Regression-Discontinuity-Based Control-Function Approach for Income Tax Effects on Male Work Hours

Young-sook Kim*
Korean Women's Development Institute
Seoul 122-707, South Korea
youngkim@kwdimail.re.kr

Myoung-jae Lee
Dept. Economics, Korea University
Seoul 136-701, South Korea
myoungjae@korea.ac.kr

- In finding the *effect of income tax on work hours*, to deal with the endogeneity of income tax, we take advantage of jumps in income tax—a regression discontinuity (RD) idea.

- *Whereas the usual RD estimators amount to local IVE, we propose a 'local' control function (CF) approach that combines the 'global' CF approach with RD: RD to find a legitimate local instrument, and CF to remove the income-tax endogeneity.*

- A general econometric methodology for the RD-based CF approach is developed, and compared with the global CF approach and the usual RD approach.

- The parameters estimated by the RD-based CF approach are more structural than those estimated by the global CF approach, because the former can allow "smooth endogeneity" that the latter cannot.

- Applying the method to find the income tax effect on male work hours (the income tax law provides income cutoffs needed for RD), we found no effect, or a very small effect of -3% if any, using Korean data.

1 Introduction

Effects of *income tax on work hours* are of interest, as income tax can be changed exogenously with tax rate or lump-sum tax. The main econometric difficulty is the endogeneity of the (previous period) income tax.

One way to find an instrument for income tax is using RD, as discontinuity in income tax occurs when the income passes a threshold in the tax law. The dummy variable for income being above the threshold makes a valid local instrument.

RD estimators can be viewed as local IVE, but given an instrument, there are several ways to deal with endogeneity. Among those, 'control function (CF)' approach adding a CF to removes the endogeneity is possibly advantageous as the error term SD may get smaller.

Structural form (SF) equations for *work hours* h and *income tax* τ are

$$h = \alpha_\tau \tau + u_h \quad \text{and} \quad \tau = \alpha_s 1[c \le s] + u_\tau \qquad \text{where } 1[A] = 1 \text{ if } A \text{ holds and 0 otherwise;}$$

$s =$ *is the (lagged) income,* c *is a cutoff* in the income tax law.

Using $\tau = E(\tau|s) + \varepsilon_\tau$ with $\varepsilon_\tau \equiv \tau - E(\tau|s)$ (τ is fixed given (s, ε_τ)),

$$E(h|s, \varepsilon_\tau) = \alpha_\tau \tau + E(u_h|s, \varepsilon_\tau) \simeq \alpha_\tau \tau + E(u_h|c, \varepsilon_\tau) \quad \forall s \simeq c$$

under the continuity of $E(u_h|s, \varepsilon_\tau)$ at $s = c$.

Assuming, e.g., $E(u_h|c, \varepsilon_\tau) = \alpha_{\varepsilon 1}\varepsilon_\tau + \alpha_{\varepsilon 2}\varepsilon_\tau^2$, ε_τ replaced by an estimator $\hat{\varepsilon}_\tau$ gives

$$E(h|s, \varepsilon_\tau) \simeq \alpha_\tau \tau + \alpha_{\varepsilon 1}\hat{\varepsilon}_\tau + \alpha_{\varepsilon 2}\hat{\varepsilon}_\tau^2 \qquad \forall s \simeq c.$$

Estimate this by the LSE of h on $(\tau, \hat{\varepsilon}_\tau, \hat{\varepsilon}_\tau^2)$ using some local observations around $s = c$.

The continuity of $E(u_h|s, \varepsilon_\tau)$ *and discontinuity of* τ *at* $s = c$ *is the RD idea, and using* $\hat{\varepsilon}_\tau$ *and* $\hat{\varepsilon}_\tau^2$ *to remove the* τ*-endogeneity is the CF idea.* The usual RD estimator is the IVE to $h = \alpha_\tau \tau + u_h$ with $1[c \le s]$ as an IV using local observations; the usual "global" CF approach is essentially the same as the above LSE except using all observations.

