Liquidity Premium and the Substitution Between Money and Treasuries

Wenhao Li*

August 26, 2018

Abstract

This paper quantifies the substitution between money and treasuries, and its implications on the time-series variations in the liquidity premium of treasuries. I estimate a liquidity-in-the-utility model with both money and treasuries. The estimated level of substitution between money and treasuries is well below perfect substitution and contrasts to Nagel (2016). Furthermore, subsample analyses reveal an increasing level of substitution over time, implying the increasing importance of considering treasuries as a form of money. My structural approach identifies the level of substitution, while reduced-form regressions are inconclusive. Results imply that an increase in treasury supply by the government directly affects the treasury liquidity premium, and also partly imposes downward pressure on the nominal interest rate, causing counteractive reactions from the central bank.

Keywords: Liquidity premium; elasticity of substitution; money; treasuries.

JEL classification: E41, E44, E63, G12.

*Graduate School of Business, Stanford University, 655 Knight Way, Stanford. I am grateful to Arvind Krishnamurthy, Sebastian Di Tella, Kenneth J. Singleton, Svetlana Bryzgalova, Monika Piazzesi, and Pablo Kurlat for helpful suggestions.

Electronic copy available at: https://ssrn.com/abstract=3239293

1. Introduction

Money and treasuries both provide liquidity services. An important economic question is whether they provide similar or different liquidity services. The boundary between treasuries and money becomes increasingly blurry as new technology makes transactions with treasury securities much easier. Because when the Fed does open market operation, it changes the composition of money and treasuries, in order to evaluate the impact of such policy, we need to know the level of substitution between money and treasuries. When the government changes its supply of treasuries, it will also cause pressure on the nominal interest rate, given a certain level of substitution between money and treasuries. The central bank thus has to accommodate such shift by open market operations. Furthermore, the building blocks of monetary models have cash-in-advance or money-in-the-utility assumptions, where money is useful for transactions while bonds are not[1]. Thus studying the substitutability between money and treasuries might inform better macro models for monetary policy.

Two recent papers provide answers to this question. On the one hand, Krishnamurthy and Vissing-Jorgensen (2012) demonstrates that treasuries have liquidity premium for their liquidity service to investors just like money, and treasury supply matters for the treasury liquidity premium, which is the yield spread between a safe but less liquid asset and treasuries. Their results are not conclusive in determining whether or not treasuries are substitutable with money because money is not studied in the paper. On the other hand, Nagel (2016) approaches this question more directly and arrives at the conclusion: money and treasuries are perfect substitutes. I revisit the question on the substitutability between money and treasuries. Similar to Nagel (2016), I provide a model with substitution between money and treasuries. Differently, I will structurally estimate the level of substitution and demonstrate empirically that money and treasuries are only partially substitutable.

Suppose money and treasuries are perfectly substitutable[2] as Nagel (2016) finds, then the liquidity premium of treasuries, interpreted as the cost of liquidity service provided by treasuries, should be proportional to the nominal interest rate, which is the cost of liquidity service provided by money. Changes in treasury supply will directly impose pressure on the nominal interest rate, which is fully accommodated by the central bank to keep the interest rate on target. Thus the impact of treasury supply on the nominal interest rate is neutralized by the central bank, and the supply of treasuries loses its explanatory power on the treasury

[1]For a summary of monetary models, refer to Walsh (2017). In the monetary search literature (Lagos and Wright, 2005; Rocheteau and Wright, 2005; Lagos et al., 2017), bonds might be partially used as money for transactional purpose.

[2]Being perfectly substitutable is different from being exactly the same. It is possible that money and treasuries are perfectly substitutable, but each dollar of money provides one unit of transactional service while each dollar of treasuries provides half a unit of transactional service.

Electronic copy available at: https://ssrn.com/abstract=3239293

liquidity premium once we know the level of nominal interest rate.

In contrast, suppose money and treasuries are imperfect substitutes, then the impact of a larger supply of treasuries can be decomposed into two components: (1) Pressure on nominal interest rate, which is neutralized by the central bank and similar to the perfect substitution case. (2) A lower marginal utility of holding treasury, which is the treasury liquidity premium. However, if the Federal Reserve increases the FFR, given that the fiscal authority does not directly respond to interest rate policies, almost all of the movements in FFR are forged into changes in the liquidity premium. If both treasury supply and FFR account for 50% of variations, with a medium level of substitution, we expect that 75% variations in the liquidity premium should be explained by the FFR, while the rest 25% is explained by the Treasury supply. Thus in general the FFR has more explanatory power on the liquidity premium even if treasury supply and FFR have equal amount of variations.

In this paper, I provide a structural approach to estimating the level of substitution between money and treasuries, which subsumes the corner case of perfect substitution. The model has a CES aggregator between money and treasuries with

$$\text{liquidity} = (\lambda_t \cdot \text{treasuries}^\rho + (1 - \lambda_t) \cdot \text{money}^\rho)^{\frac{1}{\rho}}$$

where λ_t is the relative demand for treasuries, and ρ is the level of substitution. Having $\rho = 1$ means perfect substitution and $\rho = 0$ means not substitutable at all. I use GMM to account for the strong interaction effect between FFR and treasury supply in determining the liquidity premium. This new approach provides both more accurate estimations of the substitution effect and higher explanatory power of the liquidity premium. My main finding is that

$$\hat{\rho} = 0.277$$

using monthly data of bank deposits (as a measurement of money) and treasuries after 1959. The estimated model has 23% increase in R^2 in explaining the liquidity premium compared to a linear regression with the same input. To provide further confidence of the estimated level of substitution, I use stationary bootstrap on the residual of liquidity premium based on a model with perfect substitution between money and bonds. Then I find that the likelihood of getting $\hat{\rho} \leq 0.277$ is 0.04%. As a result, it is quite unlikely that the underlying truth is perfect substitution and the GMM estimation reveals a low level of substitutability.

Using yearly data of deposits quantity that covers a longer period of time, I recover the general time trend of substitution. The estimated substitutability is about 0 from 1934 to 1985. Then it increases to 0.344 in the period of 1960-1995. In the most recent period 1980-2016, the substitutability is 0.596. In brief, the substitutability increases over time.

Electronic copy available at: https://ssrn.com/abstract=3239293

The reasons for such a trend might include a more developed repo market where treasuries can be used as "cash" for transactions. Technological progress might make treasuries more money-like in the future.

The GMM estimation results are in contrast to the linear regression results, mainly due to the lack of statistical power of a linear regression to recover an interaction effect between treasury supply and monetary policy. The intuition that a simple linear regression might fail to recover a significant interaction effect is quite general and robust, given that one factor moves at a larger range and a higher frequency than the other. I provide two exercises to illustrate this general intuition, by simulating a model with imperfect substitution between money and treasuries. First, I extract the residual of liquidity premium from predictions of the estimated model and apply stationary bootstrap on the residuals to generate new data samples. The bootstrap is performed for 4000 times to gain statistical power. In a linear regression, the t-stat on the coefficient of treasury supply is small and has a distribution on both sides of zero, implying large estimation errors. However, with a GMM approach, the t-stat on the estimated elasticity of substitution is consistently significant and in the correct direction. Second, I simulate the whole model to generate liquidity premium, only assuming that FFR has a large range of variations and higher frequency of movements. Although the supply of treasury matters for the liquidity premium in the underlying model, a linear specification fails to recover the significance and frequently provide wrong directions, even if the residual error that generates the liquidity premium goes to zero.

In terms of measures, an appropriate definition of money is critical in recovering the substitution between treasuries and money. Importantly, the quantity of money should have a negative relationship with the nominal interest rate so that the nominal interest rate is indeed the cost of holding money. In the data, both checking and savings deposits have a significant and negative relationship with the FFR, while time deposit and money market mutual funds have a significant but positive relationship with the FFR. In fact, households typically invest in money market mutual funds and time deposits to gain access to treasury securities or similar exposure, which leads to a strong and mechanical complementarity between treasuries and these components of a broader money definition. Without time deposits and money market mutual funds, the main components in a broad money definition only includes savings deposits, checking deposits, paper currencies, and bank reserves. Currencies in circulation are mostly held outside the U.S. and by investors without direct access to the treasury securities. In addition, the amount of currency in circulation is much smaller than the amount of checking and savings deposits. Bank reserves, on the other hand, cannot be accessed by households and non-depository institutions. Consequently, I only treat the substitution between saving/checking deposits and treasuries as the economically meaningful

Electronic copy available at: https://ssrn.com/abstract=3239293

object for estimation.

Although treasury securities are held by disperse investors, most variations in treasury supply is directly or indirectly subsumed by households. The major players include foreign investors, mainly sovereign funds, state and local governments, and the Federal Reserve, which together account for about 80% in 2017. However, these investors have very inelastic demand or policy related demand that does not fluctuate frequently (Krishnamurthy and Vissing-Jorgensen, 2012). Households hold treasuries by either direct holding, or through savings bonds or money market mutual funds, which together compose 14% of treasury outstanding. Only less than 3% of treasuries are held by depository institutions. Consequently, the main substitution between treasuries and money are mostly for investors without direct access to reserves, which once again validates the approach to eliminate reserves in estimating the substitution between money and treasuries.

The substitution between money and treasuries is time-varying. In general, with a more active and important repo market in recent decades, we expect treasuries are treated closer to money. My estimations on different subsamples confirm the above intuition: the level of substitution indeed increases over time, with respect to both checking and saving deposits. This time-varying substitution conveys an important message that the structure of the financial market and financial innovations will change the moneyness of treasuries, and blur the boundary between money and bonds.

This paper is related to an emerging literature on the time-variation of the liquidity premium. The liquidity premium rises in distress episodes Longstaff (2002); Brunnermeier (2009), and declines with liquidity supply (Krishnamurthy and Vissing-Jorgensen, 2012; Greenwood et al., 2015). Nagel (2016) connects the liquidity premium to the cost of holding money, which is the FFR, and find that the supply of treasury loses its explanatory power once we control for the FFR. Compared to the previous literature, the main contribution of this paper is to provide a systemic quantification of the substitution between money and treasuries, which sheds lights on how the liquidity premium of treasuries is affected by the FFR and treasury supply. This paper also provides new intuitions on the appropriate methodology to the estimation of the substitution between money and treasuries.

The substitution between treasuries and money is a new topic but it is closely connected to the monetary aggregation literature, which starts as early as Fisher (1922). Interest in this literature diminished after the Keneysian theory became dominant, but revived during the 1980s, evidenced by Barnett (1980), Barnett et al. (1984), Judd and Scadding (1982), and others. The general intuition from this literature is that for different components of a money definition, we cannot aggregate them linearly for economic applications because they are not perfect substitutes. The same intuition applies between treasuries and money

Electronic copy available at: https://ssrn.com/abstract=3239293

studied in this paper.

The rest of the paper is structured as follows. Section 2 presents a simple liquidity in the utility model and empirical specifications. Section 3 structurally estimates the model and provides the level of substitution between money and treasuries. Section 4 discusses why a structural estimation yields dramatically different results than a reduced-form approach and what is an appropriate approach for estimation. Section 5 concludes.

