

# **ECON 211B: Homework 5**

Due: Thursday, March 14, 2017

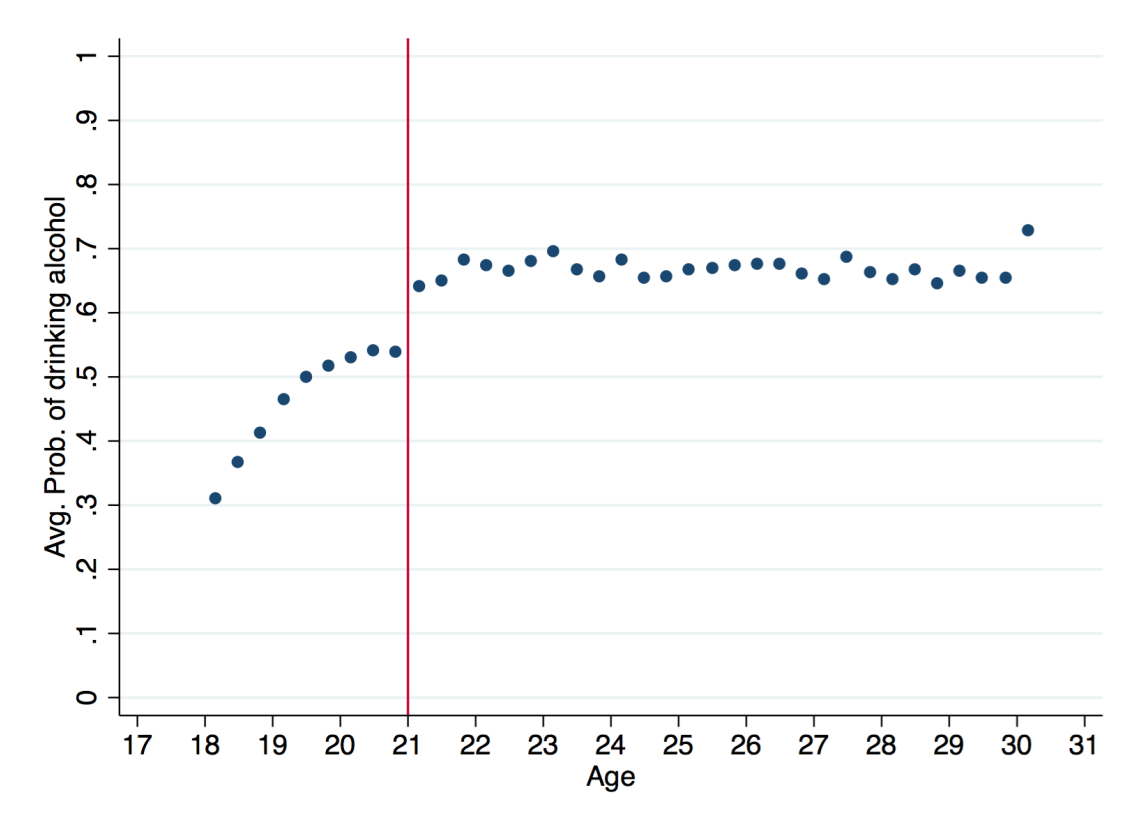
*Carlos E. Dobkin*

**David Sungho Park**

## Problem 1

Figure 1 displays the first stage relationship between the minimum legal drinking age (MLDA) and the average probability of drinking. It contains local averages with equal bin sizes. Age variable was calculated with *days\_21* ignoring any leap year (i.e. treating every year as 365 days). The average probability of drinking alcohol is local averages of *drinks\_alcohol*, which is a binary indicator variable. Several nonoverlapping bin sizes that do not span the cutoff, i.e. 21 years of age, have been tested, and the bin size with four-twelfth of a year was selected for it best portraits the first stage relationship with a clear jump at cutoff.