2 Global CF Approach

Consider a SF for $\ln h$ and a RF for $\ln \tau$:

$$\begin{aligned} \text{work hour SF} \quad &: \quad \ln h_i = \beta_\tau \ln \tau_i + x_{hi}' \beta_x + u_{hi} \\ \text{income tax RF} \quad &: \quad \ln \tau_i = E(\ln \tau | x_i) + v_{\tau i} = x' \eta_i + v_{\tau i} \quad \text{where} \quad v_\tau \equiv \ln \tau - E(\ln \tau | x) \end{aligned} \tag{A$_1$}$$

x_h and x are exogenous regressors (x strictly including x_h), and u_h and v_τ are errors.

A$_1$ includes the linearity assumption for $E(\ln \tau | x_i)$—not essential. As x strictly includes x_h, the components of x not in x_h are the instruments for $\ln \tau$. All endogenous regressors in the $\ln h$ SF should be substituted out.

Take $E(\cdot|x, v_\tau)$, neither $E(\cdot|x)$ nor $E(\cdot|x_\tau, v_\tau)$, on the $\ln h$ SF to obtain

$$\begin{aligned} E(\ln h | x, v_\tau) &= \beta_\tau \ln \tau + x_h' \beta_x + E(u_h | x, v_\tau) = \beta_\tau \ln \tau + x_h' \beta_x + \beta_{v1} v_\tau + \beta_{v2} v_\tau^2 \\ \text{assuming} \quad & E(u_h | x, v_\tau) = E(u_h | v_\tau) = \beta_{v1} v_\tau + \beta_{v2} v_\tau^2; \end{aligned} \tag{A$_2$}$$

if desired, v_τ^3, $x v_\tau$,... can appear in (A$_2$).

Do LSE of $\ln \tau$ on x in (A$_1$) to get $\hat{v}_\tau = \ln \tau - x' \hat{\eta}$, and then the LSE of $\ln h$ on $(\ln \tau, x_h, \hat{v}_\tau, \hat{v}_\tau^2)$. This is a global CF approach under (A$_1$) and (A$_2$). The CF $\hat{v}_\tau$ $\hat{v}_\tau^2$ and removes the endogeneity of $\ln \tau$, and thus polynomials of $\ln \tau$ and interaction terms between $\ln \tau$ and x_h can be used.

Although $\ln \tau$ is the lagged income tax, instead, the current income tax may appear in a predictable form, say $\ln \tau^e$; the current income tax withheld may serve as $\ln \tau^e$. Then

$$\begin{aligned} &\ln h = \beta_\tau \ln \tau^e + x_h' \beta_x + \tilde{u}_h \\ &\Longrightarrow \ln h = \beta_\tau \ln \tau + x_h' \beta_x + u_h \qquad \text{where} \quad u_h \equiv \tilde{u}_h + \beta_\tau (\ln \tau^e - \ln \tau). \end{aligned}$$

3 Brief Review on Regression Discontinuity (RD)

RD refers to a treatment d $(= \ln \tau)$ and a response y $(= \ln h)$ such that d depends on a running/score variable s with $E(d|s)$ *discontinuous* at $s = c$, and y is related to d through

$$E(y|s) = \psi_d E(d|s) + m(s); \tag{3.1}$$

ψ_d is the treatment effect, and $m(s)$ is an unknown function of s *continuous* at $s = c$.

Take $\lim_{s \downarrow c}$ and $\lim_{s \uparrow c}$ on (3.1) to obtain ($E(\cdot|c^+)$ denotes the upper/right limit)

$$E(y|c^+) = \psi_d E(d|c^+) + m(c^+) \quad \text{and} \quad E(y|c^-) = \psi_d E(d|c^-) + m(c^-) \tag{3.2}$$

$$\Longrightarrow \ \psi_d = \{E(y|c^+) - E(y|c^-)\}/\{E(d|c^+) - E(d|c^-)\} \quad \text{(solving for } \psi_d\text{)}. \tag{3.3}$$

Two types of RD's: Sharp RD (SRD) with d determined only by s so that $E(d|s) = d$ (e.g., $d = 1[c \leq s]$), and Fuzzy RD (FRD) with d determined by s and an error κ so that $E(d|s) \neq d$(e.g., $d = 1[c \leq s]1[0 \leq s + \kappa]$). FRD includes SRD with $1[c \leq s]$ as a limiting case when $E(d|c^+) - E(d|c^-) = 1$.