2. A Model of Substitution Between Money and Treasuries

In this section, I will set up a simple model with both money and treasuries. The model is a generalized version of Nagel (2016), which is built on money in the utility framework in the monetary literature. Then I will discuss model properties and how to measure different components of the model.

2.1. Model Setup

The model is composed of households, commercial banks, a government, and a central bank. The objective of the model is to describe the substitution between money and treasuries for the households. As a result, I will not describe the general equilibrium component of how the central bank affects nominal interest rate and how price level is determined. Furthermore, I will assume that the deposit rate i_t^d is a function of the nominal interest rate i_t and other state variables of the economy, but not model the exact mechanism of the connection[3].

A representative household has utility function

$$E_0[\sum_{t=1}^{\infty} \beta^t \left(u(C_t) + \alpha \cdot v(Q_t)\right)] \tag{1}$$

which composes both real consumption C_t and liquidity holding Q_t. The bundled liquidity is composed by both deposit and bonds

$$Q_t = \left((1-\lambda_t)(\frac{D_t}{P_t})^{\rho} + \lambda_t(\frac{B_t}{P_t})^{\rho}\right)^{\frac{1}{\rho}} \tag{2}$$

[3]In the data, for both checking and savings deposits, the deposit spread $i_t - i_t^d$ is well approximated by i_t and time series variations are explained by more than 90% by i_t. The relationship is robust even after the introduction of interest on extra reserves (IOER). One explanation of such a robust connection is the monopoly power of commercial banks (Drechsler et al., 2017).

Electronic copy available at: https://ssrn.com/abstract=3239293

where D_t is the nominal deposits holding, B_t is the nominal bond holding (I will use "bonds" interchangeably with "treasuries" to correspond to the notation B_t in the model), and P_t is the price level. The λ_t component reflects a time-varying demand[4] for bonds compared to deposits. The elasticity of substitution between bonds and deposits is in general more stable than the demand fluctuations, and thus modeled as a constant[5] $\rho \in [0, 1]$. When $\rho = 1$, deposits and bonds are perfect substitutes, and the liquidity bundle is just linear in both deposits and bonds,

$$Q_t|_{\rho=1} = (1 - \lambda_t)(\frac{D_t}{P_t}) + \lambda_t(\frac{B_t}{P_t}) \tag{3}$$

When $\rho \to 0$, deposits and become neither substitutes nor complements,

$$Q_t|_{\rho=0} = (\frac{D_t}{P_t})^{1-\lambda_t}(\frac{B_t}{P_t})^{\lambda_t} \tag{4}$$

The intuition is that in such a Cobb-Douglas aggregator, the portfolio allocation on deposits and bonds are fixed fractions of the total budget, and thus the amount and cost of deposit holding are irrelevant to the amount and cost of bond holding.

Households are subject to the budget constraint

$$\begin{aligned} &\underbrace{D_{t-1}(1 + i^d_{t-1})}_{\text{deposit return}} + \underbrace{B_{t-1}(1 + i^b_{t-1})}_{\text{bond return}} + \underbrace{A_{t-1}(1 + i_{t-1})}_{\text{lending return}} + \underbrace{I_t}_{\text{income}} = \\ &\underbrace{P_t C_t}_{\text{consumption}} + \underbrace{D_t + B_t + A_t}_{\text{investment}} + \underbrace{T_t}_{\text{transfer}} \end{aligned} \tag{5}$$

where i_t is the risk free borrowing and lending rate (we do not restrict household borrowing), i^d_t is the deposit rate, and i^b_t is the bond yield. In each period t, the presentative household collects returns from deposits, bonds, and lending, and earns income. Then the household consumes as well as invests in deposits, bonds, and lending. The final term is a non-distortionary wealth transfer from government to the households, which can be either positive or negative.

From the budget constraint, it is clear that given a level of consumption, by substituting $\epsilon > 0$ amount lending into deposits at time t, the household is earning an additional liquidity benefit $\partial(\alpha v(Q_t))/\partial D_t$, but at the opportunity cost of losing the yield $i_t - i^d_t$ on asset returns. Similarly, by substituting lending into bonds, the household is earning an additional liquidity

[4]The time variations in demand for treasury compared to deposits mainly come from the "flight-to-quality" effects by large investors without direct deposit insurance. In distress periods, the demand for treasuries increases dramatically compared to deposits, driving up the liquidity premium.

[5]In the empirical exercise, I will estimate time variations in ρ by taking different subsamples. However, we cannot estimate a time-varying ρ_t due to the lack of statistical power.

Electronic copy available at: https://ssrn.com/abstract=3239293

benefit $\partial(\alpha v(Q_t))/\partial B_t$, but at the opportunity cost of losing the yield $i_t - i_t^b$. The above intuitions are reflected in the following first order conditions[6] on deposits and bonds for $\rho > 0$:

$$\alpha v'(Q_t)Q_t^{1-\rho}\rho(1-\lambda_t)(D_t/P_t)^{\rho-1} = u'(C_t)\frac{i_t - i_t^d}{1+i_t} \tag{6}$$

$$\alpha v'(Q_t)Q_t^{1-\rho}\rho\lambda_t(B_t/P_t)^{\rho-1} = u'(C_t)\frac{i_t - i_t^b}{1+i_t} \tag{7}$$

The first order conditions for $\rho = 0$ is

$$\alpha v'(Q_t)(1-\lambda_t)\frac{Q_t}{(D_t/P_t)} = u'(C_t)\frac{i_t - i_t^d}{1+i_t} \tag{8}$$

$$\alpha v'(Q_t)\lambda_t\frac{Q_t}{(B_t/P_t)} = u'(C_t)\frac{i_t - i_t^b}{1+i_t} \tag{9}$$

We note that regardless of whether $\rho > 0$ or $\rho = 0$, by dividing the first order condition of bonds over the first order condition of deposits on both sides, we have a general form

$$\frac{i_t - i_t^b}{i_t - i_t^d} = \frac{\lambda_t}{1-\lambda_t}(\frac{B_t}{D_t})^{\rho-1} \tag{10}$$

for all ρ. There are several advantages of using (10) instead of the first order conditions directly for model estimation. First, it is difficult to obtain a good estimation on the marginal utility functions $u(\cdot)$ and $v(\cdot)$. Even the consumption process C_t is not easy to accurately measure. Using (10) solves all these problems. Second, equation (10) is a robust relationship regardless whether $\rho = 0$, while the first order conditions when $\rho > 0$, as in (6) and (7), are different from first order conditions when $\rho = 0$, as in (8) and (9). The government is assumed to directly control total bond supply B_t, while the central bank is assumed to directly control nominal interest rate i_t.

2.2. Data and Measurements

I will estimate a reformulated version of equation (10) throughout the paper:

$$\text{liquidity premium}_t = \tilde{\lambda}_t \cdot (\frac{B_t}{D_t})^{\rho-1} \cdot \text{deposit spread}_t + \varepsilon_t \tag{11}$$

with $\tilde{\lambda}_t = \lambda_t/(1-\lambda_t)$ the relative demand for treasury compared to deposits, and ε_t a measurement error. The information we need to collect includes the quantity of deposits,

[6] Detailed derivations are provided in Appendix A.

Electronic copy available at: https://ssrn.com/abstract=3239293

the quantity of treasuries, the relative preference in holding bonds λ_t (interpreted as the flight-to-liquidity in distress periods), liquidity premium and deposit spread.

The liquidity premium measure should exclude credit risks as much as possible. Thus I will use the spread between 3 month GC Repo rate and 3 month treasury rate.[7] The Repo 3 month term loan is collateralized by treasuries and thus is virtually free of credit risks. However, it is less liquid because the investment has to be locked in for 3 months. Before 1991, the 3 month term repo is not available. The best substitute is the 3 month banker's acceptance, since the Federal Reserve was actively supporting this market for a long history.

The amount of treasury debt is measured as treasuries held by the domestic private investors, because most foreign investors, including sovereign funds, have an inelastic demand for U.S. treasuries, which cannot be considered as substitution towards deposits. Furthermore, the federal reserve holding and intra-government holding of treasuries should not be included as well. By excluding these components, I get a better measure of the effective treasury supply to households who substitute between treasuries and deposits. The data on both foreign treasury holding and Federal Reserve treasury holding date back to 1970s. Since the two components are quite small before 1970s, I will concatenate the above measure with the total debt before 1970 to get a long time series from 1920 to 2016.

Next, I will discuss which components of a broad money definition should be included to study the substitution with respect to treasuries. The call report provides rates on checking account, savings account, and small time deposit account after 1986. As shown in Figure 1, the small time deposit rate in 80% of the period 1986-2016 is above the fed funds rate, raising concerns that it does not provide the transactional services that the checking and savings accounts do. Furthermore, the small time deposit quantity has a strong positive correlation with the nominal interest rate, which is to the contrary of the other two types of deposits. Apart from deposits, money market mutual funds are included in broader definitions of money. Since households purchase MMF shares to gain exposure to treasuries or treasury-like exposure, there is a natural complementarity between MMF and treasuries. Since the focus of this paper is to quantify the substitution between money and treasuries based on the transactional benefit, it is inappropriate to add MMF into the definition of money. Finally, U.S. dollar bills are held mostly outside U.S. by investors without substitution between treasuries and money, and the total amount of U.S. investor held dollar bills is less than 3% of checking and saving deposits. In summary, we will use checking and saving deposits as the definition of money in this paper.

[7] I will use the data from Nagel (2016), and augment it by the 3 month GC Repo rate after 2013. A key reason of not using FFR - Treasury 3 month is because the fed funds rate is contaminated by credit risks of the interbank market, which is relatively large compared to the magnitude of liquidity premium. Furthermore, the fed funds rate is an overnight rate, and the maturity difference weakens the measurement.

Electronic copy available at: https://ssrn.com/abstract=3239293

According to the money aggregation literature (Barnett, 1980; Spindt, 1985; Goldfeld and Sichel, 1990), we need to take into account the substitution between different forms of money. As a result, I will estimate the aggregation between saving and checking as a CES aggregator

$$d_t = 2(\delta(d_{\text{checking},t})^\kappa + (1-\delta)(d_{\text{saving},t})^\kappa)^{\frac{1}{\kappa}} \tag{12}$$

where the multiplier 2 is to make sure that the aggregate quantity does not shrink to a 1/2 mechanically due to definition, e.g. when $\delta = 1/2$ and components are perfectly substitutable with $\kappa = 1$, we expect $d_t = d_{\text{checking},t} + d_{\text{saving},t}$, instead of $d_t = 0.5d_{\text{checking},t} + 0.5d_{\text{saving},t}$.

The methodology to estimate relationship (12) is the same as estimating the substitution between money and deposits. By taking first order conditions on both checking and saving deposits, and divide both sides, we get

$$\frac{i_t - i_{\text{checking},t}}{i_t - i_{\text{saving},t}} = \frac{\delta}{1-\delta}\left(\frac{d_{\text{checking},t}}{d_{\text{saving},t}}\right)^{\kappa-1} \tag{13}$$

Using the same GMM method as detailed in the next section, I get estimation[8] $\kappa = 1$, and $\delta = 2/3$, and the model explains about 84% variations in the checking deposit spread. It is mainly for the internal coherence of the methodology to estimate the elasticity of substitution between checking and savings deposits. A simple aggregation with $\kappa = 1$ and $\delta = 1/2$, which is used in the definition of different money aggregates, result in a similar level of substitutability between treasuries and "money".