Figure 1: Relationship between MLDA and alcohol drinking

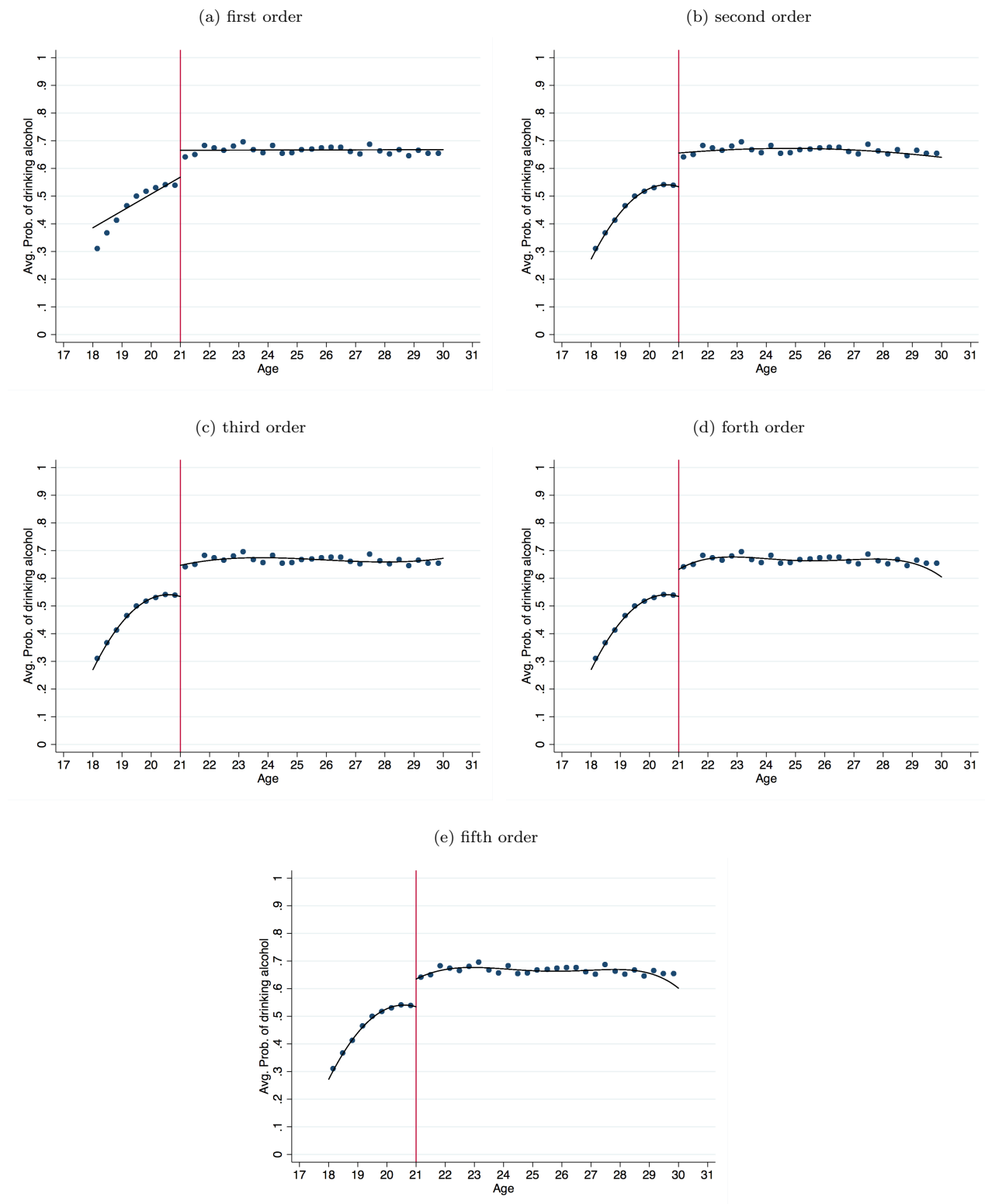


Note: The graph contains the local averages in nonoverlapping bins with the width of one-third of an year (i.e. approximately 121.7 days).

## Problem 2

Figure 2 and Table 1 show the results of regression continuity estimation with different order polynomials, from 1st to 5th order. Figure 2 includes the fitted values of the regression over the scatter plot from problem 1. We can visually check that the second order polynomial specification sufficiently fits the data. I use second order henceforth. The bandwidth are automatically selected by the *rdrobust* package. I also use a bandwidth of 1.39 henceforth.

Figure 2: Regression discontinuity fitted values



Note: The graph contains the fitted values from regression discontinuity estimation by each polynomial order along with local averages in nonoverlapping bins with the width of one-third of an year (i.e. approximately 121.7 days).

Table 1: Regression discontinuity estimation by different polynomial order

	coefficient	standard error	bandwidth
1st order	0.09	0.02	1.33
2nd order	0.08	0.03	1.39
3rd order	0.08	0.04	1.50
4th order	0.10	0.05	1.43
5th order	0.10	0.05	1.61

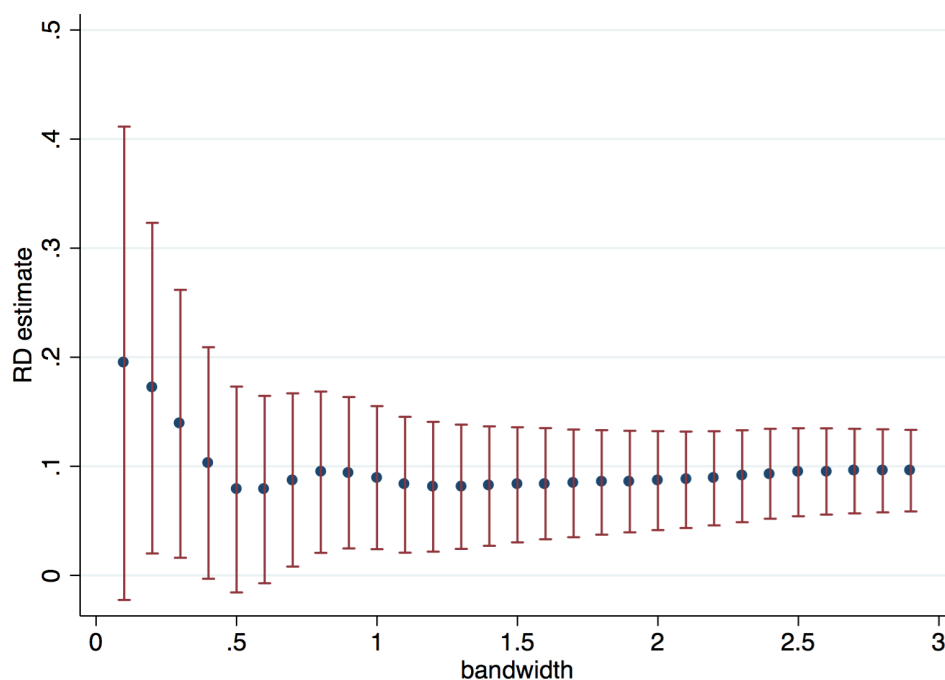
\* The row names are the order of polynomial with which the regression discontinuity estimation was run.

\*\* The last column is the bandwidth automatically selected by default method, which is the one common MSE-optimal bandwidth selector.

### Problem 3

Figure 3 plots how RD estimates and their 95% confidence intervals respond to bandwidth selection. Given that the automatically chosen bandwidth in Problem 2 is 1.39, I test for 30 bandwidths in the range from 0.1 to 2.9. It is safe to say that the RD estimates are stable where the bandwidth is equal to or greater than 1. Therefore, as long as the bandwidth is not too small (i.e. less than 0.7), the RD estimates are robust to bandwidth selection.

Figure 3: RD estimation by bandwidth selection



Note: The graph contains the RD estimates with 2nd order polynomial. The figure also displays the 95% confidence intervals.

## Problem 4

Table 2 contains the results of testing for any sharp changes in the characteristics of the sample. Except for *married*, there are not sufficient statistical evidence that the pretreatment variables have a discontinuity at the cutoff. All the specifications for RD estimation are according to the results in previous Problems.

Figures 4 and 5 are scatter plots with regression discontinuity fitted values. Again, except for *married*, we cannot see an evident jump at 21.

