GV300 - Quantitative Political Analysis

University of Essex - Department of Government

Lorenzo Crippa

Week 22 – 24 February, 2020

Question 1

Question 1 – (a)

1. (a) (5 marks) What does exclusion restriction mean in the context of instrumental variable regression?

Question 1 – (a)

1. (a) (5 marks) What does exclusion restriction mean in the context of instrumental variable regression?

It means that the chosen instrument has an effect on the outcome variable **only** through the instrumented variable (the treatment variable)

Question 1 – (b)

1. (b) Generate variables `X`, `T`, `Y`, `Z` and `zNot` according to the instructions.

Question 1 – (b)

1. (b) Generate variables `X`, `T`, `Y`, `Z` and `zNot` according to the instructions. Explain and demonstrate graphically why your variable `Z` is an ideal instrumental variable.

Question 1 – (b)

1. (b) Generate variables `X`, `T`, `Y`, `Z` and `zNot` according to the instructions. Explain and demonstrate graphically why your variable `Z` is an ideal instrumental variable. Why would it be harder to demonstrate that `Z` is an ideal instrumental variable when we worked with observational data?

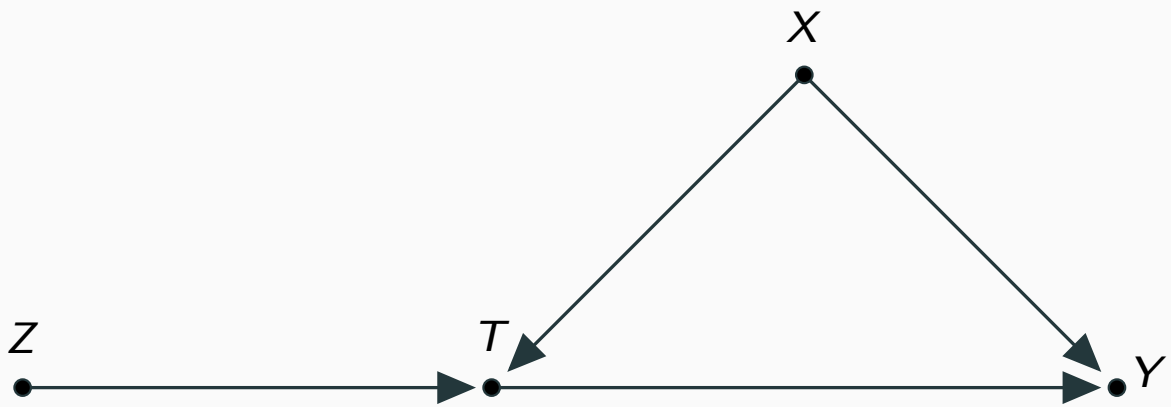
Question 1 – (b)

1. (b) Generate variables `X`, `T`, `Y`, `Z` and `zNot` according to the instructions. Explain and demonstrate graphically why your variable `Z` is an ideal instrumental variable. Why would it be harder to demonstrate that `Z` is an ideal instrumental variable when we worked with observational data? Create a variable `zNot`, which is not an ideal instrument. How would such a variable need to look like?