Thus we can define aggregate deposits as a weighted average

$$d_t = 2\left(\delta d_{\text{checking},t} + (1-\delta)d_{\text{saving},t}\right) \tag{14}$$

and the aggregate deposit spread as

$$i_t - i_t^d = \frac{d_{\text{checking},t}}{d_t}(i_t - i_{\text{checking},t}) + \frac{d_{\text{saving},t}}{d_t}(i_t - i_{\text{saving},t}) \tag{15}$$

Deposit spread data are available after 1986. However, this short period of data is not enough for the main purpose of the paper to estimate the substitutability between treasuries and deposits, since the quantity of treasuries moves slowly over time, with less variations than the amount of deposits. To extend data availability, one practical way is to project the deposit spread defined in equation (15) to the fed funds rate, which is available back to 1920s. To see whether the approximation is reasonable, I project the monthly deposit rate

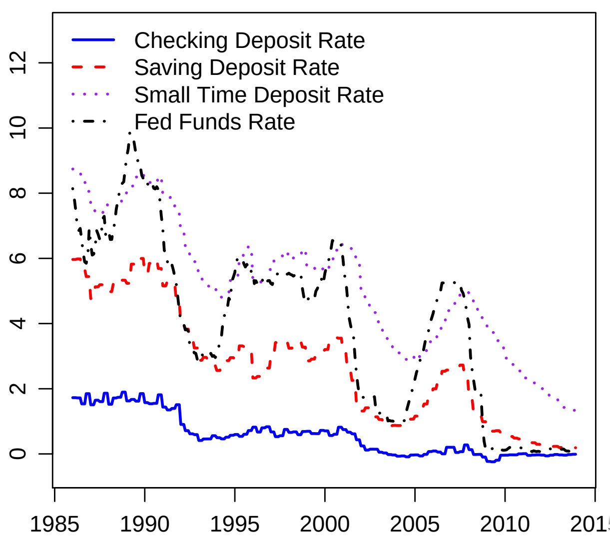
[8]Details of the estimation are provided in Appendix B.2.

Electronic copy available at: https://ssrn.com/abstract=3239293

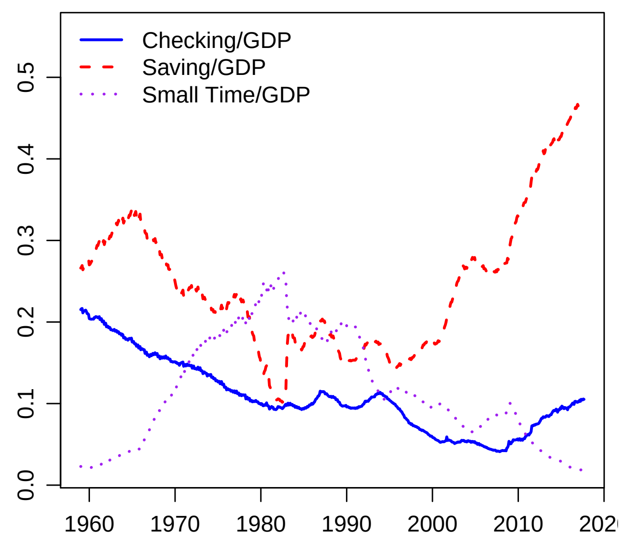
data after 1986 onto the fed funds rate, which results in an approximation

$$i_t - i_t^d \approx 0.57 i_t \tag{16}$$

that explains about 95% time series variations in the deposit spread[9]. Given the high explanatory power, I apply the approximation to extend the deposit spread to 1920s. Then the data limitation comes from deposit quantities, which are available at a monthly frequency starting from 1959 in FRED. One way to extend this quantity is to use the FDIC historical deposits data, which are in yearly from 1934. I find this historical data has similar trends as the FRED data, but different levels in general. So I can only separately use these two types of data to estimate the elasticity of substitution between treasuries and bonds. Given that monthly data is richer, I will use the monthly data from 1959 to estimate the aggregate level of substitutability. But since the yearly data has a longer history, I will use it to study the changes of substitutability over time.



(a) Rates



(b) Volume

Fig. 1. Comparison of Deposit Rates and Volume

Finally, I use VIX index to approximate the flight-to-liquidity premium of treasuries[10]. I

[9]Such high explanatory power is robust after the 2008 crisis with interest on reserve. One explanation for the stable relation is commercial bank monopoly power on deposits, as shown in Drechsler et al. (2017). Since the deposit spread is a reflection of monopoly power, and the monopoly power is relatively stable over time, we expect a reasonably stable relation between deposit spread and the nominal interest rate.

[10]This approach follows Nagel (2016), with the idea that treasuries are fully guaranteed by the government, while checking and saving deposits are not insured when above the FDIC insurance limit. When the economy is full of uncertainty and investors are more risk averse, the "flight-to-quality" effect drives up the demand for treasury, and thus the liquidity premium.

Electronic copy available at: https://ssrn.com/abstract=3239293

will take a form

$$\tilde{\lambda}_t = \frac{\lambda_t}{1-\lambda_t} = \beta_\lambda \cdot \mathrm{VIX}_t$$

and estimate the coefficient β_λ.

3. Model Estimations

In this section, I will estimate the elasticity of substitution between treasuries and deposits, using equation (11). Then I calculate both Type I error and Type II error, which both confirm that the substitutability between money and treasuries is in general very far from 1. Finally, I provide subsample analyses that reveal an increasing substitutability over time.

3.1. Baseline Estimations of the Substitutability

Denote the residual for any parameter (ρ, β_λ) as

$$\varepsilon_t = lp_t - \delta i_t \cdot \beta_\lambda \mathrm{VIX_t}(\frac{b_t}{d_t})^{\rho-1}$$

where I use lp_t to denote the liquidity premium at time t. To get robust estimations and control for the time-series correlations in the residual, I will mainly use the GMM method for estimation. Other approaches, such as nonlinear least squares, yield similar results and are reported in the appendix.

I derive the moment conditions from two sources. The first set of moment conditions come from minimizing the sum of residual error of model fitting,

$$\min_{\rho,\beta_\lambda} \sum_{t=1}^{T} \varepsilon_t^2 \tag{17}$$

As a result, the moment conditions include

$$\frac{1}{T}\sum_{t=1}^{T} \varepsilon_t \cdot \delta i_t \cdot \mathrm{VIX_t}(\frac{b_t}{d_t})^{\rho-1} = 0 \qquad \text{(FOC on } \beta_\lambda\text{)}$$

$$\frac{1}{T}\sum_{t=1}^{T} \varepsilon_t \cdot \delta i_t \cdot \mathrm{VIX_t}(\frac{b_t}{d_t})^{\rho-2} = 0 \qquad \text{(FOC on } \rho\text{)}$$

The second set of moment conditions comes from the simple idea that the residual from

Electronic copy available at: https://ssrn.com/abstract=3239293

model prediction should be close to zero on average,

$$\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t = 0 \qquad \text{(Residual restriction)}$$

The three moment conditions in (FOC on β_λ), (FOC on ρ), and (Residual restriction) result in an over-identified system with one degree of freedom that can be used to apply the J-test for model fitting. A high value of J-statistic, or smaller p-value, implies worse model fitting. The advantage of using such an overidentified system, instead of only using (FOC on β_λ) and (FOC on ρ) as in a nonlinear least square, includes the superior stability and additional testing power on the model itself. Numerically, having (Residual restriction) helps the gmm solver to find global solutions, since the nonlinearity in (FOC on β_λ) and (FOC on ρ) results in local optimums that are hard to escape without additional restrictions.

Then I apply the two-step GMM method and an HAC residual covariance structure of 12 lags.[11] Results are shown in Table 1. We get an estimation of the substitution effect

$$\hat{\rho} = 0.277$$

which is quite far from the perfect substitution scenario of $\rho = 1$. The model fits the data well, since the J-test has a very large p-value, indicating that it is difficult to reject the model. The estimation is quite robust, starting from different initial values.

Table 1: GMM Estimations on the Model in (11)

Parameters	Estimated Value (Volatility)
ρ	0.277 (0.157)
β_λ	0.010 (0.0005)
Observations	696
p-value of J-test	0.818

[11] I implement the GMM using the gmm package in R. The gmm function in that package provides different options in terms of Heteroskedasticity and Autocorrelation Consistent Covariance (HAC) matrix estimations. Options include Truncated, Bartlett, Parzen, Tukey-Hanning, and Quadratic spectral. The first three corresponds to methods proposed by White (1980), Newey and West (1987), Gallant (2009), respectively. The last two are proposed by Andrews (1991). The typical Newey-West is chosen by setting kernel="Bartlett".

Electronic copy available at: https://ssrn.com/abstract=3239293

Table 2: Model Explanatory Power with Different Components

Model Inputs	Fractions of Variations Explained
Only Bonds/Deposits	0.523
Only FFR	0.765
Only VIX	0.480
Bonds/Deposits + FFR	0.802
Bonds/Deposits + VIX	0.573
FFR + VIX	0.807
Full Model	0.851

To provide intuitions on how much each component of the model contributes to the explanatory power, I list the explained fractions of variations with different model components in Table 2, where the fraction of variations is calculated as 1 - residual variance / total variance. With only bonds/deposits data as inputs (FFR and VIX are set as their time series average, respectively), the fraction of variations explained is 52.3%, while other two components individually are also very effective as well. With both Bonds/Deposits and FFR, the model explains 80.2% variations in the liquidity premium. With the full model, the model can explain 85.1% variations in the liquidity premium. As shown in Figure 2, the model generated liquidity premium with all three components as inputs closely tracks the data counterparts. The fed funds rate component of the model in (11) helps explain the ups and downs at yearly frequency, while the quantity of treasury/quantity of deposits modifies the general shape of the predicted liquidity premium and helps explain the liquidity premium at a frequency of several years. Finally, the variations in demand for treasury, approximated by the VIX index, changes at monthly frequency and help explain spikes in the liquidity premium.

Finally, to illustrate the benefit of using a structural model, I compare the R^2 of model predictions and a linear regression of liquidity premium on treasury/deposits ratio, FFR, and VIX. With a linear regression of liquidity premium on FFR, VIX, and log(treasury/deposits), I get an $R^2 = 0.62$. Using either log(treasury/deposits) or treasury/deposits yield the same results. However, both are far below the $R^2 = 0.85$ from the structural model. The 23% improvement clearly shows the power of the model.