Two popular RD estimators using local observations around $s = c$. The first is LSE to

$$y = \psi_d d + m(s) + e \quad \text{with} \quad e \equiv y - E(y|s) - \psi_d\{d - E(y|s)\} \tag{3.4}$$

that is a rewritten version of (3.1); $m(s)$ is replaced by a polynomial or spline function.

Since $E(e|s) = 0$ holds by construction, LSE to (3.4) is fine for SRD. But for FRD, d may be endogenous due to $COR(\kappa, e) \neq 0$, in which case IVE can be applied with $1[c \leq s]$ as an "automatic" instrument for d.

The second RD estimator is a 'local linear kernel estimator' for (3.3), which is known to be the same as IVE to an artificial linear model (with $1[c \leq s]$ as an IV for d)

$$y_i = \gamma_d d_i + \gamma_0 + \gamma_l 1[s_i < c](s_i - c) + \gamma_r 1[c \leq s_i](s_i - c) + \nu_i;$$

the error ν is defined as y minus the other terms on the right-hand side.

4 RD-Based CF Approach

Consider a $\ln h$ SF and recall the $\ln \tau$ RF:

$$\ln h = \alpha_\tau \ln \tau + \varepsilon_h \qquad \text{and} \qquad \ln \tau = E(\ln \tau | s) + \varepsilon_\tau$$
$$\implies E(\ln h | s, \varepsilon_\tau) = \alpha_\tau \ln \tau + E(\varepsilon_h | s, \varepsilon_\tau) \qquad \text{as } \tau \text{ is fixed given } (s, \varepsilon_\tau).$$

Endogenous regressors with $E(\cdot|s, \varepsilon_h)$ continuous at $s = c$ can be subsumed in ε_h; this $\ln h$ SF does not require substituting out the endogenous regressors, differently from the global CF approach. All following discussion is done only for a local neighborhood of $s = c$.

Relative to the income tax break, a small change in $E(\varepsilon_h|s, \varepsilon_\tau)$ due to s can be ignored—a RD idea:

$$E(\varepsilon_h|s, \varepsilon_\tau) \text{ is a continuous function of } s \text{ for all } \varepsilon_\tau$$
$$\implies \quad E(\varepsilon_h|s, \varepsilon_\tau) \simeq E(\varepsilon_h|c, \varepsilon_\tau) \implies E(\ln h|s, \varepsilon_\tau) \simeq \alpha_\tau \ln \tau + E(\varepsilon_h|c, \varepsilon_\tau). \qquad (4.1)$$

Specifically, assume

$$E(\varepsilon_h|s, \varepsilon_\tau) = \alpha_{c0} + \alpha_{c1}\varepsilon_\tau + \alpha_{c2}\varepsilon_\tau^2 \implies E(\ln h|s, \varepsilon_\tau) = \alpha_\tau \ln \tau + \alpha_{c0} + \alpha_{c1}\varepsilon_\tau + \alpha_{c2}\varepsilon_\tau^2. \quad (A_3)$$

Compared with the global CF approach dropping x as in $E(u_h|x, v_\tau) = E(u_h|v_\tau)$, the continuity of $E(\varepsilon_h|s, \varepsilon_\tau)$ in s is much weaker.

Replacing s with c in (4.1) is a 'local-constant' approximation. To improve the local approximation, a linear function of s might be used in (A_3)—recall $m(s)$ in (3.1):

$$E(\varepsilon_h|s, \varepsilon_\tau) = \alpha_{c0} + \alpha_{c1}\varepsilon_\tau + \alpha_{c2}\varepsilon_\tau^2 + \alpha_{cs} \ln s. \qquad (A_3')$$

Nonparametric RD-based CF approach obtains a nonparametric residual $\tilde{\varepsilon}_\tau = \ln \tau - \tilde{E}(\ln \tau|s)$ first, and then does LSE for (A_3) of $\ln h$ on $(\ln \tau, 1, \hat{\varepsilon}_\tau, \hat{\varepsilon}_\tau^2)$. Obtain $\tilde{E}(\ln \tau|s)$ with one-sided kernels to prevent the local averaging from spanning over c.