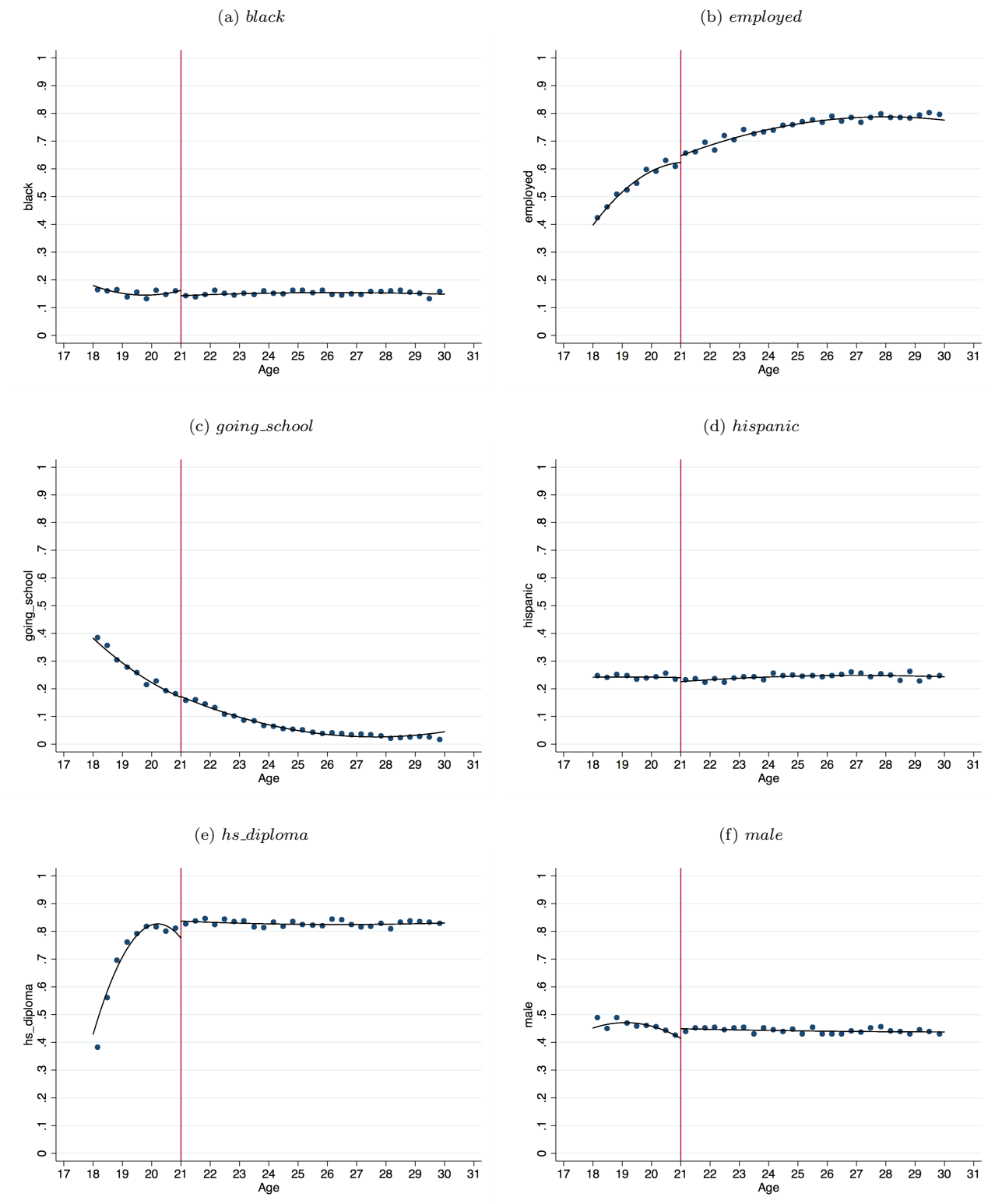
Table 2: Regression discontinuity estimation for

	RD estimate	standard error	p-value
<i>hs_diploma</i>	0.02	0.02	0.35
<i>hispanic</i>	0.01	0.03	0.56
<i>white</i>	-0.01	0.03	0.85
<i>black</i>	-0.02	0.02	0.44
<i>uninsured</i>	-0.00	0.03	0.91
<i>employed</i>	0.03	0.03	0.28
<i>married</i>	-0.04	0.02	0.03
<i>working_lw</i>	0.03	0.03	0.24
<i>going_school</i>	0.01	0.02	0.81
<i>male</i>	0.02	0.03	0.44

\* The row names are pretreatment variables.