Electronic copy available at: https://ssrn.com/abstract=3239293

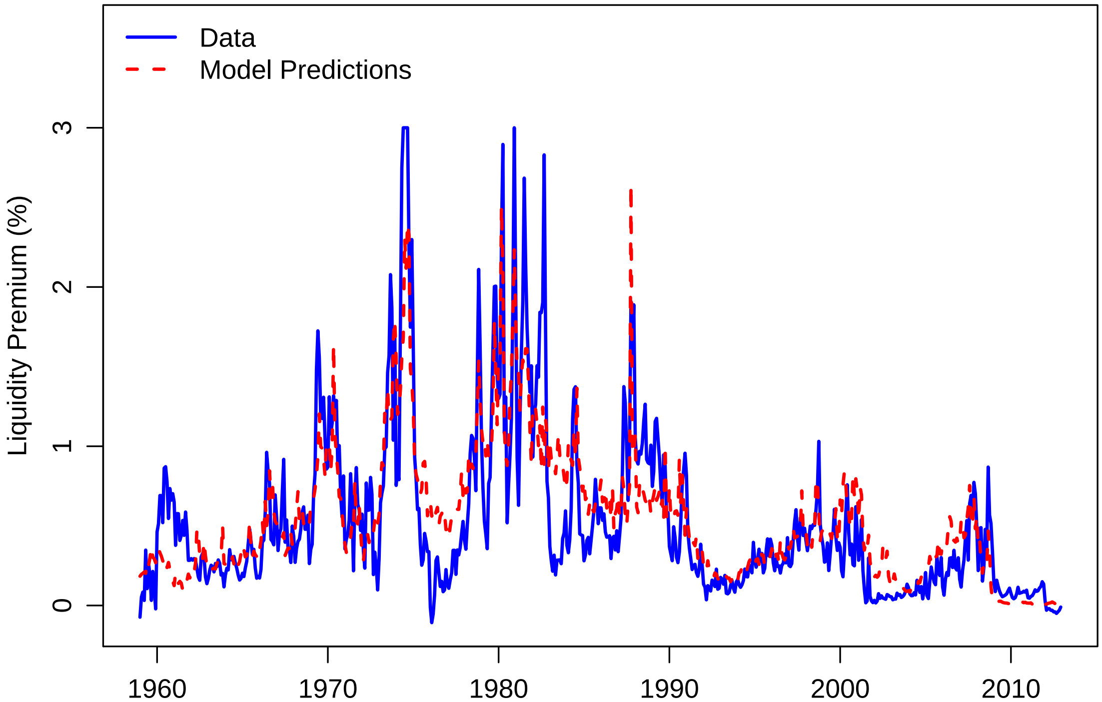


Fig. 2. Model Predicted Liquidity Premium

3.2. Confidence in A Low Level of Substitutability

In order to see how far the estimated ρ is from the perfect substitution, I apply the Wald test on the null hypothesis:

$$H0 : \rho = 1$$

I find a Wald statistic $W = 21$, with p-value=4.4×10^{-6}. Consequently, the model strongly rejects the null hypothesis that deposits and treasuries are perfect substitutes. Furthermore, the estimation that $\hat{\rho}$ is about 0.3 and significantly different from 1 is robust to different estimation algorithms[12], such as nonlinear least squares that directly minimizes the sum of squared errors in (17), as well as an exact-identified GMM approach with only (FOC on β_λ) and (FOC on ρ).

I also check the probability that the actual model is $\rho = 1$, but the GMM estimation wrongly get $\hat{\rho} \leq 0.277$. I use stationary bootstrap on the residuals of the model with $\rho = 1$ for 5000 times, and then calculate the likelihood of getting $\hat{\rho} \leq 0.277$. It turns out that

$$P(\hat{\rho} \leq 0.277|\rho = 1) = 0.04\%$$

which is a very tiny probability. As a result, it is quite unlikely that the truth is $\rho = 1$.

[12]Refer to Appendix B.3 for details.

Electronic copy available at: https://ssrn.com/abstract=3239293

3.3. *Trends of Substitution*

Next, using the yearly data from 1934-2016, I study the time series variations in the level of substitutability. The estimated substitutability is shown in Table 3. From 1934-1970, the substitution is almost zero. The same for 1950-1985. Then from 1960-1995, the substitution is 0.344, and increases to 0.398 for the period 1970-2005. Finally, after 1980, the substitution increases to 0.596. The general trend is quite clear. We also realize that the p-values of J-tests are quite high, indicating close fit of the model to the data.

Table 3: Changes of Substitutability Over Time, Yearly Data from 1934-2016

	Estimated Value (Volatility)				
Parameters	(1) 1934-1970	(2) 1950-1985	(3) 1960-1995	(4) 1970-2005	(5) 1980-2016
ρ	0.019	0.000	0.344***	0.398***	0.596***
	(0.381)	(0.154)	(0.124)	(0.076)	(0.120)
β_λ	0.019***	0.019***	0.017***	0.016***	0.013***
	(0.004)	(0.002)	(0.001)	(0.001)	(0.001)
Observations	36	35	35	35	37
p-value of J-tests	0.847	0.338	0.796	0.432	0.973

Note: *p<0.1; **p<0.05; ***p<0.01

4. Comparisons of Different Approaches

The analyses in the last section reveal that the model implied substitution between treasuries and money (deposits) should be 0.277, which is closer to the non-substitutable case $\rho = 0$ than the fully substitutable case $\rho = 1$. This result contradicts to the claim in Nagel (2016) that money and near money assets are close to perfect substitutes. Since the argument in Nagel (2016) is based on the insignificance of log(treasuries/GDP) in a linear regression on the liquidity premium, the key is to understand the difference between that approach and the structural approach taken in this paper.

There are two main differences. The first one is that I use log(treasuries/deposits) instead of log(treasuries/GDP), and thus input more direct information into the estimation. This reduces noise and improves the significance of the treasury supply. The second one is more important: the nominal interest rate and treasury supply strongly interacts in affecting the

Electronic copy available at: https://ssrn.com/abstract=3239293

liquidity premium. When the interest rate is high, the deposit spread is larger, which implies lower equilibrium deposits holding. Due to the compositional effect of deposit and bonds in the utility as in 2, a lower deposits holding implies a higher marginal value of additional bond holding. Connecting the above, a higher interest rate implies a higher marginal value of bond holding. Given that the marginal value of bond holding is the same of the liquidity premium (interpreted as the marginal cost of bond holding), we expect that the bond holding has a larger marginal impact on the liquidity premium with a higher interest rate. As a result, the interest rate and the bond supply interact to determine the liquidity premium. Since FFR varies from 15% to 0.15%, a 100 times difference that is much larger than the 3 times difference of bond supply in the data, and that the liquidity premium is determined by their interaction, FFR captures most of variations if we only do a linear regression.

In what follows, I will compare three methods to illustrate the differences.

1. Linear regression using bond/GDP: liquidity premium $\sim$ FFR + log(Bond/GDP) + VIX. The same as Nagel (2016).
2. Linear regression using bond/deposit: liquidity premium $\sim$ FFR + log(Bond/Deposits) + VIX. This is an improved version based on the model, with more information from the data.
3. GMM: directly estimate (11) using GMM with moment conditions (FOC on β_λ), (FOC on ρ), and (Residual restriction). This method takes the model exactly as given and extracts the best parameter values that match the data.

To gain statistical power on the three different methods, I will generate new data in two ways, both taking the estimated values $\rho = 0.277$ and $\beta_\lambda = 0.010$ as given. First, I bootstrap the model residual, and then calculate the distribution of estimated values with the three methods. This illustrates whether each method is able to extract the underlying truth, given the right hand side variables in the data. Second, I simulate each element of the model to study the fundamental properties in the data that generate the differences among the three methods, which are about the slow moving quantities and the large range of FFR variations.

4.1. Bootstrap

Given that the underlying model has $\rho = 0.277$ and $\beta_\lambda = 0.010$, I bootstrap the residuals in the model and then summarize the statistical properties of different estimation methods.

Since Nagel (2016) only uses the bond/GDP data that date back to 1920s, not bond/deposits data that only date back to 1959, I extend the bond/deposits ratio measure back to 1920s by setting the bond/deposits ratio as a constant at the 1959 level. This approach guarantees similar length of data sample and makes the exercise more comparable. It is a conservative

Electronic copy available at: https://ssrn.com/abstract=3239293

approach since setting bond/deposits to a constant will be equivalent to eliminating variations in this measure and reduce the statistical power of the model using bond/deposits.

For the specific bootstrap methodology, a simple bootstrap works for i.i.d. residuals. However, in the data, the residual of liquidity premium has a strong time series correlation, although stationarity and weakly dependence seem both satisfied. In the statistics literature, the method designed for solving such problems is the stationary bootstrapping (Politis and Romano, 1994), which preserves stationarity and dot not lose the time series properties, with only the assumption that the time series is stationary and weakly dependent.

The bootstrap is performed for 4000 times[13]. In each time, three methods are applied to extract the impact of the supply variable, which is the coefficient of log(Bond/GDP) in the first approach, the coefficient of log(Bond/Deposits) in the second approach, and $\rho - 1$ in the third approach.

The t-stats in a linear regression used in Nagel (2016) is not significant, as shown in Figure 3, while the t-stats become larger in absolute value when we measure bond/deposits instead of bond/GDP, and much better when we use GMM. The average t-stat is -1.43 with the first linear regression method, -4.60 with the second linear regression, while -6.77 using GMM. Thus we might wrongly conclude that bonds and deposits are almost perfect substitutes based on the insignificance of the supply factor using the first linear regression method, even if the underlying truth is quite far from perfect substitution.

Next, we note that the estimations from first two methods of linear regressions do not reveal the underlying parameter $\rho - 1$, as shown in Figure 4, where the black dotted line shows the underlying value $\rho - 1$ used for simulation. On the other hand, the GMM estimated coefficient is on average -0.72, implying $\hat{\rho} = 0.28$, which is very close to the underlying parameter 0.277. Using the first method, we might even get a positive coefficient on the supply of treasuries, as shown by the above zero cases of Figure 4.

4.2. Pure Simulations

To test weather the key features in the data, including slow moving government bonds and deposits quantities and much wider range of changes in FFR, lead to the insignificance of the first approach, I will simulate all elements of the model in (11) and perserve the key features. In the data, the bond/GDP ratio is very slow moving and has a frequency of about 1 year, which is much slower than the changes in fed funds rate and VIX. Deposits to GDP ratio have a similar frequency. However, FFR changes at a much higher frequency and a wider range. To preserve these features, in the simulations, I set the range of simulated bond/GDP,

[13] It is enough to have 4000 times, and results are similar if I use more rounds.

