimates are insignificant, which is out of my expectation and a little counter-intuitive. Another unexpected result is the opening of metro system increases the mental CESD scores about 1.1 points, which is quite significantly ($p - value < 0.01$). The implication in staggered DID regression results may be that shorter commute time cost doesn't has obvious benefit or improvement for individuals' physical health while depresses their mental state on the contrary. In fact, the results are still influenced by many other unobserved factors, a point is that the comfort level of different commute approaches will have a large effect on commuters' physical and mental health[12].

5 Discussions

5.1 Robustness Check

Because the dependent variable *Self-Evaluated Health State* is a categorical variable, including five levels: poor (5 points), fair (4 points), good (3 points), very good (2 points), quite good (1 points). I use a multinominal model to estimate and predict the probability of an individual's health state level mainly according to his/her one-way commute time. The five levels have an order, therefore, I employ Oredered Logit Model (more information about the principle of this model is in Appendix) as robustness check specification.

Let the SEHS be dependent variable y, which has five different ordered value (1 to 5), I set up the model as below:

$$\begin{aligned} P_i = P(y=i) \quad L_K &= \mathrm{Ln}\left(\frac{P(y \leqslant k)}{P(y \geqslant k+1)}\right) \qquad i = 1,2,3,4,5 \\ \text{Logit} \; &= \mathrm{Ln}\,\frac{p_1 + \cdots + p_j}{1-(p_1+\cdots+p_j)} = \alpha_j + \sum \beta_j x_j, j = 1,2,3,4,5 \end{aligned} \tag{4}$$

The result of Ordered Logit Regression is reported by Table 5. According to the estimates of coefficients, the prediction model can be written as follows (all of the coefficients are highly significant, $p - value < 0.01$):

[12] According to some research outcomes, commuting by subways is not so salutary: high noise levels can harm hearing (cause hearing loss) (Lee et al., 2017); terrible ventilation and high indoor particle matters (PM) concentrations will be detrimental to individual's respiratory system (He et al., 2018); and the closed environment and depressing atmosphere will let commuters' mood down. These characteristics are specific for metro system, hence, the possible positive health effects may be offset by above negative effects, leading the estimates biased and underestimated.

$$\begin{aligned}
&\text{Logit} = \text{Ln}\,\frac{\mathrm{p}_1 + \cdots + \mathrm{p}_\mathrm{j}}{1-(\mathrm{p}_1 + \cdots + \mathrm{p}_\mathrm{j})} = \alpha_\mathrm{j} + \sum \beta_\mathrm{j}\mathrm{x}_\mathrm{j}, \mathrm{j} = 1,2,3,4 \\
&\alpha_1 = -0.358 \quad \alpha_2 = 0.809 \quad \alpha_3 = 2.555 \quad \alpha_4 = 3.561 \\
&\sum \beta_\mathrm{j}\mathrm{x}_\mathrm{j} = 0.079\text{subway} + 0.040\text{age} - 0.127\text{gender} + 0.040\text{urban} - 0.018\text{eduy}
\end{aligned} \tag{5}$$

Table 5 also presents the odds ratio of key predictors, the odds ratio of *subway* is 1.082 ($p-value < 0.01$), meaning that the respondents living in cities with metro systems will have higher probability to report higher SEHS level, corresponding to worse health state.

As for the staggered DID model, I replace previous dependent variable with the satisfaction to the job (labelled as "qg401" in CFPS), to do robustness check first. After controlling demographic and life habits variables, the coefficient is 0.114 ($p-value < 0.05$) under TWFE, which is significant enough to believe the shortening of commute time will give someone a positive stimulus to his/her job.