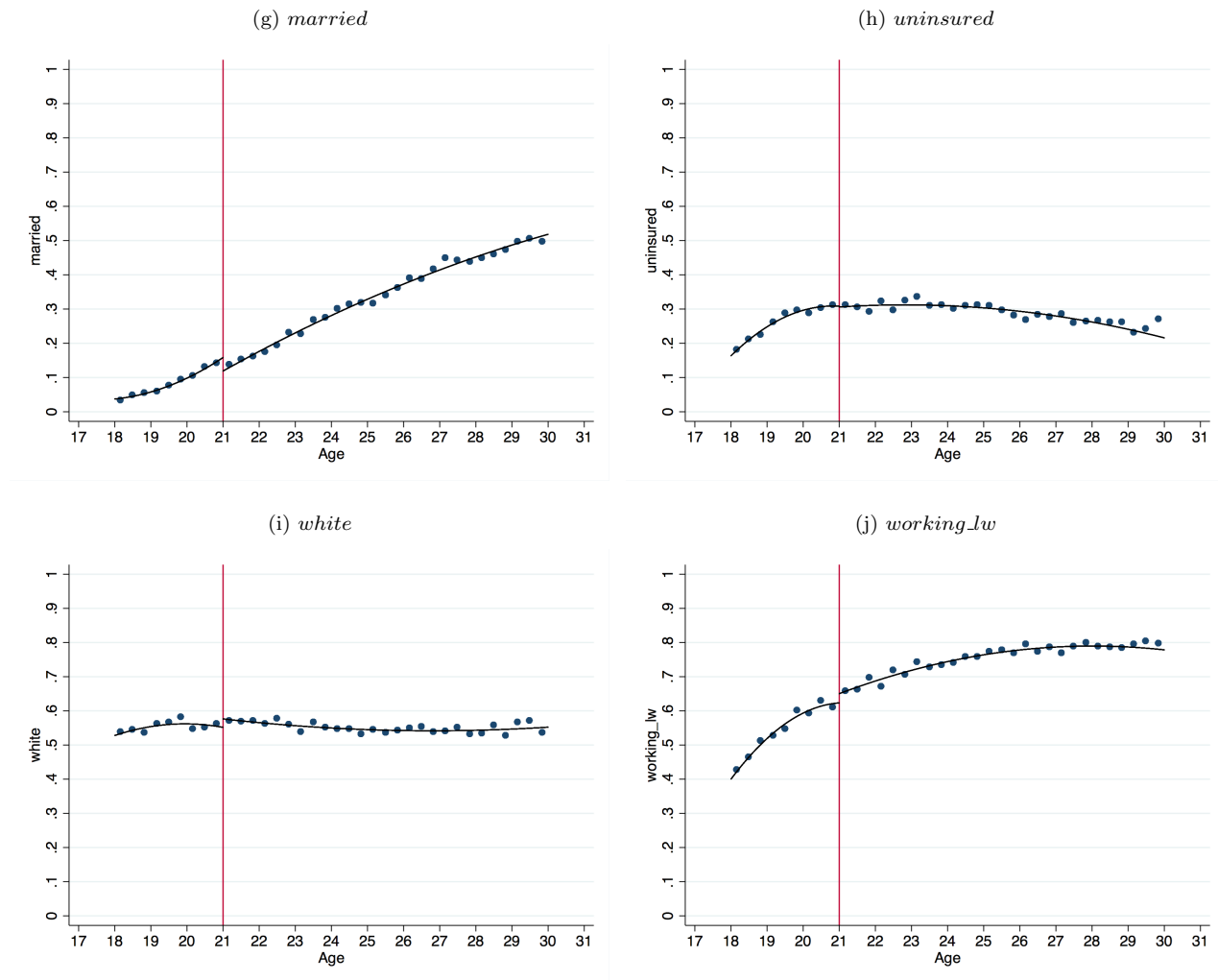
\*\* The last column is the p-value for the null hypothesis of no discontinuity.

Figure 4: RD examination for pretreatment variables



Note: The graph contains the fitted values from regression discontinuity estimation by each polynomial order along with local averages in nonoverlapping bins with the width of one-third of an year (i.e. approximately 121.7 days).

Figure 5: RD examination for pretreatment variables (continued)

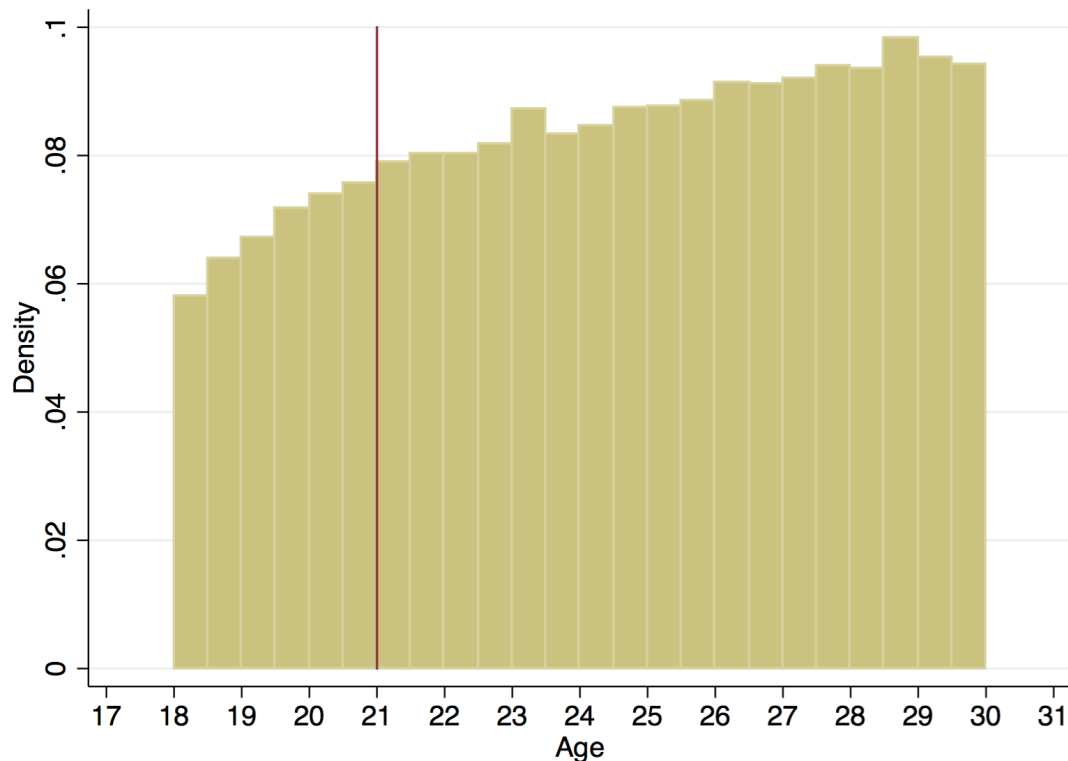


Note: The graph contains the fitted values from regression discontinuity estimation by each polynomial order along with local averages in nonoverlapping bins with the width of one-third of an year (i.e. approximately 121.7 days).

## Problem 5

Figure 6 shows the density of people responding to the NHIS survey. The number of respondents gradually increase by age, but there is no significant jump at age 21.

Figure 6: Density of survey respondents



Note: The graph contains the density of the number of people in each bin. The bin size is one-half of an year.

## Problem 6

According to the second row of Table 1, MLDA reduces the probability of drinking by 0.08, i.e. 8 percent. We can interpret this number as the fraction of population that changed their behavior (from not drinking to drinking) due to MLDA. This value is confirmed by the first stage estimation results conducted in Problem 11.

## Problem 7

Figure 7 plots the death rates due to all causes by age profiles.

## Problem 8

Figure 8 plots the death rates due to motor vehicle accident (MVA) causes.