Electronic copy available at: https://ssrn.com/abstract=3239293

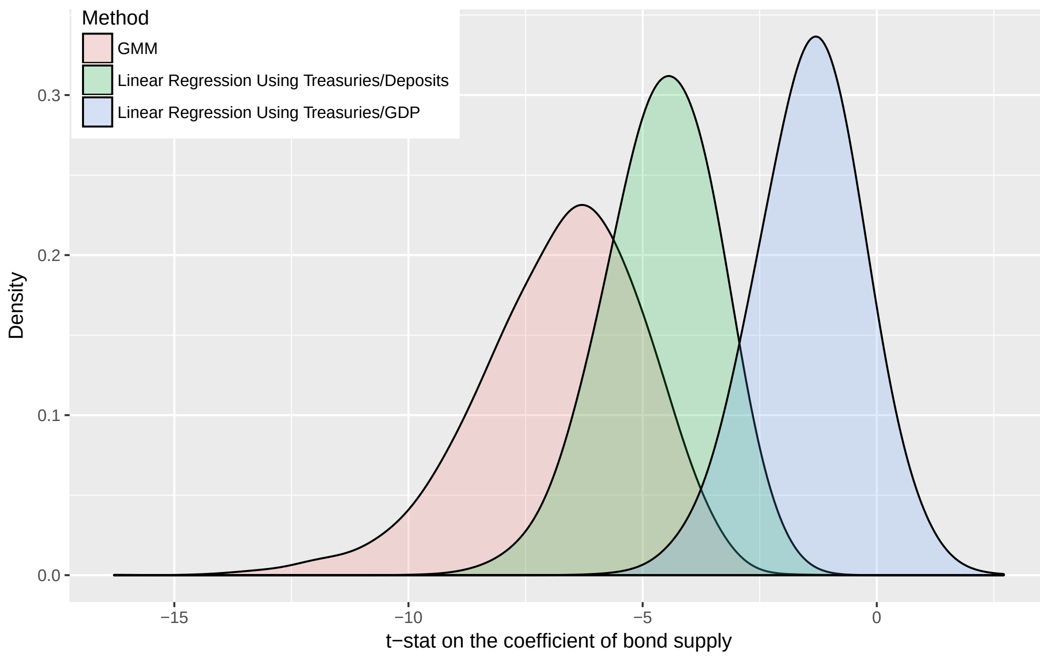


Fig. 3. Density of t-stat in Three Approaches under Stationary Bootstrapping

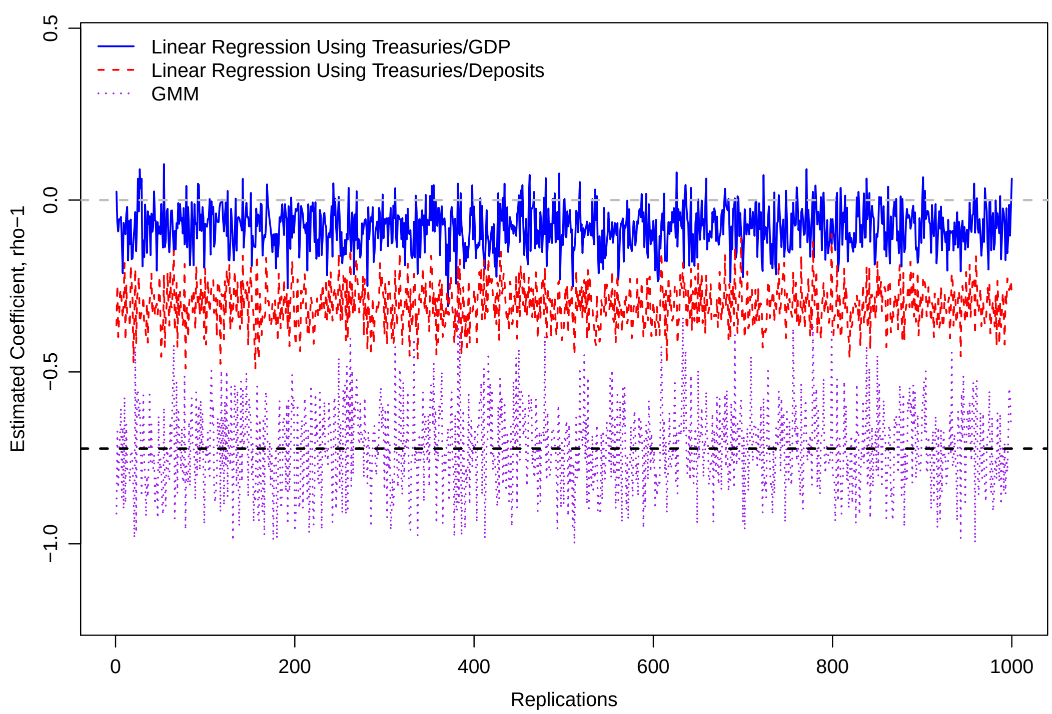


Fig. 4. Comparison of Estimated Coefficients on Bond Supply in Three Approaches under Stationary Bootstrapping

Electronic copy available at: https://ssrn.com/abstract=3239293

deposits/GDP and FFR to the data counterparts. Specifically, for the bond/GDP data, I first generate a yearly time series using

$$x_{1,T+1} = 0.99x_{1,T} + 0.05e_{1,t+13}, \quad T \in \{0, 1, 2, \cdots\}$$

The yearly time series autocorrelation is set to 0.99 to reflect the high series correlation in the data. Then I make the following transformation

$$y_{1,T} = \frac{e^{x_{1,T}}}{e^{x_{1,T}} + 1}$$

to scale the series into $(0, 1)$. Finally, to match the range of movements in the data, I scale the time series $y_{1,\cdot}$ into

$$b_{T,\text{yearly}} = y_{1,T} \cdot \frac{\max(\text{bonds/GDP in data})}{\max(y_{1,\cdot})}$$

Finally, I expand the yearly time series into monthly time series by setting each month the same as the corresponding year value, and get monthly bond/GDP simulated data b_t.

Following the same approach, I generate

$$x_{2,T+1} = 0.99x_{2,T} + 0.05e_{2,t+13}, \quad T \in \{0, 1, 2, \cdots\}$$

and then transform it into monthly deposits/GDP data d_t. Then I get the bond/deposits ratio data b_t/d_t.

Next, the demand λ_t is simulated as the exponential of the following monthly AR(1) process

$$x_{3,t+1} = 0.99x_{3,t} + 0.1e_{3,t+1}$$

$$\lambda_t = e^{x_{3,t}}$$

To get FFR, I simulate

$$x_{4,t+1} = 0.98x_{4,t} + 0.1e_{4,t+1}$$

and then take the exponential of this series to get

$$y_{4,t} = e^{x_{4,t}}$$

The FFR is defined as

$$i_t = y_{4,t}\frac{\max(\text{FFR}_\cdot)}{\max(y_{4,\cdot})} - \min(y_{4,\cdot})\frac{\max(\text{FFR}_\cdot)}{\max(y_{4,\cdot})} + \min(\text{FFR}_\cdot)$$

Electronic copy available at: https://ssrn.com/abstract=3239293

which then results in a time series with maximum value of FFR of about 15, and minimum value of 0.15. The noise part is an AR(1) process with smaller time series correlation,

$$\varepsilon_{t+1} = 0.5\varepsilon_t + \sigma \cdot e_{5,t+1}$$

where σ is a controlled parameter. Finally, I generate liquidity liquidity premium following equation (11), with $\rho = 0.277$.

In all of the above, I assume $e_{i,t}$ are independent through time t and classes i. An alternative way is to estimate the variance-covariance matrix of all variables in the data and then generate random variables based on this covariance matrix. However, since the point of this exercise is to illustrate what features in the data makes a linear regression inconclusive, assuming independent time series is cleaner. Results are quite robust to imposing covariance structures in the data.

Then I apply the three methods in section 4.1 to get estimated coefficients. To be close to the actual regressions, I make sure all the simulated data are monthly and each simulation run has 1000 months. Then I compare three scenarios: (1) High volatility scenario: $\sigma = 0.01$. (2) Medium volatility scenario: $\sigma = 0.006$. (3) Low volatility scenario: $\sigma = 0.002$. By reducing the residual volatility, we are able to compare whether different approaches can converge to the underlying truth as the residual error goes to zero.

In the first scenario $\sigma = 0.01$, estimated t-stat of the bond supply coefficient are shown in Figure 5. We find that the coefficient might go above zero in a linear regression with log(bond/GDP), as found in Nagel (2016), and in general quite insignificant with an average t-stat -1.2. When we use log(bond/deposits) for the linear regression, results are slightly improved and the average t-stat is -1.7. With a GMM regression, the average t-stat is -5.2, and the coefficients are mostly below zero.

Then the residual volatility is reduced to $\sigma = 0.006$ in the second scenario. I find that differences between linear regressions and the GMM estimations are much starker, and there is very little improvement in linear regression results. As shown in Figure 6, the average t-stat of bonds/deposits in a linear regression using log(bonds/GDP) is still -1.2, while the average t-stat $\hat{\rho}$ in a GMM regression is -8.3. If we reduce the residual volatility further to $\sigma = 0.002$, as shown in Figure 7, the average t-stat in a GMM regression becomes -24.8, while the t-stats for linear regressions remain similar.

Finally, comparing the average estimations, the GMM results on average predicts $\rho = 0.277$, and thus reliably recovers the underlying truth.

Electronic copy available at: https://ssrn.com/abstract=3239293

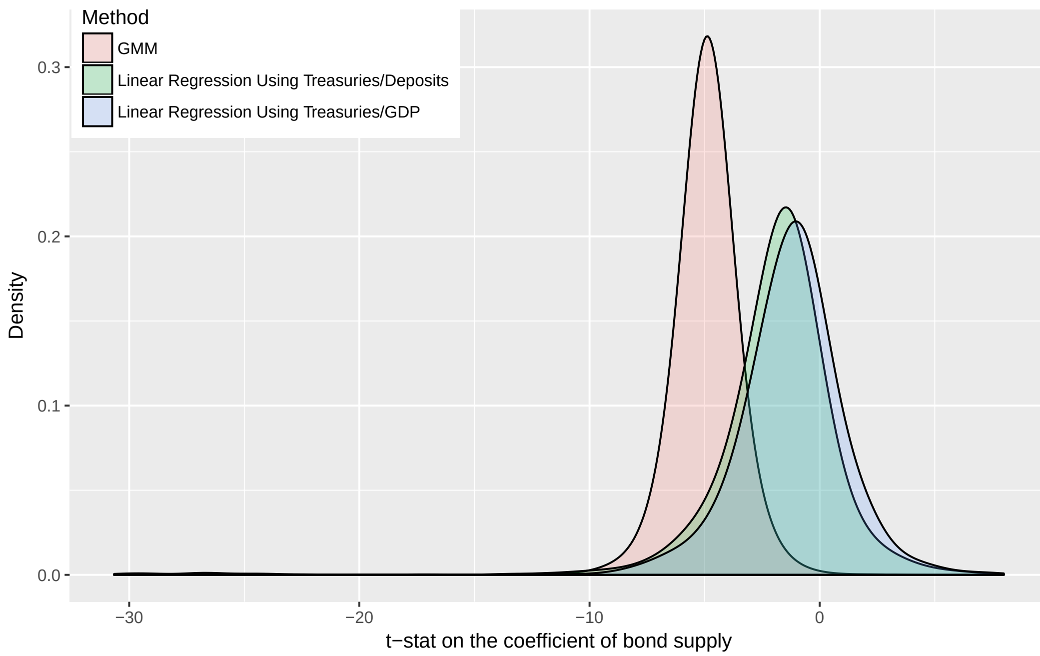


Fig. 5. T-stats of Liquidity Supply Coefficients with Purely Simulated Data, with $\sigma = 0.01$