For the staggered DID model, parallel trend and dynamic effects of treatment can be tested by including leads and lags (Jacobson et al., 1993):

$$y_{it} = \beta_0 + \sum_{t=-q}^{-1} \beta_t \cdot \text{ Treat }_{it} + \sum_{t=0}^{m} \beta_t \cdot \text{ Treat }_{it} + \delta_0 z_{it} + \alpha_i + \tau_t + u_{it} \tag{6}$$

Figure 1 shows the result of parallel test using above Event Study Model. The figure illustrates that, after the opening of metro systems, the MHS increases significantly. Before the policy shock, the estimates are insignificant (the 95% confidence interval intersects with the 0 line) in both periods. While after the shock, the estimates become significantly positive even if I still can't reject the assumption of significance from the estimates of first (1_{st}) and fourth (4_{th}) period after the policy shock, which reduce the confidence for the solidness of parallel trend.

As a result of the limited periods in this research (only five waves), the parallel test's graphic result is not convincing enough. As a supplement, I conduct joint significance test for ex ante and ex post trend respectively. The F-test value of joint significance test for leads is 0.089 and $p - value > 0.1$, meaning the parallel trend is solid enough before the policy shock. While the F-test value of joint significance test for lags is 3.701 and $p - value < 0.01$, implying the ex post trend probably exists. In short, the staggered DID model in Section 4.2 can be considered to pass the parallel trend test.

To analyze whether the treatment of metro opening has a solid effect on the mental health state of commuters, I conduct placebo test using a typical strategy of "Randomly Generate Experimental Group". In this research, for all observations, there are 12,014 living in cities with metro systems, about 13.12% of total sample. Therefore, I choose 12,014 observations randomly (both by city and by year) as treatment group and generate a new binary variable, $subway\_new_{it}$), to denote if the individual is treated (1 means with subways, 0 means no subways). Then I repeat the regression in Section 4.2 500 times, and combine all of the estimates together to draw a graph.

Figure 2 illustrates the result of placebo test graphically. The coefficient estimated in Section 4.2 is denoted by a red dotted line at 1.114 (with all of the controls and TWFE). While the estimates regressed by "Pseudo Policy Dummy Variable ($subway\_new_{it}$)" (the blue dots) nearly follow a normal distribution. Moreover, the p-values of most estimates are more than 0.1, meaning the estimates are insignificant statistically. In a nutshell, it is nearly impossible to get such significant results in Section 4.2 by accident, the effects from other policies and random factors can be ignored. The result is in line with the expectation.

5.2 Heterogeneous Analysis

To find out whether the casual effects have different characteristics in different groups of commuters, I employ regression by groups mainly according to age (three groups: "Young" is 16 to 35 years old, "Middle-Aged" is 36 to 50, while "Old" is 51 to 65) and gender (two groups: Male and Female). Table 6 shows the results of heterogeneous analysis, to make the table looking cleaner and neater, I omit the number of observations and R-Square values in each regression of both panels. In addition, Panel A chooses the same controls (both demographic and life habits) and TWFE model as the Panel B, which are also omitted in Table 6. Comparing the estimates of different groups, it can conclude that: i. the effect metro opening on the physical self-reported health state is insignificant for any gender or any age group; ii. the effect on mental health is quite significant for male and middle-aged group, men's MHS will increase about 1.78 ($p-value < 0.01$) after the metro opening while middle-aged commuters' MHS will increase about 1.87 ($p-value < 0.01$) on average. But the effect is totally insignificant for other groups.

The gender heterogeneity may be related to the difficulty to switch the working and residential places, which is different for women and men. According to the result from Roberts et al. (2011), the detrimental effect on the psychological health of women is more significant than men because of women's greater sensitivity and larger responsibility for day-to-day household tasks, including childcare and housework. Hence, facing the shock of the opening of metro system, men will try to get higher salary and decrease their rent costs, they have stronger tendency to move to further places to live and will be more probable to choose to take the subway. Longer distance will worse their mental health. Nevertheless, women will balance between work and family more, they may be less interested in changing their jobs and migrate to other places. Migration within the same city is a possible reason, just like the footnote 11 said.