In the nonparametric approach, accounting for the first-stage estimation error's effect on the second stage is difficult, although this problem vanishes under the null of no endogeneity. For this, adopt *semiparametric RD-based CF approach using* $\hat{\varepsilon}_\tau$ *that is the residual with* $E(\ln \tau|s)$ *specified as (log-) linear.*

The break location c and break magnitude of $E(\ln \tau|s)$ may be known from the tax law. But due to deductions and exemptions, the actual value of c and break magnitude may change. Plot $E(\ln \tau|s)$ versus s to find them as follows. Once the form of $E(\ln \tau|s)$ is found, a linear model may be used.

With K denoting a kernel and a a bandwidth, an estimator for a jump at s_o is the difference between a right-sided kernel estimator and a left-sided kernel estimator:

$$J_N(s_o) \equiv \frac{\Sigma_i K\{(s_i - s_o)/a\} 1[s_o \leq s_i] \ln \tau_i}{\Sigma_i K\{(s_i - s_o)/a\} 1[s_o \leq s_i]} - \frac{\Sigma_i K\{(s_i - s_o)/a\} 1[s_i < s_o] \ln \tau_i}{\Sigma_i K\{(s_i - s_o)/a\} 1[s_i < s_o]}. \tag{4.2}$$

$J_N(s_o) \to^p$'the jump magnitude at s_o'.

RD-based CF approach has a local "automatic" instrument, while the global CF approach needs a "global" instrument. Although variables such as the number of household members below/above certain age threshold may be used as income-tax-law-provided IV's, such IV's are invalid if they directly affect $\ln h$.

A concern is manipulating income to move into a lower tax bracket. Suppose u taking on $0, 1$ is an individual trait to comply with laws, $P(u = 0) = P(u = 1) = 0.5$, and

$$f(s|u = 0) = 1[c - 1 \leq s < c] \quad \text{and} \quad f(s|u = 1) = \phi(s - c) \quad \text{where } \phi \text{ is } N(0, 1) \text{ density}$$
$$\implies f(s) = 0.5\{1[c - 1 \leq s < c] + \phi(s - c)\} \implies f(c^+) - f(c^-) = -0.5.$$

The break of $f(s)$ at c occurs because those with $u = 0$ manipulated their s to perfection. Suppose $f(s|u = 0)$ being a continuous density tilted heavily to the left of c, which means that those with $u = 0$ could not perfectly manipulate their s although they could to a large extent. Then, $f(c^+) - f(c^-) = 0$. Hence, the concern can be gauged by estimating $f(s)$.

5 Empirical Analysis

5.1 Data

The Korean National Survey of Tax and Benefit by the Korea Institute of Public Fiance. Three panel waves (2009-2011) are available; $N = 2874$ married households with working males.

The weekly work hours is 48, and yearly income tax is 1390 (in 1000 Won $\simeq$1\$) much lower than 'before-tax income times tax rate' due to deductibles, exemptions and special tax treatments. Household monthly income includes wife's income, rental income, financial income, transfer income,...

Table 1: Descriptive Statistics ($N = 2874$; money amounts in 10,000 Won)

	Variables	Mean (SD)	Min, Max
Head	weekly work hours h	47.7 (10.9)	10, 126
	yearly income tax τ	138.7 (425.9)	0, 13472
	yearly income s	4394 (2559)	136, 48752
	x_h ($\ln h$ equation covariates):		
	age	41.5 (8.40)	22, 87
	college education	0.57 (0.50)	0, 1
	graduate education	0.09 (0.29)	0, 1
	regular worker	0.98 (0.14)	0, 1
Household	single-income family	0.67 (0.47)	0, 1
	# family members	3.78 (0.88)	2, 7
	own house	0.66 (0.47)	0, 1
	x other than x_τ:		
	# members with age $\geq$60	0.16 (0.47)	0, 3
	# children with age $\leq$20	0.91 (0.96)	0, 4
	$q_1 \equiv 1[400 \leq$house.month.income$< 600]$	0.29 (0.45)	0, 1
	$q_2 \equiv 1[600 \leq$house.month.income$]$	0.23 (0.42)	0, 1