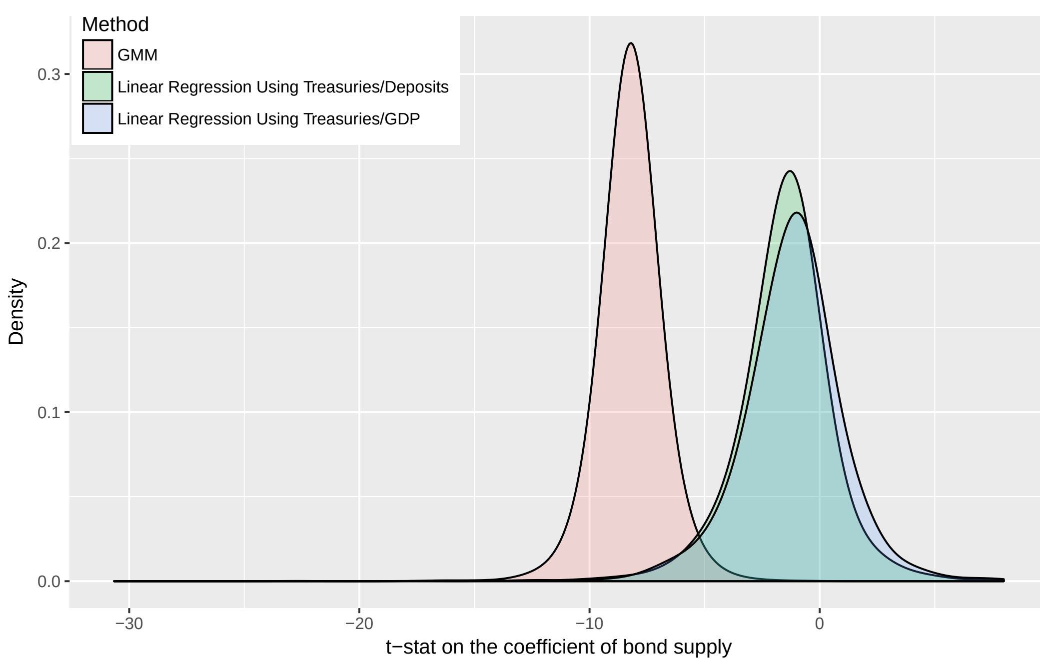


Fig. 6. T-stats of Liquidity Supply Coefficients with Purely Simulated Data, with $\sigma = 0.006$

Electronic copy available at: https://ssrn.com/abstract=3239293

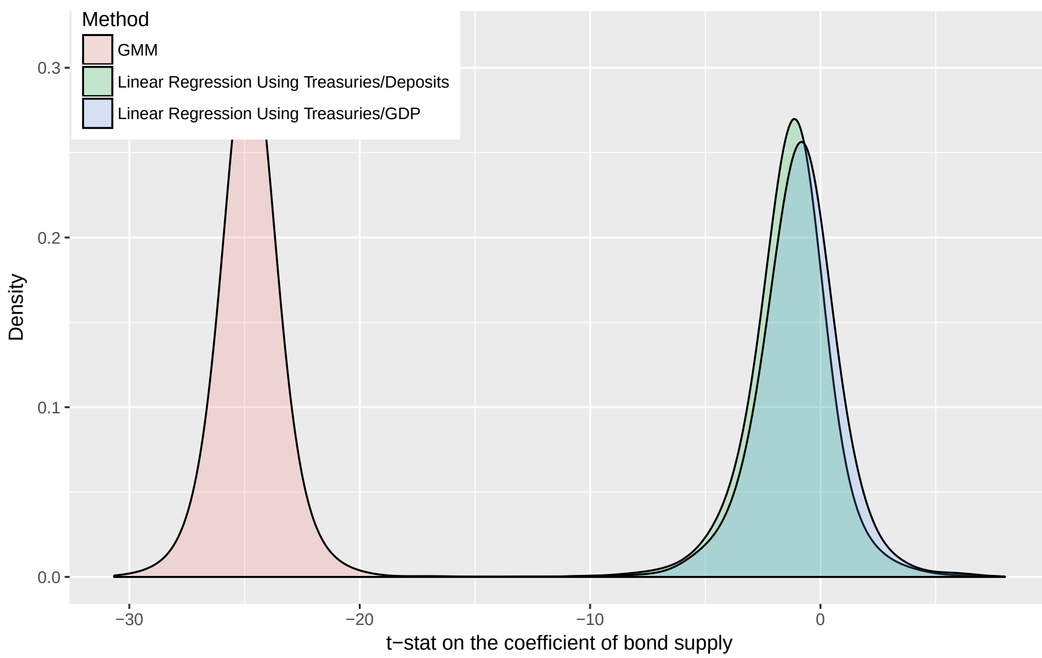


Fig. 7. T-stats of Liquidity Supply Coefficients with Purely Simulated Data, with $\sigma = 0.002$

Electronic copy available at: https://ssrn.com/abstract=3239293

5. Conclusions

In this paper, I quantify the substitution between money and treasuries and find that the level of substitution is well below the perfect substitution concluded by Nagel (2016). The main difference comes from two aspects: (1) I use more relevant information, including deposits quantity and deposits rates. (2) By quantifying the structural model, I take into account the interaction between treasury supply and the FFR. Since the FFR changes at a much higher frequency and a larger range, a linear regression of the liquidity premium on the FFR and log(treasury/GDP) will only recover the significance of FFR, but has larger estimation errors on the treasury supply. The resulting level of substitution is 0.277, which is well below 1, or perfect substitution, and the likelihood of getting a low substitution when the underlying truth is perfect substitution is negligible.

Applying the same estimation methods on subsamples, I discover an increasing level of substitution over time. The substitution increases from close to 0 to about 0.6, from the period 1934-1970 to the period 1980-2016. This is consistent with the technological advances that facilitate transactions with treasuries. This increased level of substitution makes it more important to consider the "moneyness" of treasuries in monetary policy since the effectiveness of open market operations depends on the composition and substitution between money and treasuries. Furthermore, incorporating treasuries in the liquidity component of modern monetary models, such as New Keynesian DSGE models, might result in better model performance.

This paper helps to improve understanding of the treasury liquidity premium, which is a good indicator of financial distress (Longstaff, 2002), connected to risk premia and risk-taking (Drechsler et al., 2018; Li, 2018), and closely related to U.S. exchange rates (Jiang et al., 2018). With a medium level of substitutability between money and treasuries, the liquidity premium is determined both by the supply of treasuries and the nominal interest rate. As a result, we expect that both monetary policy and supply of treasuries by the government have additional influence on the risk premia and risk-taking, as well as exchange rate movements. These interesting questions are left for future research.

Electronic copy available at: https://ssrn.com/abstract=3239293

References

Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica: Journal of the Econometric Society*, pages 817–858.

Barnett, W. A. (1980). Economic monetary aggregates an application of index number and aggregation theory. *Journal of Econometrics*, 14(1):11–48.

Barnett, W. A., Offenbacher, E. K., and Spindt, P. A. (1984). The new divisia monetary aggregates. *Journal of Political Economy*, 92(6):1049–1085.

Brunnermeier, M. K. (2009). Deciphering the liquidity and credit crunch 2007-2008. *Journal of Economic perspectives*, 23(1):77–100.

Drechsler, I., Savov, A., and Schnabl, P. (2017). The deposits channel of monetary policy. *The Quarterly Journal of Economics*, 132(4):1819–1876.

Drechsler, I., Savov, A., and Schnabl, P. (2018). A model of monetary policy and risk premia. *The Journal of Finance*, 73(1):317–373.

Fisher, I. (1922). *The making of index numbers: a study of their varieties, tests, and reliability*. Number 1. Houghton Mifflin.

Gallant, A. R. (2009). *Nonlinear statistical models*, volume 310. John Wiley & Sons.

Goldfeld, S. M. and Sichel, D. E. (1990). The demand for money. *Handbook of monetary economics*, 1:299–356.

Greenwood, R., Hanson, S. G., and Stein, J. C. (2015). A comparative-advantage approach to government debt maturity. *The Journal of Finance*, 70(4):1683–1722.

Jiang, Z., Krishnamurthy, A., and Lustig, H. (2018). Foreign safe asset demand and the dollar exchange rate. *National Bureau of Economic Research*.

Judd, J. P. and Scadding, J. L. (1982). The search for a stable money demand function: A survey of the post-1973 literature. *Journal of Economic Literature*, 20(3):993–1023.

Krishnamurthy, A. and Vissing-Jorgensen, A. (2012). The aggregate demand for treasury debt. *Journal of Political Economy*, 120(2):233–267.

Lagos, R., Rocheteau, G., and Wright, R. (2017). Liquidity: A new monetarist perspective. *Journal of Economic Literature*, 55(2):371–440.

Electronic copy available at: https://ssrn.com/abstract=3239293

Lagos, R. and Wright, R. (2005). A unified framework for monetary theory and policy analysis. *Journal of political Economy*, 113(3):463–484.

Li, W. (2018). Public liquidity supply, bank run risks, and financial crises. *Working paper*.

Longstaff, F. A. (2002). The flight-to-liquidity premium in us treasury bond prices.

Nagel, S. (2016). The liquidity premium of near-money assets. *The Quarterly Journal of Economics*, 131(4):1927–1971.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical association*, 89(428):1303–1313.

Rocheteau, G. and Wright, R. (2005). Money in search equilibrium, in competitive equilibrium, and in competitive search equilibrium. *Econometrica*, 73(1):175–202.

Spindt, P. A. (1985). Money is what money does: Monetary aggregation and the equation of exchange. *Journal of Political Economy*, 93(1):175–204.

Walsh, C. E. (2017). *Monetary theory and policy*. MIT press.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, pages 817–838.

Electronic copy available at: https://ssrn.com/abstract=3239293

Appendix A. Model Derivations

The Lagrangian of the household optimization problem is

$$\sum_{t=1}^{\infty} \beta^t \left(\begin{array}{l} u(C_t) + \alpha \cdot v(Q_t) + \lambda_t \left(-P_t C_t - D_t - B_t - A_t + \cdots\right) \\ +\lambda_{t+1} \left(D_t(1+i_t^d) + B_t(1+i_t^b) + A_t(1+i_t) + \cdots\right) \end{array} \right)$$

The first order condition on consumption is

$$u'(C_t) = \lambda_t P_t \tag{18}$$

The first order condition on lending is

$$\lambda_{t+1}(1+i_t) - \lambda_t = 0 \tag{19}$$

The first order condition on deposit holding is

$$\alpha \frac{\partial v(Q_t)}{\partial D_t} - \lambda_t + \lambda_{t+1}(1+i_t^d) = 0 \tag{20}$$

The first order condition on bond holding is

$$\alpha \frac{\partial v(Q_t)}{\partial B_t} - \lambda_t + \lambda_{t+1}(1+i_t^b) = 0 \tag{21}$$

Combining Equations (18), (19), and (20), we get

$$\alpha \frac{\partial v(Q_t)}{\partial B_t} = \frac{u'(C_t)}{P_t} \frac{i_t - i_t^d}{1+i_t}$$

When $\rho > 0$, by expanding the definition of Q_t, we have

$$\alpha v'(Q_t) Q_t^{1-\rho} \rho (1-\lambda_t)(D_t/P_t)^{\rho-1} = u'(C_t) \frac{i_t - i_t^d}{1+i_t}$$