As for the age heterogeneity, old commuters generally don't change their commute approach even though the metro system opens, they often have difficulty to adapt the fast pace and crowded atmosphere while young commuters' adaptability will be much higher than other groups. Therefore, the middle-aged group will have to transfer to the new approach of metro system with lower adaptability towards the less comfortable environment of subways (refer to footnote 12), leading to the significant negative effect on their mental health scores.

5.3 Possible Mechanism

In this part, I take advantage of other information extracted from the questions in CFPS questionnaire to explore possible mechanism behind the somewhat weird casual relationship. I employ five complementary regressions, using five new explained variables: i. the duration of entertainment (mainly on TV and movies, per week); ii. the duration of noon break (in minutes); iii. the duration of sleeping (per workday); iv. the frequency of doing sports (times per week) and v. the time taken on chores/housework (per workday).

Table 7 reports the results of these regression. The coefficients in Column (2), (3) and (5) are insignificant ($p - value > 0.1$), meaning that metro opening has not significantly changed commuters' duration of noon break, the time taken on sleeping and chores. While the individuals living in cities with metro have less entertainment and higher frequency of exercising. Statistically, the time for entertainment each week will decrease about 0.6 hours ($p - value < 0.1$), more than half an hour. Hence, the shortening of entertainment (which is an effective approach to relax oneself and release someone's pressure) may provide a reasonable explanation for the negative effect the metro opening on mental health in Table 4. Meanwhile, people living in cities with metro systems have higher (about 0.3, $p - value < 0.01$) frequency of exercising. This provides new evidence for a possible channel, it seems reasonable that commuters' physical health level indeed get improvement not only by less commute time but also by using the time saved to do more exercises (even though the offset effect mentioned in the interpretation part of Section 4.2 can still not to be ignored).

5.4 A Modified Model

There are lots of mathematics models in urban economics which can be applied in the analysis of commuting.

The earliest and most famous one is called Monocentric Model (or AMM Model, named by the capital letters of three economists), whose assumption is a city has sole central business district (CBD) and workers will economize on commuting and trade off the cost of renting and commuting against the salary (Alonso, 1964; Mills, 1972; Muth, 1969).

Apart from the "sole CBD" assumption, AMM model also assumes workers in this city have completely the same preference and residential conditions. The cost for commuting from residential places to CBD can not be neglected for workers in AMM model.

According to the main results and conclusions of this research, I will provide a modified version of classical AMM model by adding a new parameter λ in the consumers' utility function. λ measures the negative utility of commuting (both for physical and mental health) per unit distance (kilometer for instance). The modified model[13] can be written as follows:

$$\begin{aligned} \max \quad & Z^{1-\alpha}H^{\alpha} - \lambda x \\ s.t. \quad & w \leq tx + HR(x) + Z \end{aligned} \tag{7}$$

Applying the Lagrange method to transform this optimization question into a series of equations conditions:

$$\begin{aligned} & L = Z^{1-\alpha}H^{\alpha} - \lambda x + \xi(w - tx - HR(x) - Z) \\ & \frac{\partial L}{\partial Z} = (1-\alpha)Z^{-\alpha}H^{\alpha} - \xi = 0 \\ & \frac{\partial L}{\partial H} = \alpha Z^{1-\alpha}H^{\alpha-1} - \xi R(x) = 0 \\ & \frac{\partial L}{\partial x} = -\lambda - \xi t - HR(x) = 0 \\ & \frac{\partial L}{\partial \xi} = w - tx - \xi HR'(x) - Z = 0 \end{aligned} \tag{8}$$

Because the function form of $R(x)$ is not given, it is difficult to solve above system of equations.

[13]In this model, the consumers' utility function is $U(Z, H, x)$, H is the housing (or area) good, closer to CBD meaning a higher payment, Z denotes all the other goods whose price have been standardized as unit 1 (using Cobb-Douglas utility function, $H > 0$ and $Z > 0$). Each consumer/worker can get a salary of w for working at CBD, the commute distance is x, the rent price for per unit housing is denoted as a bid-rent function $R(x)$.