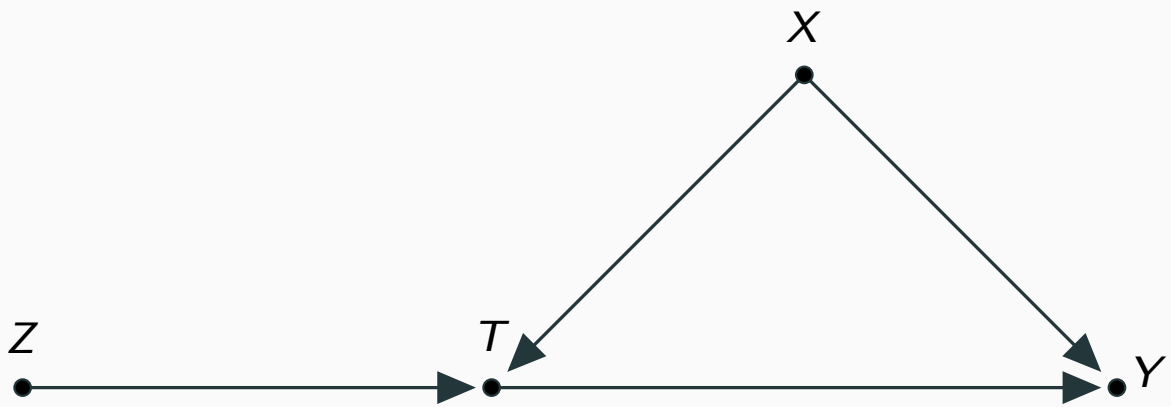
Question 1 – (b)

1. (b) Generate variables `X`, `T`, `Y`, `Z` and `zNot` according to the instructions. Explain and demonstrate graphically why your variable `Z` is an ideal instrumental variable. Why would it be harder to demonstrate that `Z` is an ideal instrumental variable when we worked with observational data? Create a variable `zNot`, which is not an ideal instrument. How would such a variable need to look like? Explain and demonstrate graphically how `Z` and `zNot` differ making `zNot` a not ideal instrument.

Question 1 – (b): DGP

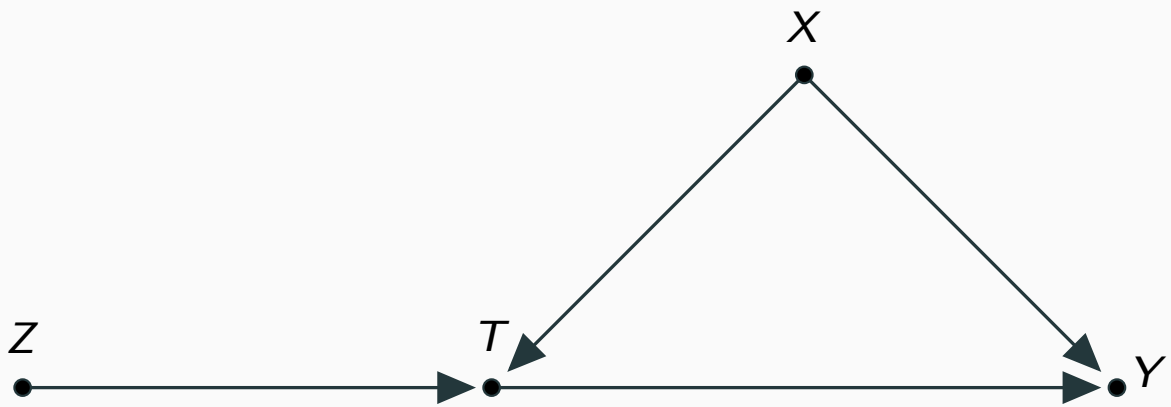


Question 1 – (b): DGP



Z is a good instrument because it determines T and meets the exclusion restriction.

Question 1 – (b): DGP



Z is a good instrument because it determines T and meets the exclusion restriction.

With observational data it would be difficult to find a good instrument as Z because it is very difficult that the exclusion restriction is met.