Similarly, from (18), (19), and (21) and the definition of Q_t, we have

$$\alpha v'(Q_t) Q_t^{1-\rho} \rho \lambda_t (B_t/P_t)^{\rho-1} = u'(C_t) \frac{i_t - i_t^b}{1+i_t}$$

When $\rho = 0$, plugging in the specific form of Q_t, the derivatives are

$$\frac{\partial v(Q_t)}{\partial B_t} = \lambda_t \frac{Q_t}{(B_t/P_t)} \frac{1}{P_t}$$

Electronic copy available at: https://ssrn.com/abstract=3239293

$$\frac{\partial v(Q_t)}{\partial D_t} = (1-\lambda_t)\frac{Q_t}{(D_t/P_t)}\frac{1}{P_t}$$

Thus the first order conditions on deposits and bonds can be simplified as

$$(1-\lambda_t)\frac{Q_t}{(D_t/P_t)} = u'(C_t)\frac{i_t - i_t^d}{1+i_t}$$

$$\lambda_t\frac{Q_t}{(B_t/P_t)} = u'(C_t)\frac{i_t - i_t^b}{1+i_t}$$

Appendix B. Data Construction and Robustness Checks

B.1. Details on Data Construction

B.1.1. Quantities

- Checking deposits: all checking accounts in commercial banks. Data series downloaded from FRED, with identifier "TCDSL", from 1959-2016. Checking deposits under the FDIC insurance threshold are insured by the FDIC.
- Savings deposits: all saving accounts in commercial bankds, including money market deposit accounts. Data series downloaded from FRED, with identifier "SAVINGSL", from 1959-2016. Savings deposits under the FDIC insurance threshold are insured by the FDIC.
- Bonds: treasury holding by private investors (FRED identifier "FDHBPIN"), minus foreign holding (FRED identifier "FDHBFIN"), with both data from 1970 to 2016. Before 1970, the measure is the total amount of federal debt, the same as Nagel (2016).
- Nominal GDP: gross domestic product. For 1947-2016, monthly data are downloaded from FRED with identifier "GDP". For 1929-1946, yearly data are downloaded from FRED with identifier "GDPA", and transformed into monthly data where each month in the same year has the same GDP. Then I concatenate the two data series to get nominal GDP from 1929 to 2016.
- Money at zero maturity (MZM): M2 less small-denomination time deposits plus institutional money funds. Data downloaded from FRED with identifier "MZMNS", from 1959 to 2016.

B.1.2. Rates

- Fed funds rate (FFR): Monthly effective fed funds rate in percentage, with FRED identifier "FEDFUNDS", from 1954 to 2016. From 1920 to 1954, monthly data is from Nagel (2016).

Electronic copy available at: https://ssrn.com/abstract=3239293

- Liquidity premium: From 1991 to 2016, liquidity premium is measured as the yield spread between 3-month Repo collateralized by treasuries and 3-month treasury. Data are downloaded from Bloomberg. From 1920 to 1991, liquidity premium is measured as the yield spread between 3 month banker acceptance and 3 month treasuries, which is from Nagel (2016). Historically, the banker acceptance market was actively supported by the federal reserve.
- Checking deposit rate: checking rate is calculated as the commercial bank checking interest expense over total checking volume, using data from call report. The monthly checking rate is provided by Drechsler et al. (2017) from 1986 to 2013. It cannot date back further because of data limitation from the call report.
- Saving deposit rate: calculated in a similar way as checking deposit rate, and monthly data is provided by Drechsler et al. (2017). As checking deposit rate, the saving deposit rate can only date back to 1986.
- MZM own rate: Weighted average of the rates received on the interest-bearing assets included in MZM. Data downloaded from FRED, with label "MZMOWN", from 1974 to 2016.

B.2. Aggregation Among Saving and Checking Deposits

We estimate the substitution between saving and checking deposits using equation

$$\frac{i_t - i_{\text{checking},t}}{i_t - i_{\text{saving},t}} = \frac{\delta}{1-\delta}\left(\frac{d_{\text{checking},t}}{d_{\text{saving},t}}\right)^{\kappa-1}$$

In the data, we find that the checking deposit spread is well approximated by a constant multiplying the saving deposit spread, which implies that the left-hand-side of the above equation should be a constant, although the ratio of checking and saving is changing over time. Thus a good estimation if $\kappa = 1$.

If we use GMM to estimate the above model, results are quite close to $\kappa = 1$, as shown in Table 4. I will pick the economically meaningful rounding value $\kappa = 1$, and rounded value $\delta/(1-\delta) = 2$, which implies $\delta = 2/3$. Thus we are confident to use the following aggregation:

$$d_t^{\rho} = \delta d_{1,t} + (1-\delta) d_{2,t}$$

and the aggregate d_t^{ρ}, with the weighted spread as

$$s = \delta s_{1,t} + (1-\delta) s_{2,t}$$

Electronic copy available at: https://ssrn.com/abstract=3239293

Table 4: GMM Estimation for the Substitution Between Saving and Checking Deposits

	Dependent variable:
	Estimated Value (Volatility)
κ	1.074***
	(0.018)
$\delta/(1-\delta)$	2.094***
	(0.144)
Observations	336
p-value of J-test	0.06
Note:	*p<0.1; **p<0.05; ***p<0.01

B.3. Robustness of the Estimation Method

In the main text, I have used two-step GMM to estimate an over-identified system with moment conditions (FOC on β_λ), (FOC on ρ). As robustness checks, I will use slightly different methods for estimation. The first one is a nonlinear least square (NLS) with objective function (17). The second one is still GMM, but with only moment conditions (FOC on β_λ) and (FOC on ρ). Since it is not easy to control for time series correlations in a typical NLS package, I use stationary bootstrap (Politis and Romano, 1994) to calculate the standard errors, which guarantees stationarity for the generated time series and also preserves time series correlations.

Results are shown in Table 5. We find that the estimated values are quite similar to what we have found using GMM with three moment conditions. Thus results are robust to estimation methods. The main approach that I use in the paper is chosen because of superior numerical stability and feasibility of J-test for model fit.

B.4. Details of the Bootstrap Estimations

In this section, I provide more detailed results on the bootstrap estimations. In Figure 4, I plot the estimated coefficients under three methods, based on the bootstrap data. As shown in the solid blue line, the linear regression setup used in Nagel (2016) results in coefficient estimations close to zero, although the underlying model implies a negative one. Incorporating deposits into the regression improves results, shown by the dotted red line. However, neither of them is close to the desired value of $\rho - 1$, denoted by the dotted black line. Using the GMM approach that I have proposed provides estimations that are on average

Electronic copy available at: https://ssrn.com/abstract=3239293

Table 5: Estimations with Different Methods

	(1) NLS	(2) GMM
ρ	0.285^{**}	0.280^{**}
	(0.116)	(0.117)
β_λ	0.010^{***}	0.010^{***}
	(0.0004)	(0.0003)
Observations	696	696

Notes: Stationary bootstrap is used to estimate the volatility of coefficient estimation. The first column uses nonlinear least squares (NLS) with objective function (17), while the second column uses generalized methods of moments (GMM) with two first order conditions, (FOC on β_λ) and (FOC on ρ), which are derived from (17).

almost the same as the underlying value of $\rho - 1$. In summary, Figure 4 shows the superiority of the structural approach taken in this paper.

Next, we can also simulate all components of the model to better understand the mechanism of an insignificant supply factor in a linear regression. Plots of the estimated coefficients under different residual volatility σ are shown in Figure 8, Figure 9, and Figure 10. Similar to the plots of the t-stats, the GMM estimated coefficient has much lower volatility after σ goes down, while linear regression estimated coefficient remains as volatile. Thus both of the two linear regression approaches are fundamentally inconclusive about whether treasuries and money are perfect substitutes.

B.5. Illustrations of Linear Regressions and Interaction effects

To illustrate how different measures improve the results, and the strength of the interaction effects, In Table 6, I do regressions with different measures of bonds, as well as including the interaction effects. We find that a better measure indeed helps improve predictions a lot, and the interaction effect is strongly significant.

B.6. Log-Scale Regressions

When bonds and deposits are fully substitutable, we have $\rho = 1$, and

$$\frac{i_t - i_t^b}{(1-\delta)i_t} = \frac{\lambda_t}{1-\lambda_t}$$

Electronic copy available at: https://ssrn.com/abstract=3239293

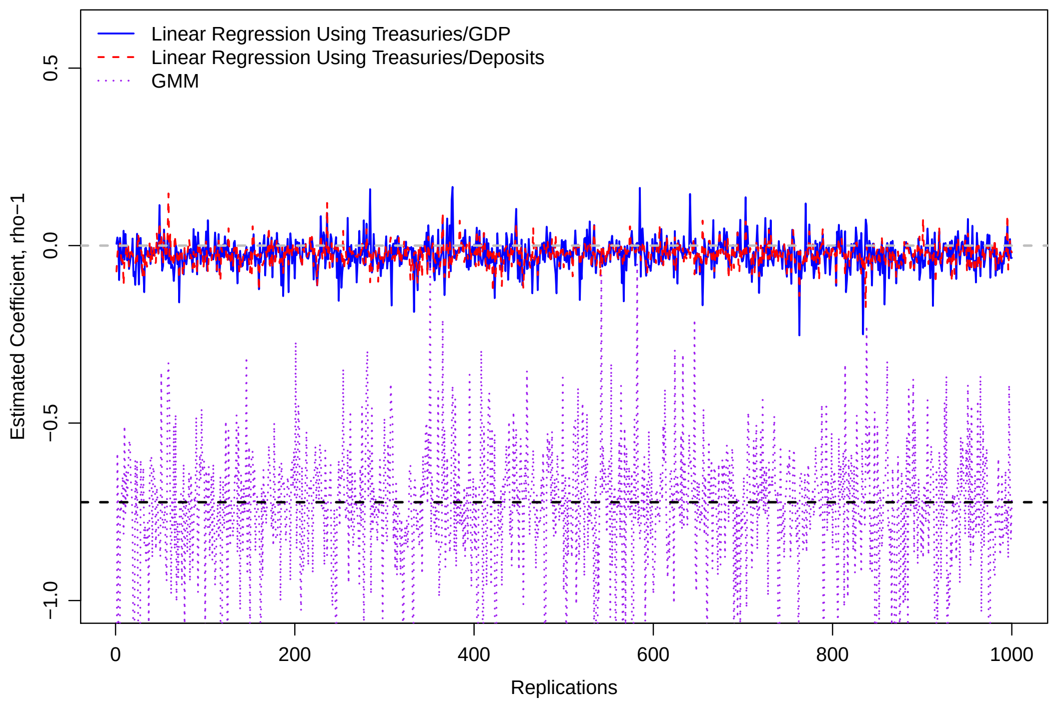


Fig. 8. Estimated Coefficients of Liquidity Supply in Pure Simulations with $\sigma = 0.01$