To compare the results of new version and the model without λ, I employ MATLAB to do a simple simulation. The given parameters are: $\alpha = 0.2$, $w = 10000$, $t = 10$, $\lambda = 5$. The bid-rent function is diminishing with decreasing marginal price $R(x) = p + m * e^{-x}$, where $p = 500$, $m = 1500$[14].

The result is: in modified model, the optimal commuting distance is 5.02 km while in traditional model the result is 6.39 km[15]. This shows that workers will reduce the commute distance and time to get better physical and mental health state, which can be a part of their better utility, even though they may bear a higher rent price. This also embodies the rationality and importance to include the negative utility of commuting in AMM model.

In reality, extreme commuting shows that there are many other factors influencing commuting, the actual commute time and distance will be much greater than that predicted by AMM Model (Hamilton and Röell, 1982). In 21th century, Lucas and Rossi-Hansberg set up a new endogenous Model called LRH Model based on the externality (Lucas and Rossi-Hansberg, 2002; Lucas Jr et al., 2001; Rossi-Hansberg, 2004). In LRH model, the monocentric assumptions are relaxed and the free bidding for firms and families in different regions of a city is leaded in, which is an extraordinary progress.

6 Conclusion

This paper analyzes the health effect of commuting using CFPS data. The results shows the health effect of commuting exists and cannot be ignored. By using the metro opening as a policy shock, the physical health effect is offset to some extent while the mental health state even worse by this new approach of commuting, which is solid under different robustness checks. This research provides some heterogeneity analysis and possible mechanism, the middle-aged male workers will be effected most while the exercise frequency and the time taken on entertainment, changed by the opening of metro system, can be possible mechanism.

[14] The distance is in kilometers by default.

[15] The process of solving the traditional model is almost the same as the modified model, just omitting the red components in Equation 7 and 8.

The modified AMM model proves the importance of considering the negative utility from commuting in economic and urban planning analysis.

There are many shortcomings worthy of improvement for this research. Firstly, the estimates are not solid enough and needed more robustness test and more convincing interpretations. Secondly, the application of big data (which has much more observations and big data is more accurate and reliable than data collected from questionnaire), such as *Amap* or mobile-communication data, is a good potential research direction. Thirdly, the theoretical economics model is worthwhile to be more detailed and complete, combining with the labor market and firm choice model for example.

References

Alonso, W. (1964). Location and land use. Harvard university press.

Bellemare, M. F. and Wichman, C. J. (2020). Elasticities and the inverse hyperbolic sine transformation. *Oxford Bulletin of Economics and Statistics*, 82(1):50–61.

Brodeur, A., Gray, D., Islam, A., and Bhuiyan, S. (2021). A literature review of the economics of covid-19. *Journal of Economic Surveys*, 35(4):1007–1044.

Cohen, A. J., Anderson, H. R., Ostro, B., Pandey, K. D., Krzyzanowski, M., Künzli, N., Gutschmidt, K., Pope III, C. A., Romieu, I., Samet, J. M., et al. (2004). Urban air pollution. *Comparative quantification of health risks: global and regional burden of disease attributable to selected major risk factors*, 2:1353–1433.

Davis, K. (1956). The amazing decline of mortality in underdeveloped areas. *The American Economic Review*, 46(2):305–318.

Hamilton, B. W. and Röell, A. (1982). Wasteful commuting. *Journal of political economy*, 90(5):1035–1053.

He, S., Jin, L., Le, T., Zhang, C., Liu, X., and Ming, X. (2018). Commuter health risk and the protective effect of three typical metro environmental control systems in beijing, china. *Transportation Research Part D: Transport and Environment*, 62:633–645.

Jacobson, L. S., LaLonde, R. J., and Sullivan, D. G. (1993). Earnings losses of displaced workers. *The American Economic Review*, pages 685–709.