Figure 7: Deaths due to all causes

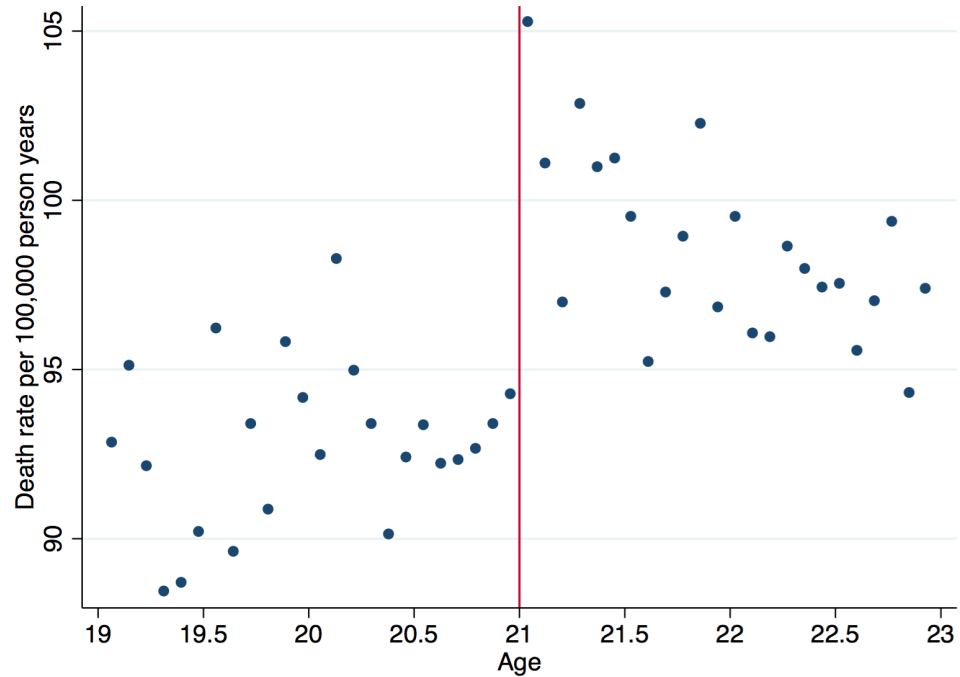
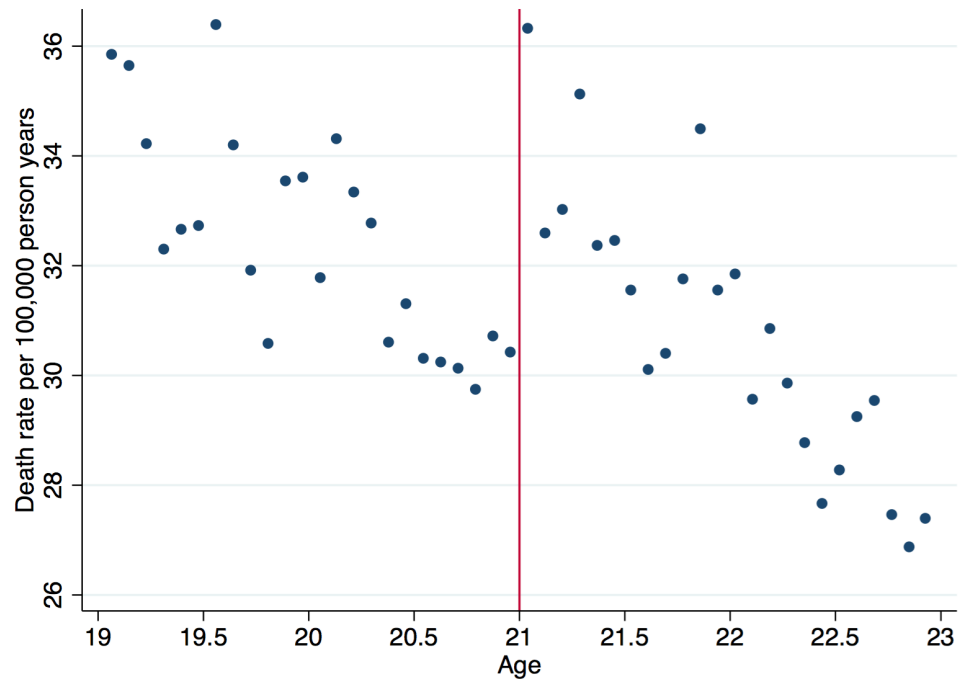