5.2 Global CF Approach

Table 2: Global CF: First-Stage LSE for $\ln \tau$

Variables	Estimate	Bootstrap 95% CI
# family members	-0.27	-0.40, -0.16
own house	-0.071	-0.26, 0.11
age	0.14	0.052, 0.25
age^2/100	-0.15	-0.27, -0.049
college education	0.17	-0.015, 0.34
graduate education	0.29	-0.066, 0.63
regular worker	0.57	0.027, 1.05
single-income family	0.79	0.60, 0.95
# members with age $\geq$60	-0.17	-0.41, 0.078
# children with age $\leq$20	0.008	-0.11, 0.12
q_1	1.91	1.73, 2.08
q_2	2.97	2.73, 3.22

Table 3: Global CF: Second-Stage LSE for $\ln h$

$\ln \tau$	0.009	-0.19, 0.20
$q_1 \ln \tau$	0.013	0.00, 0.025
$q_2 \ln \tau$	0.012	0.00, 0.025
q_1	-0.11	-0.46, 0.27
q_2	-0.11	-0.66, 0.49
# family members	0.007	-0.053, 0.062
Own house	-0.022	-0.046, -0.001
Age	-0.013	-0.043, 0.016
Age^2/100	0.014	-0.018, 0.046
college education	-0.032	-0.068, 0.005
graduate education	-0.09	-0.156, -0.022
regular worker	0.053	-0.087, 0.21
single-income family	-0.012	-0.17, 0.15
$\hat{v}_\tau$	-0.013	-0.20, 0.18
$\hat{v}_\tau^2$	0.000	-0.003, 0.002

5.3 Usual RD-IVE

Using $J_N(s_o)$, Figure 1 shows multiple breaks: 2400, 3000, 3600, ... The tax drops right after break points are likely due to efforts to lower the tax, e.g., by separate/joint-filing tax returns or shifting around deductibles within the couple.

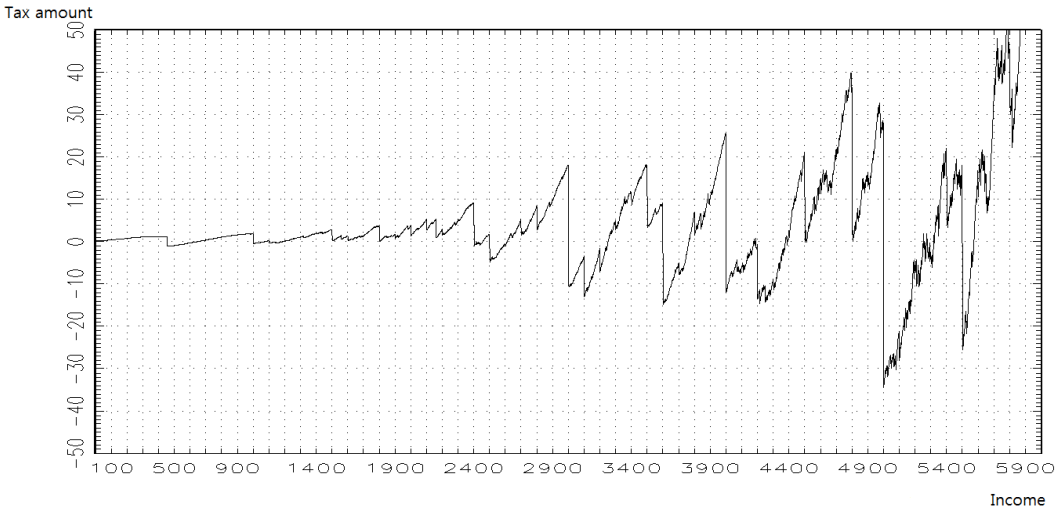

Figure 1: Mean Income Tax as a Function of Income (10,000 won)