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., and Stone, A. A. (2004). A survey method for characterizing daily life experience: The day reconstruction method. *Science*, 306(5702):1776–1780.

Kain, J. F. (1968). Housing segregation, negro employment, and metropolitan decentralization. *The quarterly journal of economics*, 82(2):175–197.

Lee, D., Kim, G., and Han, W. (2017). Analysis of subway interior noise at peak commuter time. *Journal of audiology & otology*, 21(2):61.

Liu, Dow, W., Fu, and JohnAkin (2004). Health human capital and income growth in china. *China Economics Quarterly*, 4(1):18.

Lu, Y., Shi, X., Sivadasan, J., and Xu, Z. (2019). How does improvement in commuting affect employees? evidence from a natural experiment. *The Review of Economics and Statistics*, pages 1–47.

Lucas, R. E. and Rossi-Hansberg, E. (2002). On the internal structure of cities. *Econometrica*, 70(4):1445–1476.

Lucas Jr, R. E. et al. (2001). Externalities and cities. *Review of Economic Dynamics*, 4(2):245–274.

Mills, E. S. (1972). Studies in the structure of the urban economy. Johns Hopkins Press.

Monte, F., Redding, S. J., and Rossi-Hansberg, E. (2018). Commuting, migration, and local employment elasticities. *American Economic Review*, 108(12):3855–90.

Muth, R. F. (1969). Cities and housing: The spatial pattern of urban residential land use. The University of Chicago Press.

Nie, P. and Sousa-Poza, A. (2018). Commute time and subjective well-being in urban china. *China Economic Review*, 48:188–204.

Roberts, J., Hodgson, R., and Dolan, P. (2011). "it's driving her mad": Gender differences in the effects of commuting on psychological health. *Journal of health economics*, 30(5):1064–

1076.
Rossi-Hansberg, E. (2004). Optimal urban land use and zoning. *Review of Economic Dynamics*, 7(1):69–106.
Sun and He (2022). Home-work separation, time squeeze and enterprises' production efficiency. *China Economics Quarterly*, (4):22.
Sun, Wu, Yin, and Chen (2019). The impact of commuting duration on R esidents' health: Evidence from china family panel studies. *Urban Development Studies*, 26(3):3.
Van de Poel, E., O'Donnell, O., and Van Doorslaer, E. (2012). Is there a health penalty of china's rapid urbanization? *Health economics*, 21(4):367–385.
Wang, Gong, and Li (2008). Healthy human capital, healthy investment and economic growth: a case study of cross provincial data in china. *Management World*, (3):13.
Wang, Liu, and Zhao (2018). Traffic jam and haze pollution: A new perspective based on jobs-housing balance. *Finance & Trade Economics*, 39(1):14.
Wang and Wei (2018). Influences of commuting time on job performance. *Urban Problems*, (7):91–103.
Wu (2017). Resolving "commuting paradox" : How commute time influences subjective well-being. *Acta Psychologica Sinica*, (11):97–107.
Xie, Y. and Lu, P. (2015). The sampling design of the china family panel studies (cfps). *Chinese journal of sociology*, 1(4):471–484.
Zheng, Xu, and Gu (2014). Rethinking "jobs-housing balance" : Providing more choices rather than imposing constraints. *Academic Monthly*, 46(5):11.

Tables

Table 1: Descriptive Analysis

Variable	Definition	With Metro			No Metro		
		Mean (1)	Std. Dev. (2)	Obs. (3)	Mean (4)	Std. Dev. (5)	Obs. (6)
CT	one-way commute time	25.22	24.98	6,063	18.46	21.20	21,043
SEHS	self-evaluated health state	2.91	1.06	11,980	2.92	1.21	79,175
MHS	mental health score by CES-D	31.52	7.18	7,108	33.11	7.81	40,900
age	age in years	42.12	12.01	12,014	42.96	12.26	79,554
urban	urban-rural status (urban=1)	0.75	0.43	11,746	0.38	0.49	77,957
male	gender (male=1)	0.55	0.50	11,941	0.54	0.50	78,924
edu	education years	10.30	4.55	10,999	7.49	4.74	72,476
log(income)	log form of annual household income	10.47	0.96	7,553	9.87	1.10	31,134

Notes: Data Source: China Family Panel Studies, 2012-2020.