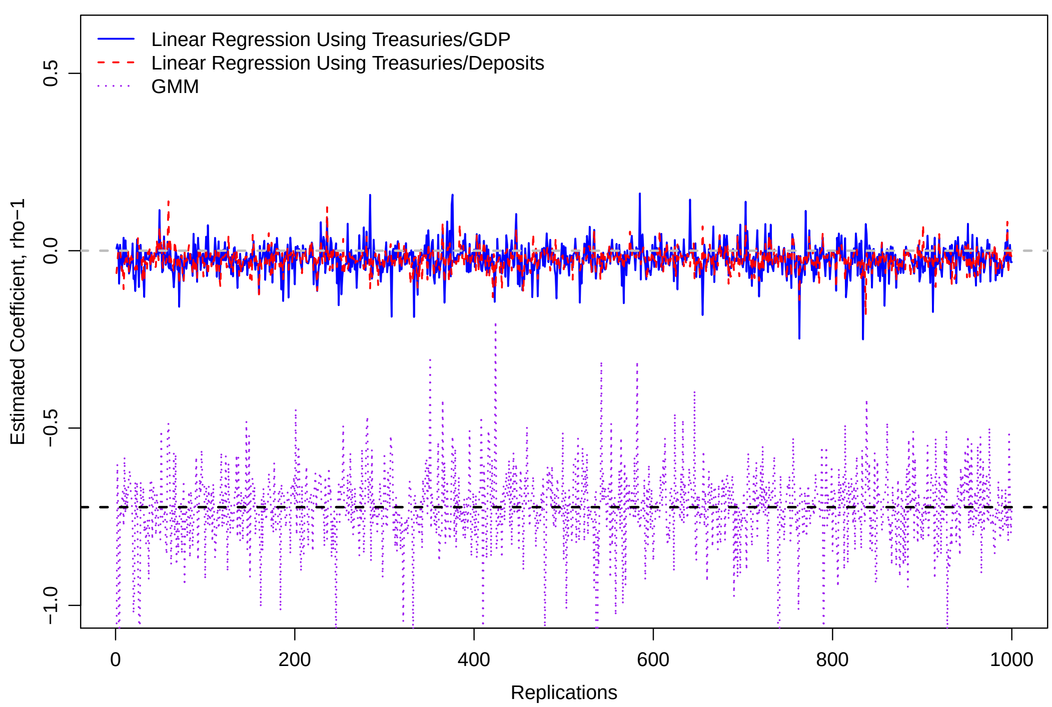


Fig. 9. Estimated Coefficients of Liquidity Supply in Pure Simulations with $\sigma = 0.006$

Electronic copy available at: https://ssrn.com/abstract=3239293

Table 6: Regressions with Different Measures and Interaction Effects

	Dependent variable:					
	100∗Liquidity Premium					
	(1)	(2)	(3)	(4)	(5)	(6)
FFR	11.15***	10.85***	10.36***	3.01	3.77	16.68***
	(1.43)	(1.43)	(1.07)	(3.69)	(2.52)	(2.89)
VIX	1.23***	1.19***	2.34***	1.03***	1.05***	2.08***
	(0.26)	(0.26)	(0.62)	(0.23)	(0.24)	(0.57)
log(Bonds/GDP)	0.21					
	(5.18)					
log(Bonds′/GDP)		0.06				
		(5.44)				
log(Bonds'/Deposits)			−34.58***			
			(11.58)			
FFR*log(Bonds/GDP)				−6.23*		
				(3.25)		
FFR*log(Bonds′/GDP)					−5.08**	
					(2.02)	
FFR*log(Bonds′/Deposits)						−7.16***
						(2.57)
Constant	−30.12***	−27.33***	−13.72	−20.28***	−18.83***	−41.79***
	(9.02)	(7.85)	(10.57)	(7.42)	(7.21)	(12.24)
Observations	1,032	1,092	696	1,032	1,092	696
R^2	0.57	0.57	0.62	0.59	0.59	0.65

Notes: Newey West standard errors with 12 lags are used to control time series correlations. *p<0.1; **p<0.05; ***p<0.01. Bonds are the total government bonds, while bonds′ exclude Federal Reserve holding and foreign government holding. Deposits are measured as the effective deposits constructed by aggregating both saving and checking deposits.

Electronic copy available at: https://ssrn.com/abstract=3239293

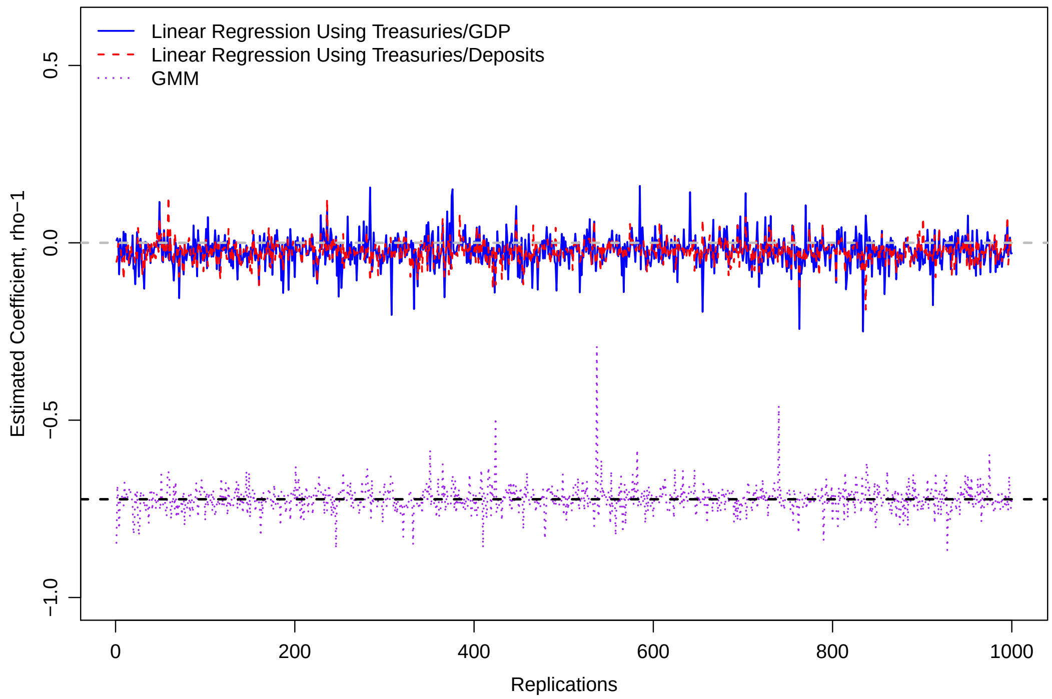


Fig. 10. Estimated Coefficients of Liquidity Supply in Pure Simulations with $\sigma = 0.002$

To test the relationship, we need

$$\log(i_t - i_t^b) \approx \log(1-\delta) + \log(i_t) + \log(\frac{\lambda_t}{1-\lambda_t})$$

When bonds and deposits are neither substitutes nor complements, we have $\rho = 0$, and

$$\log(i_t - i_t^b) = -\log(u_c(C_t)) + \log(1+i_t) + \log(\alpha) + \log(\lambda_t) - \log(b_t)$$

When $i_t \to 0$, the variations in i_t can be neglected, and thus the liquidity premium should be only related to the supply of bonds. Based on the above observations on the two extreme cases, Nagel (2016) designs linear regressions and find that the supply of treasury loses its explanatory power when we also include the interest rate.

If in a regression, we find that coefficients on $\log(i_t)$ is 1, while the coefficient on both the $\log(b_t)$ and $\log(GDP)$ are zero, then we are confident that there is a high degree of substitution among bonds and deposits. However, if we have a coefficient on i_t that is close to 1, then the results are more appealing to the non-substitutes case, especially when we have a significance on the treasury supply $\log(b_t)$, and GDP growth $\log(GDP)$.

In Table 7, we find strong supportive evidence in favor of a weak substitution between money (deposits) and bonds. With different measures of bonds/GDP ratio, we get mostly

Electronic copy available at: https://ssrn.com/abstract=3239293

very significant and consistent results. We note that using the original DebtToGDP measure, we get a reasonable magnitude but insignificant coefficient in column (2) of Table 7. Then I use an improved measure that excludes federal reserve holding of bonds, which then restores significance. Using the total public liquidity supply as a measure of liquidity also restores significance, as shown in column (1). Although the bond to GDP measure in column (2) is not significant, it becomes significant when we separate bond and GDP in column (4), and the signs are exactly to what we expect from the model. As another direct evidence, I aggregate the improved DebtToGDP measure into yearly level, and do a difference regression in column (5). Results are very significant.

Next, we notice that before the Federal Reserve directly control the fed funds rate (before 1980), the Fed Funds rate is not significant in the regression, while after 1980, it becomes very significant.

B.7. Money of Zero Maturity

Money of zero maturity is the most popular measure of money after 2005 for making macroeconomic inferences, especially on inflation. However, this measure contains money market mutual funds, which might be strong substitutes with treasuries, since their interest rates move very close to each other. Due to this strong substitution effect, I use the "nlminb" optimization algorithm that imposes a constraint of $\rho \in [0, 1]$. Results are shown in Table 8. We find that the substitution effect is quite strong, and in column (4), the estimated value is exactly on the boundary of 1, implying very strong substitution effects between money of zero maturity and treasuries.

Electronic copy available at: https://ssrn.com/abstract=3239293

Table 7: Liquidity Premium Regressions

	Dependent variable:			
	log(RepoT)		log(RepoT) Diff	
	(1)	(2)	(3)	(4)
log(Bond/GDP)	-0.79^{***}			
	(0.28)			
log(GDP)	0.80^{***}			
	(0.30)			
log(Bond/Deposits)		-0.63^{**}		
		(0.25)		
log(FFR)	0.43^{***}	0.46^{***}		
	(0.08)	(0.08)		
log(VIX)	0.45^{**}	0.49^{***}		
	(0.19)	(0.17)		
Liq_Log_Diff			-1.69^{***}	
			(0.65)	
GDP_Log_Diff			-0.34	
			(1.61)	
LiqGDP_Log_Diff				-1.62^{***}
				(0.52)
FFR_Log_Diff			0.54^{***}	0.46^{**}
			(0.18)	(0.21)
VIX_Log_Diff			1.09^{***}	1.29^{***}
			(0.39)	(0.34)
Constant	-3.88^{***}	-3.81^{***}	0.15^{***}	0.02
	(0.77)	(0.51)	(0.04)	(0.03)
Observations	962	996	86	89
R^2	0.46	0.44	0.28	0.26
Adjusted R^2	0.46	0.44	0.24	0.24
Residual Std. Error	0.80	0.81	0.69	0.69
F Statistic	205.29^{***}	258.86^{***}	7.74^{***}	10.07^{***}
Note:		*p<0.1; **p<0.05; ***p<0.01		

Electronic copy available at: https://ssrn.com/abstract=3239293

Table 8: Estimations using MZM

Parameters	Estimated Value (Volatility)			
	(1) 1959-2012	(2) 1959-1990	(3) 1970-2000	(4) 1980-2012
ρ	0.884	0.808	0.948*	1.000***
	(0.622)	(0.879)	(0.570)	(0.289)
β_λ	0.009***	0.009***	0.009***	0.008***
	(0.001)	(0.002)	(0.001)	(0.001)
Observations	468	192	312	396
p-value of J-tests	0.173	0.202	0.292	0.032

Note: *p<0.1; **p<0.05; ***p<0.01

Electronic copy available at: https://ssrn.com/abstract=3239293