Table 4: LSE-Based Break Test

# obs. (c)	318 (2400)		336 (3000)		168 (3600)	
Variables	Estimate	t-value	Estimate	t-value	Estimate	t-value
$1[c \leq s]$	15.80	2.27	29.79	2.05	7.82	0.54
1	3.78	0.61	9.69	0.75	29.87	2.17
$1[s < c](s - c)$	-0.04	-1.12	-0.027	-0.37	0.13	0.58
$1[c \leq s](s - c)$	-0.02	-0.99	-0.043	-0.92	-0.15	-1.29

Table 5: RD (IVE) for $\ln h$

# obs. (c)	318 (2400)		336 (3000)	
$\ln \tau$	-0.032	-0.59	0.057	0.53
q_1	-0.019	-0.45	-0.046	-0.76
q_2	-0.011	-0.12	-0.045	-0.84
$1[s < c](s - c)$	0.00	-0.17	0.00	-1.20
$1[c \leq s](s - c)$	0.00	-0.36	0.00	-0.04

5.4 RD-Based CF Approach

Table 6 shows the nonparametric RD-based CF approach results with the residual $\tilde{\varepsilon}_\tau$. Although $\tilde{\varepsilon}_\tau$ is insignificant, $\tilde{\varepsilon}_\tau^2$ is significant; the t-values were obtained under no endogeneity though. Table 8 is the second-stage for the semiparametric RD-based CF approach with the log-linear model residual $\hat{\varepsilon}_\tau$.

Table 6: RD-Based CF: $\ln h$ with Nonparametric $\tilde{\varepsilon}_\tau$

# obs. (c)	318 (2400)		336 (3000)	
Variables	Estimate	t-value	Estimate	t-value
$\ln \tau$	-0.033	-1.18	-0.002	-0.042
$q_1 \ln \tau$	0.036	1.89	0.021	1.15
$q_2 \ln \tau$	0.053	0.87	0.034	1.75
q_1	-0.099	-1.97	-0.07	-1.37
q_2	-0.11	-0.83	-0.12	-2.19
$\tilde{\varepsilon}_\tau$	0.03	1.14	-0.005	-0.13
$\tilde{\varepsilon}_\tau^2$	-0.009	-2.08	-0.009	-2.32

Table 8: RD-Based CF: Second Stage LSE for $\ln h$

# obs. (c)	318 (2400)		336 (3000)	
Variables	Estimate	Bootstrap 95% CI	Estimate	Bootstrap 95% CI
$\ln \tau$	-0.025	-0.17, 0.11	-0.05	-0.18, 0.085
$q_1 \ln \tau$	0.028	-0.031, 0.096	0.014	-0.017, 0.042
$q_2 \ln \tau$	0.04	-0.090, 0.28	0.034	-0.01, 0.066
q_1	-0.07	-0.29, 0.14	-0.043	-0.13, 0.06
q_2	-0.076	-0.82, 0.15	-0.10	-0.18, -0.011
$\hat{\varepsilon}_\tau$	0.027	-0.11, 0.16	0.042	-0.096, 0.18
$\hat{\varepsilon}_\tau^2$	-0.012	-0.023, -0.002	-0.009	-0.017, -0.002

6 Conclusions

In Fuzzy Regression Discontinuity (FRD) with a treatment d, d is a fuzzy version a "sharp" treatment $1[c \leq s]$ where s is the RD running/score variable and c is a cutoff.

A fuzzy RD estimator has an IVE interpretation: $1[c \leq s]$ is an instrument for d. As well known, IVE is only one of several approaches for an endogenous regressor with an IV.

Control function (CF) approach is another prominent approach, and its main advantage is "pulling out" the CF from the model error term to possibly reduce the error term SD.

CF approach brings up the question: *whether RD can be combined with CF. This papers showed positively how to do RD-based CF approach.*

Applying the method to find the income tax effect on male work hours as the income tax law provides income cutoffs for RD, we found no effect, or a very small effect of -3% if any, using Korean data.