Table 2: the Opening Time of Metro System in Different Cities (2012-2020)

	The new cities for metro opening (compared with the last survey wave)
2012	**Shenyang** , **Chengdu**, Foshan, **Chongqing**, **Xi'an**
2014	Suzhou, Kunming, **Hangzhou**, **Wuhan**, Harbin, **Zhengzhou**
2016	**Changsha**, **Ningbo**, **Wuxi**, **Dalian**, Qingdao, Nanchang
2018	Fuzhou, **Dongguan**, Nanning, **Hefei**, Shijiazhuang, Changchun, Guiyang, Xiamen
2020	Urumqi, **Ji'nan**, **Lanzhou**, Changzhou, Xuzhou, Huhhot

Notes: The cities covered by CFPS are marked as font-bold.

Table 3: OLS Regression

Panel A: Self-Evaluated Health State						
	(1)	(2)	(3)	(4)	(5)	(6)
CT	0.001***	0.001**	0.001**	0.001**		
	(0.000)	(0.000)	(0.000)	(0.000)		
ihs_CT					0.033***	0.033***
					(0.010)	(0.010)
Controls_demo			✓	✓	✓	✓
Controls_habits				✓		✓
Individual FE		✓	✓	✓	✓	✓
Year FE		✓	✓	✓	✓	✓
N	23805	23804	21420	21420	21420	21420
R^2	0.001	0.038	0.077	0.080	0.077	0.080

Panel B: Mental Health Score						
	(1)	(2)	(3)	(4)	(5)	(6)
CT	0.007***	0.010***	0.012***	0.012***		
	(0.002)	(0.002)	(0.002)	(0.002)		
ihs_CT					0.339***	0.337***
					(0.064)	(0.064)
Controls_demo			✓	✓	✓	✓
Controls_habits				✓		✓
Individual FE		✓	✓	✓	✓	✓
Year FE		✓	✓	✓	✓	✓
N	23795	23794	21411	21411	21411	21411
R^2	0.000	0.046	0.060	0.062	0.061	0.062

Notes: Standard errors (in parentheses) are clustered at the individual level (* $p < .1$, ** $p < .05$, *** $p < .01$).

Table 4: Staggered DID Regression

	(1) SEHS	(2) SEHS	(3) SEHS	(4) MHS	(5) MHS	(6) MHS
subway	-0.039	-0.005	-0.015	1.031***	1.127***	1.114***
	(0.027)	(0.028)	(0.028)	(0.388)	(0.402)	(0.402)
Controls_demo		✓	✓		✓	✓
Controls_habits			✓			✓
Individual FE	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓
N	91155	82733	81823	48007	43244	43244
R^2	0.045	0.119	0.124	0.062	0.085	0.087

Notes: Standard errors (in parentheses) are clustered at the individual level (* $p < .1$, ** $p < .05$, *** $p < .01$).

Table 5: Oredered Logit Regression

	Odds Ratio	Coefficient
subway	1.082***	0.079***
	(0.020)	(0.018)
age	1.041***	0.040***
	(0.001)	(0.001)
gender	0.881***	-0.127***
	(0.006)	(0.007)
urban	1.041***	0.040***
	(0.005)	(0.004)
eduy	0.983***	-0.018***
	(0.001)	(0.001)
/cut1		-0.358
		(0.028)
/cut2		0.809
		(0.028)
/cut3		2.555
		(0.030)
/cut4		3.561
		(0.031)

Notes: Standard errors (in parentheses) are clustered at the individual level (* $p < .1$, ** $p < .05$, *** $p < .01$), /cut'n' means the respective threshold for five SEHS levels.