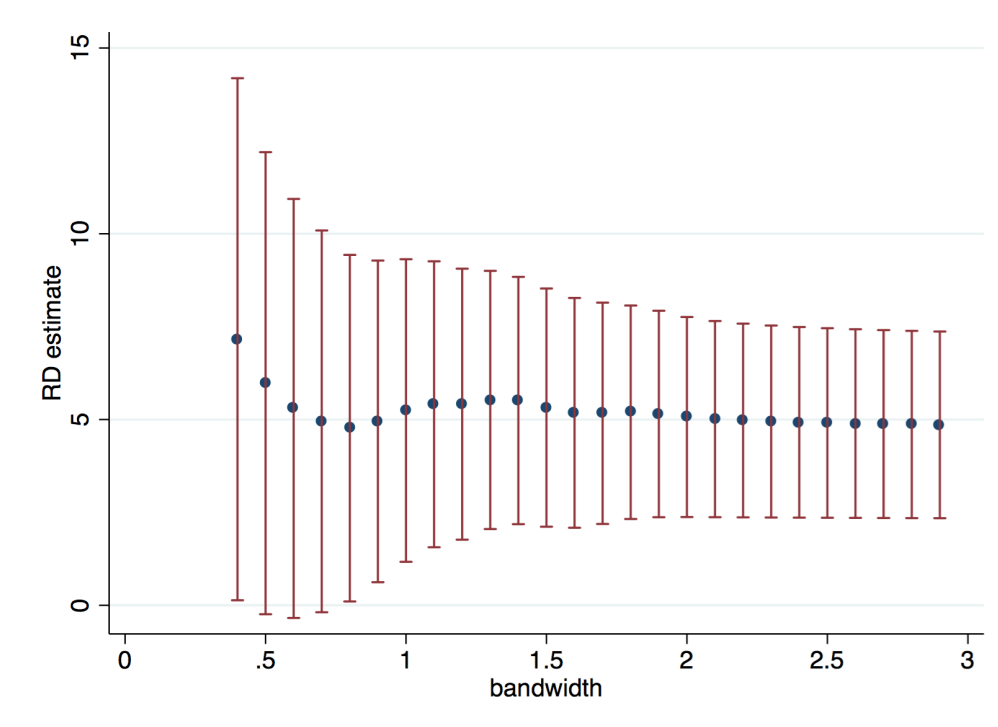
Figure 8: Deaths due to MVA causes



## Problem 9

Figure 9 plots how RD estimates of MLDA's effect on MVA-caused deaths and their 95% confidence intervals respond to bandwidth selection. I test for 26 bandwidths in the range from 0.4 to 3.0. It is safe to say that the RD estimates are stable where the bandwidth is equal to or greater than 1.2. Therefore, as long as the bandwidth is not too small (i.e. less than 1), the RD estimates are robust to bandwidth selection.

Figure 9: RD estimates by bandwidth selection



Note:

## Problem 10

Table 3 shows the impact of MLDA on death rates, both all causes and MVAs. According the RD estimation results, MLDA does reduce death rates by about 11 per 100,000 person years.

Table 3: Impact of MLDA on death rates

	coefficient	standard error	p-value	observations	R-squared
all causes	10.955	2.334	0.000	34	0.717
MVA	5.207	1.398	0.001	34	0.513

\* The bandwidth of 1.39 was implemented.

## Problem 11

Table 4 displays how we estimate the effect of MLDA on MVAs. The first stage estimation is run on the NHIS dataset. The reduced form estimation is performed on the *mortality* dataset. By taking ratios, the result tells us that MLDA reduces MVA-caused deaths by approximately 6.37 million.

Table 4: 2SLS estimation of the impact of MLDA on MVA

	First stage	Reduced form	2SLS
Coefficient	0.082	5.207	63.743
standard error	0.026	1.398	26.392
p-value	0.002	0.001	
observations	13,164	34	
R-squared	0.018	0.513	

\* Bandwidth of 1.39 was imposed.

\*\* The standard error for the 2SLS estimator was estimated by the delta method.

## Problem 12

For the MLDA to be a valid instrument, we need assumptions that

1. MLDA has no spillover effect  
 $\Rightarrow$  At a given time, MLDA assignment can be understood as age distribution. Unless there is some systematic correlation in age, e.g. baby boom period, we have no problem with this assumption.
2. Drinking alcohol has no spillover effect  
 $\Rightarrow$  This is not realistic. Drinking alcohol is one of the typical object of peer pressure.
3. MLDA is randomly assigned  
 $\Rightarrow$  At a given time, MLDA, or age, is randomly assigned.
4. MLDA has no direct effect on MVA  
 $\Rightarrow$  Drinking alcohol is not only related to driving under influence. Also, being assigned with MLDA also means that one is 21 or older. the MVAs can be dependent on the age.
5. MLDA has nonzero effect on drinking alcohol  
 $\Rightarrow$  This makes sense especially because around age 21 is when people are usually interested in drinking. There should be many who start drinking alcohol right after becoming 21.
6. No defier  
 $\Rightarrow$  It is hard to believe that someone would be a defier in our context.

## Stata Codes

```
* Change working directory
cd "/Users/DSP/Dropbox/UCSC (2016- )/1stYear_2Q/211B/Homeworks/HW5"

* NHIS data: CSV to DTA (+ misc. cleansing)
import delimited "NHIS Data.csv", clear
gen age = 21 + days_21/365
gen MLDA = 0
replace MLDA = 1 if age >= 21
save "211b_hw5_NHIS_data.dta", replace

* In addition, drop extreme observations
use "211b_hw5_NHIS_data", clear
drop if age>30 | age<18
save "211b_hw5_NHIS_data_trimmed.dta", replace

* CSV to DTA: mortality data (+ misc. cleansing)
import delimited "mortality data.csv", clear
gen MLDA = 0
replace MLDA = 1 if age >= 21
save "211b_hw5_mortality_data.dta", replace

***** Problem 1 *****

* Bin size: (consider only 1,2,3,4,6,8,12-twelfth of an year)
local list = "1 2 3 4 6 8 12"
foreach i of local list {
    use "211b_hw5_NHIS_data_trimmed", clear
    gen age_cell_`i' = autocode(age,14*12/`i',17,31)
    replace age_cell_`i' = age_cell_`i' - 1/24*`i'
    collapse (mean) drinks_alcohol, by(age_cell_`i')
    twoway (scatter drinks_alcohol age_cell_`i',msize(small)), xline(21)
    ytitle(Avg. Prob. of drinking alcohol) xtitle(Age) xscale(range(17 31)) xlabel(17(1)31)
    yscale(range(0 1)) ylabel(0(0.1)1) graphregion(fcolor(white) lcolor(white))
    graph export age_cell_`i'.png, replace
}

***** Problem 2 *****

* Estimation with different order polynomials + superimposing fitted values to scatter plot
```

```

set more off
use "211b_hw5_NHIS_data_trimmed", clear
mat T = J(5,3,.)
forval i = 1/5 {
    quietly rdrobust drinks_alcohol age, c(21) kernel(tri) deriv(0) p('i') bwselect(mserd)
    mat T['i',1] = e(tau_cl)
    mat T['i',2] = e(se_tau_cl)
    mat T['i',3] = e(h_l)
    rdplot drinks_alcohol age, c(21) kernel(tri) deriv(0) p('i') nbins(9 27)
    graph_options(title("") legend(off) xline(21) ytitle(Avg. Prob. of drinking alcohol)
    xtitle(Age) xscale(range(17 31)) xlabel(17(1)31) yscale(range(0 1)) ylabel(0(0.1)1)
    graphregion(fcolor(white) lcolor(white)))
    graph export 211b_hw5_p2_rdplot_'i'_t.png, replace
}
mat rownames T = "1st order" "2nd order" "3rd order" "4th order" "5th order"
mat colnames T = "coefficient" "standard error" "bandwidth"
frmttable using "211b_hw5_p2_t.tex",statmat(T) tex fr nocenter sdec(2,2,2,3) replace