Table 6: Heterogeneous Analysis

Panel A: Gender Heterogeneity		
	(1)	(2)
	SEHS	MHS
a. Male		
subway	0.015	1.776***
	(0.037)	(0.543)
b. Female		
subway	-0.061	0.364
	(0.041)	(0.591)

Panel B: Age Heterogeneity		
	(1)	(2)
	SEHS	MHS
a. Young (16~35)		
subway	0.002	0.635
	(0.049)	(0.645)
b. Middle-Aged (36~50)		
subway	-0.013	1.868***
	(0.045)	(0.632)
c. Old (51~65)		
subway	-0.058	0.800
	(0.062)	(0.940)
Controls_demo	✓	✓
Controls_habits	✓	✓
Individual FE	✓	✓
Year FE	✓	✓

Notes: i. Standard errors (in parentheses) are clustered at the individual level (* $p < .1$, ** $p < .05$, *** $p < .01$). ii. To make the table looking cleaner and neater, I omit the number of observations and R-Square value in each regression of both panels. iii. Panel A chooses the same controls and FE model as the Panel B, which are also omitted here.

Table 7: Mechanism Analysis

	(1) Entertainment	(2) Noon Break	(3) Sleep_wd	(4) Sports_f	(5) Chore_wd
subway	-0.631*	-0.489	-0.068	0.275***	-0.116
	(0.360)	(1.235)	(0.051)	(0.086)	(0.104)
Controls_demo	✓	✓	✓	✓	✓
Controls_habits	✓	✓	✓	✓	✓
Individual FE	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓
N	55889	41171	49446	69772	34749
R^2	0.044	0.091	0.048	0.137	0.220

Notes: Standard errors (in parentheses) are clustered at the individual level (* $p < .1$, ** $p < .05$, *** $p < .01$).