```

\*\*\*\*\* Problem 3 \*\*\*\*\*

```

* Robustness check of bandwidth selection (2nd polynomial)
set more off
use "211b_hw5_NHIS_data_trimmed", clear
gen bandwidth=.
gen rd_est = .
gen rd_ci_l = .
gen rd_ci_h = .
local i = 1
forval h = 0.1(0.1)3 {
    quietly rdrobust drinks_alcohol age, c(21) kernel(tri) deriv(0) p(2) h('h')
    replace bandwidth = 'h' in 'i'
    replace rd_est = e(tau_cl) in 'i'
    replace rd_ci_l = e(tau_cl) - 1.96 * e(se_tau_cl) in 'i'
    replace rd_ci_h = e(tau_cl) + 1.96 * e(se_tau_cl) in 'i'
    local i = 'i' + 1
}
tway (scatter rd_est bandwidth, msize(small)) (rcap rd_ci_h rd_ci_l bandwidth),
ytitle(RD estimate) ylabel(none) yscale(range(0 0.5)) xtitle(bandwidth) xscale(range(0 3))
xlabel(0(0.5)3) ylabel(0(0.1)0.5) legend(off) graphregion(fcolor(white) lcolor(white))
graph export 211b_hw5_p3_bandwidth.png, replace

```

## \*\*\*\*\* Problem 4 \*\*\*\*\*

```

* Visual inspection via plots
set more off
use "211b_hw5_NHIS_data_trimmed", replace
local x_list = "hs_diploma hispanic white black uninsured employed married working_lw
going_school male"
local list = "4"
foreach x of local x_list{
    rdplot `x' age, c(21) kernel(tri) deriv(0) p(2) nbins(9 27) graph_options(title(""))
    legend(off) xline(21) ytitle(`x') xtitle(Age) xscale(range(17 31)) xlabel(17(1)31)
    yscale(range(0 1)) ylabel(0(0.1)1) graphregion(fcolor(white) lcolor(white))
    graph export 211b_hw5_p4_`x'.png, replace
}

* Making table
set more off
use "211b_hw5_NHIS_data_trimmed", replace
mat T = J(10,3,.)
local i = 1
local x_list = "hs_diploma hispanic white black uninsured employed married working_lw
going_school male"
foreach x of local x_list {
    quietly rdrobust `x' age, c(21) kernel(tri) deriv(0) p(2) bwselect(mserd)
    mat T[`i',1] = e(tau_cl)
    mat T[`i',2] = e(se_tau_cl)
    mat T[`i',3] = ttail(e(N), abs(e(tau_cl)/e(se_tau_cl)))*2
    local i = `i' + 1
}
mat rownames T = hs_diploma hispanic white black uninsured employed married working_lw
going_school male
mat colnames T = "RD estimate" "standard error" "p-value"
frmttable using "211b_hw5_p4_cov.tex",statmat(T) tex fr nocenter sdec(2,2,2,3) replace

```

## \*\*\*\*\* Problem 5 \*\*\*\*\*

```

* Density histogram
use "211b_hw5_NHIS_data_trimmed", replace
twoway (histogram age, width(0.5))(scatteri 0 21 .1 21, c(1) m(i) lcolor(maroon)),
ytitle(Density) xtitle(Age) xscale(range(17 31)) xlabel(17(1)31) legend(off)
graphregion(fcolor(white) lcolor(white))
graph export 211b_hw5_p5_histogram.png, replace

```

## \*\*\*\*\* Problem 7 \*\*\*\*\*

```
* Scatter plot of deaths due to all causes
use "211b_hw5_mortality_data", replace
twoway (scatter all age,msize(small)), xline(21) ytitle("Death rate per 100,000 person years")
xtitle(Age) xscale(range(19 23)) xlabel(19(0.5)23) graphregion(fcolor(white) lcolor(white))
graph export death_all.png, replace

* RD estimation
set more off
use "211b_hw5_mortality_data", replace
mat T = J(2,4,.)
forval i = 1/2 {
    quietly rdrobust all age, c(21) kernel(tri) deriv(0) p('i') bwselect(mserd)
    mat T['i',1] = e(tau_cl)
    mat T['i',2] = e(se_tau_cl)
    mat T['i',3] = ttail(e(N), abs(e(tau_cl)/e(se_tau_cl)))*2
    mat T['i',4] = e(h_1)
    rdplot all age, c(21) kernel(tri) deriv(0) p('i') nbins(24 24) graph_options(title(""))
    legend(off) xline(21) ytitle(Avg. Prob. of drinking alcohol) xtitle(Age)
    xscale(range(19 23)) xlabel(19(0.5)23) graphregion(fcolor(white) lcolor(white))
    graph export 211b_hw5_p7_rdplot_all_`i'.png, replace
}
mat rownames T = "1st order" "2nd order"
mat colnames T = "RD estimate" "standard error" "p-value" "bandwidth"
frmttable using "211b_hw5_p7_all.tex",statmat(T) tex fr nocenter sdec(2,2,2,3) replace
```

## \*\*\*\*\* Problem 8 \*\*\*\*\*

```
* Scatter plot of deaths due to MVA causes
use "211b_hw5_mortality_data", clear
twoway (scatter mva age,msize(small)), xline(21) ytitle("Death rate per 100,000 person years")
xtitle(Age) xscale(range(19 23)) xlabel(19(0.5)23) graphregion(fcolor(white) lcolor(white))
graph export death_mva.png, replace

* RD estimation
set more off
use "211b_hw5_mortality_data", clear
mat T = J(2,4,.)
forval i = 1/2 {
```

```

quietly rdrobust mva age, c(21) kernel(tri) deriv(0) p('i') bwselect(mserd)
mat T['i',1] = e(tau_cl)
mat T['i',2] = e(se_tau_cl)
mat T['i',3] = ttail(e(N), abs(e(tau_cl)/e(se_tau_cl)))*2
mat T['i',4] = e(h_1)
rdplot mva age, c(21) kernel(tri) deriv(0) p('i') nbins(24 24) graph_options(title("")
legend(off) xline(21) ytitle(Avg. Prob. of drinking alcohol) xtitle(Age)
xscale(range(19 23)) xlabel(19(0.5)23) graphregion(fcolor(white) lcolor(white))
graph export 211b_hw5_p8_rdplot_mva_`i`.png, replace
}

mat rownames T = "1st order" "2nd order"
mat colnames T = "RD estimate" "standard error" "p-value" "bandwidth"
frmttable using "211b_hw5_p7_all.tex",statmat(T) tex fr nocenter sdec(2,2,2,3) replace