Figures

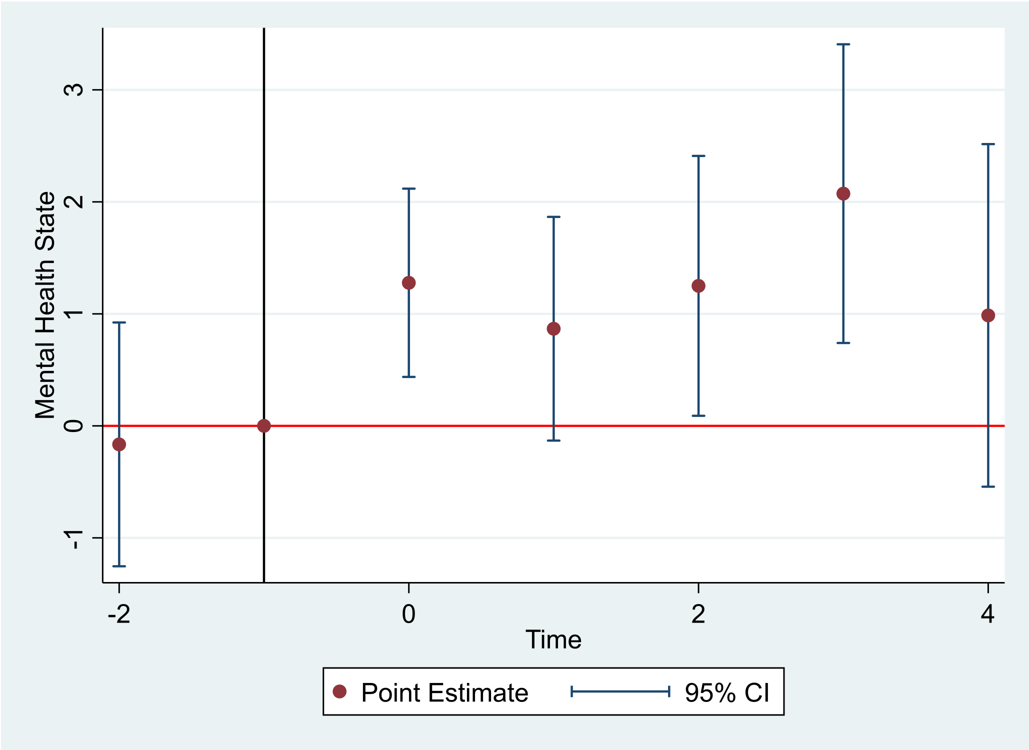

Figure 1: Parallel Test

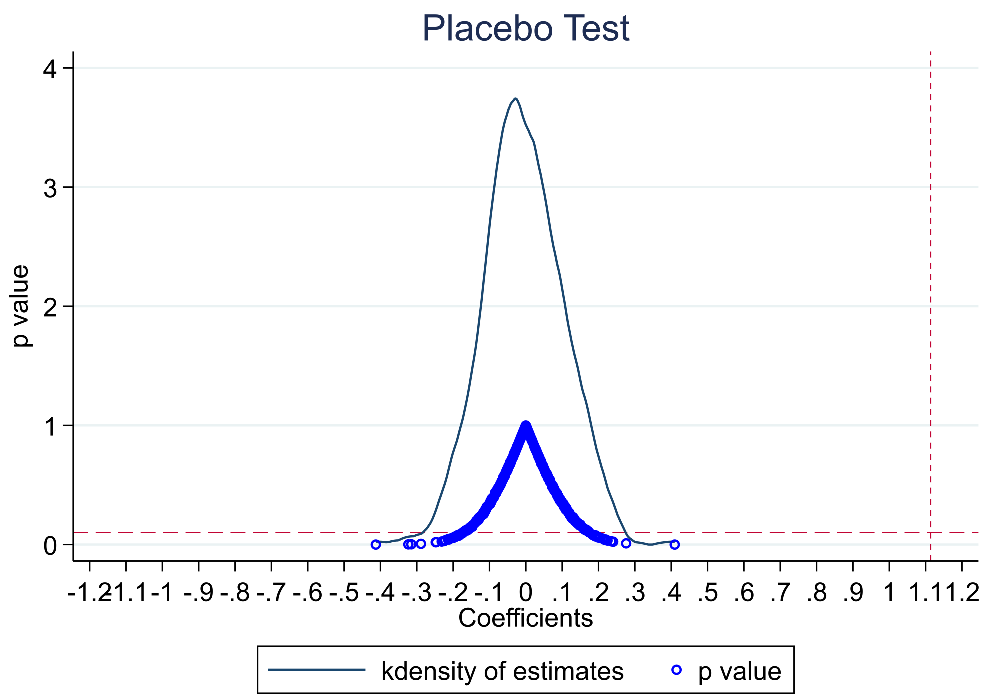

Figure 2: Placebo Test

Ordered Logit Model and Latent Variable Approach[16]

Ordered logit can be derived from a latent-variable model, similar to the one from which binary logistic regression can be derived. Suppose the underlying process to be characterized is:

$$y^* = \mathbf{x}^\top \beta + \varepsilon \tag{9}$$

where y^* is an unobserved dependent variable (perhaps the exact level of individual's physical health state); $\mathbf{x}$ is the vector of independent variables (including CT_{it}); ε is the error term, assumed to follow a standard logistic distribution; and β is the vector of regression coefficients which need to be estimated. Further suppose that while I cannot observe y^*, I instead can only observe the categories of response

$$y = \begin{cases} 0 & \text{if } y^* \le \mu_1 \\ 1 & \text{if } \mu_1 < y^* \le \mu_2, \\ 2 & \text{if } \mu_2 < y^* \le \mu_3 \\ \vdots & \\ N & \text{if } \mu_N < y^* \end{cases} \tag{10}$$

where the parameters μ_i are the externally imposed endpoints of the observable categories. Then the ordered logit technique will use the observations on y, which are a form of censored data on y^*, to fit the parameter vector β.

[16]This appendix part refers to the Wikipedia page of Ordered Logit.