```

\*\*\*\*\* Problem 9 \*\*\*\*\*

```

* Robustness check of bandwidth selection for ALL (2nd polynomial)
set more off
use "211b_hw5_mortality_data", replace
gen bandwidth=.
gen rd_est = .
gen rd_ci_l = .
gen rd_ci_h = .
local i = 1
forval h = 0.4(0.1)3 {
    quietly rdrobust all age, c(21) kernel(tri) deriv(0) p(2) h('h')
    replace bandwidth = `h' in `i'
    replace rd_est = e(tau_cl) in `i'
    replace rd_ci_l = e(tau_cl) - 1.96 * e(se_tau_cl) in `i'
    replace rd_ci_h = e(tau_cl) + 1.96 * e(se_tau_cl) in `i'
    local i = `i' + 1
}

twoway (scatter rd_est bandwidth, msize(small)) (rcap rd_ci_h rd_ci_l bandwidth),
ytitle(RD estimate) xtitle(bandwidth) xscale(range(0 2)) xlabel(0(0.5)3) legend(off)
graphregion(fcolor(white) lcolor(white))
graph export 211b_hw5_p9_bandwidth_all.png, replace

* Robustness check of bandwidth selection for MVA (2nd polynomial)
set more off
use "211b_hw5_mortality_data", replace
gen bandwidth=.
gen rd_est = .

```



```

gen rd_ci_l = .
gen rd_ci_h = .
local i = 1
forval h = 0.4(0.1)3 {
    quietly rdrobust mva age, c(21) kernel(tri) deriv(0) p(2) h('h')
    replace bandwidth = 'h' in 'i'
    replace rd_est = e(tau_cl) in 'i'
    replace rd_ci_l = e(tau_cl) - 1.96 * e(se_tau_cl) in 'i'
    replace rd_ci_h = e(tau_cl) + 1.96 * e(se_tau_cl) in 'i'
    local i = 'i' + 1
}
tway (scatter rd_est bandwidth, msize(small)) (rcap rd_ci_h rd_ci_l bandwidth),
yttitle(RD estimate) xttitle(bandwidth) xscale(range(0 2)) xlabel(0(0.5)3) legend(off)
graphregion(fcolor(white) lcolor(white))
graph export 211b_hw5_p9_bandwidth_mva.png, replace

```

\*\*\*\*\* Problem 10 \*\*\*\*\*

\* Settings for output matrix

```

mat T = J(2,5,.)
mat colnames T = "coefficient" "standard error" "p-value" "observations" "R-squared"
mat rownames T = "all causes" "MVA"

```

\* MLDA on Deaths (All causes)

```

use "211b_hw5_mortality_data", replace
gen age_c = age - 21
reg all MLDA age_c c.age_c#c.age_c c.age_c#i.MLDA c.age_c#c.age_c#i.MLDA if abs(age_c) <= 1.39
mat T[1,1] = _b[MLDA]
mat T[1,2] = _se[MLDA]
mat T[1,3] = ttail(e(N), abs(_b[MLDA]/_se[MLDA]))*2
mat T[1,4] = e(N)
mat T[1,5] = e(r2)

```

\* MLDA on Deaths (MVA)

```

use "211b_hw5_mortality_data", replace
gen age_c = age - 21
reg mva MLDA age_c c.age_c#c.age_c c.age_c#i.MLDA c.age_c#c.age_c#i.MLDA if abs(age_c) <= 1.39
mat T[2,1] = _b[MLDA]
mat T[2,2] = _se[MLDA]
mat T[2,3] = ttail(e(N), abs(_b[MLDA]/_se[MLDA]))*2
mat T[2,4] = e(N)
mat T[2,5] = e(r2)

```

```
* Export table to LaTeX
frmttable using "211b_hw5_p10.tex",statmat(T) tex fr nocenter sdec(3,3,3,0,3) replace

***** Problem 11 *****

* Settings for output matrix
mat T = J(5,3,.)
mat rownames T = "Coefficient" "standard error" "p-value" "observations" "R-squared"
mat colnames T = "First stage" "Reduced form" "2SLS"

***** using 2SLS *****

* First stage
use "211b_hw5_NHIS_data_trimmed", replace
gen age_c = age - 21
reg drinks_alcohol MLDA age_c c.age_c#c.age_c c.age_c#i.MLDA c.age_c#c.age_c#i.MLDA
if abs(age_c) <= 1.39
mat T[1,1] = _b[MLDA]
mat T[2,1] = _se[MLDA]
mat T[3,1] = ttail(e(N), abs(_b[MLDA]/_se[MLDA]))*2
mat T[4,1] = e(N)
mat T[5,1] = e(r2)

* Reduced form
use "211b_hw5_mortality_data", replace
gen age_c = age - 21
reg mva MLDA age_c c.age_c#c.age_c c.age_c#i.MLDA c.age_c#c.age_c#i.MLDA if abs(age_c) <= 1.39
mat T[1,2] = _b[MLDA]
mat T[2,2] = _se[MLDA]
mat T[3,2] = ttail(e(N), abs(_b[MLDA]/_se[MLDA]))*2
mat T[4,2] = e(N)
mat T[5,2] = e(r2)

* 2SLS = RF/FS
mat T[1,3] = T[1,2] / T[1,1]

* SE for 2SLS estimator (Delta method)
mat T[2,3] = sqrt(1/(T[1,1])^2 *(T[2,2])^2 + (T[1,2])^2/(T[1,1])^4 *(T[2,1])^2)

* Export output matrix to LaTeX
frmttable using "211b_hw5_p11_2sls_2.tex",statmat(T) tex fr nocenter sdec(3\3\3\0\3) replace
```