Question 1 – (b): bad instruments

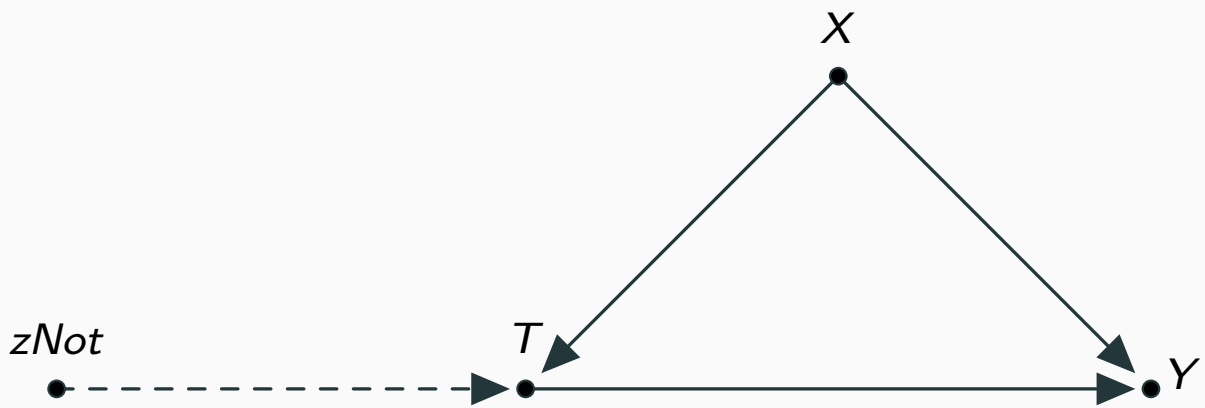
There are two ways `zNot` can be a bad instrument.

Question 1 – (b): bad instruments

There are two ways `zNot` can be a bad instrument. The first is that it has no (or weak) effect on `T`:

Question 1 – (b): bad instruments

There are two ways `zNot` can be a bad instrument. The first is that it has no (or weak) effect on `T`:

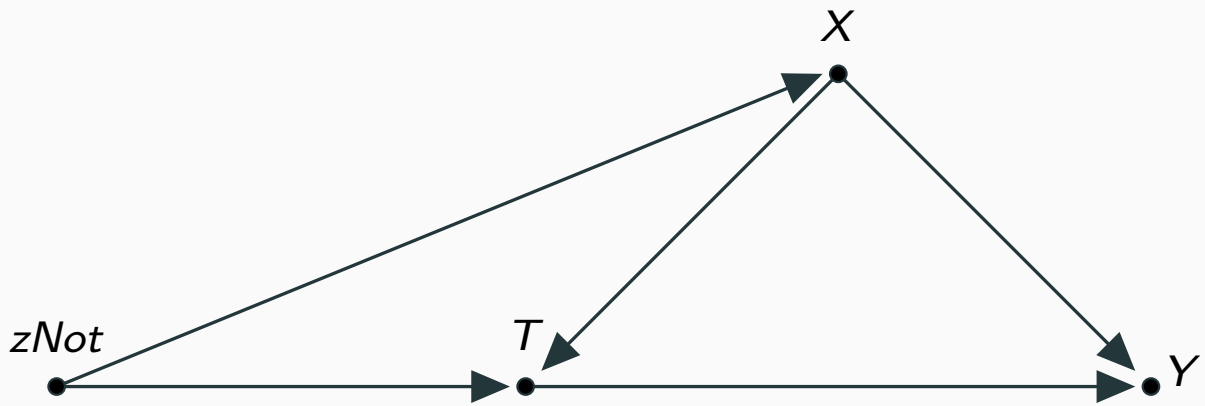


Question 1 – (b): bad instruments

The second is that the exclusion restriction is not met:

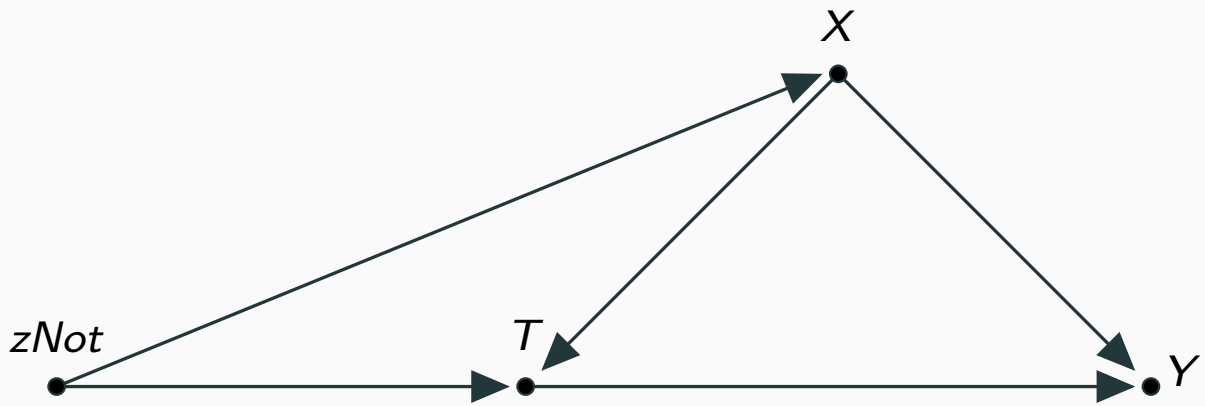
Question 1 – (b): bad instruments

The second is that the exclusion restriction is not met:



Question 1 – (b): bad instruments

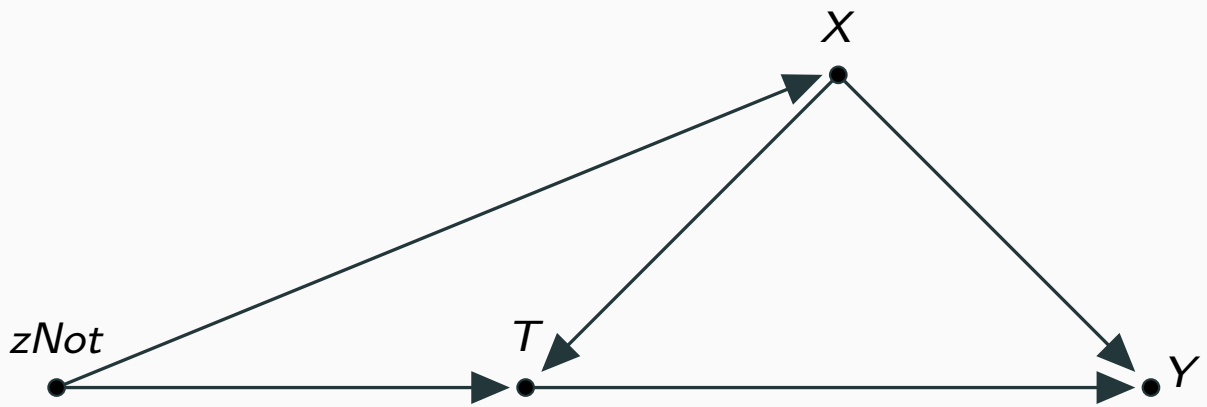
The second is that the exclusion restriction is not met:



How could we solve the problem in the picture?

Question 1 – (b): bad instruments

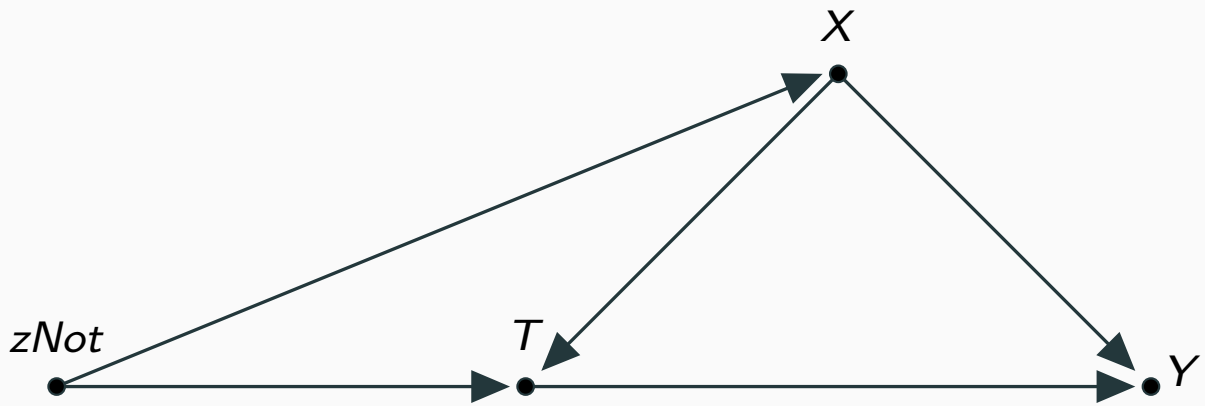
The second is that the exclusion restriction is not met:



How could we solve the problem in the picture? Controlling for `X`, provided we had information on that.

Question 1 – (b): bad instruments

The second is that the exclusion restriction is not met:



How could we solve the problem in the picture? Controlling for `X`, provided we had information on that. In real life, with observational data, we often **cannot** control for unobservable variable our instrument has an effect on.

Question 1 – (b): coding

R code to obtain these variables:

```r
X <- rpois(1000, lambda = 3)
Z <- rbinom(1000, size = 8, prob = 0.4)
T <- 2 + 3*Z - 2*X + rnorm(1000)
Y <- 1 + 2*T -3*X + rnorm(1000)

model.iv <- ivreg(Y ~ T | Z)
```

Stata code:

```
gen X = rpoisson(3)
gen Z = rbinomial(8, 0.4)
gen T = 2 + 3*Z - 2*X + rnormal()
gen Y = 1 + 2*T -3*X + rnormal()

ivreg Y (T = Z)
```

Question 1 – (b): coding

R code to obtain bad instruments (weak):

```r
zNot <- runif(1000)
T <- 2 + 3*Z - 2*X + 0*zNot + rnorm(1000)
# zNot is not a cause of T
Y <- 1 + 2*T -3*X + rnorm(1000)

model.iv.bad <- ivreg(Y ~ T | zNot)
```

Stata code:

```
gen zNot = runiform()
gen T = 2 + 3*Z - 2*X + 0*zNot + rnormal()
gen Y = 1 + 2*T -3*X + rnormal()

ivreg Y (T = zNot)
```

Question 1 – (b): coding

R code to obtain bad instruments (exclusion restriction violated):

```r
X <- 1 + 2*zNot + rnorm(1000)
T <- 2 + 3*Z - 2*X + 5*zNot + rnorm(1000)
# zNot IS a cause of T but also of X
Y <- 1 + 2*T -3*X + rnorm(1000)

model.iv.bad2 <- ivreg(Y ~ T | zNot)
```

Stata code:

```
gen X = 1 + 2*zNot + rnormal()
gen T = 2 + 3*Z - 2*X + 5*zNot + rnormal()
gen Y = 1 + 2*T -3*X + rnormal()

ivreg Y (T = zNot)
```

Results

	Linear (controls)	Good IV	Weak IV	Exclusion Restriction
(Intercept)	0.98^{***}	-8.19^{***}	-37.53	36.62^{**}
	(0.11)	(0.26)	(250.76)	(16.59)
T	1.99^{***}	2.05^{***}	7.27	-2.14
	(0.01)	(0.04)	(44.61)	(1.65)
X	-2.99^{***}			
	(0.02)			
R^2	1.00	0.89	-2.04	-2.55
Adj. R^2	1.00	0.89	-2.04	-2.55
Num. obs.	1000	1000	1000	1000

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Question 2

Question 2

Create a dataset with 50,000 observations. Create a variable called *Instrument*, which takes a value of 0 60% of the time and a value of 1 otherwise.

Question 2

Create a dataset with 50,000 observations. Create a variable called *Instrument*, which takes a value of 0 60% of the time and a value of 1 otherwise. Generate a variable *ObservableThing* $\sim N(0; 1)$ and a variable *UnobservableThing* which is also $\sim N(0; 1)$.

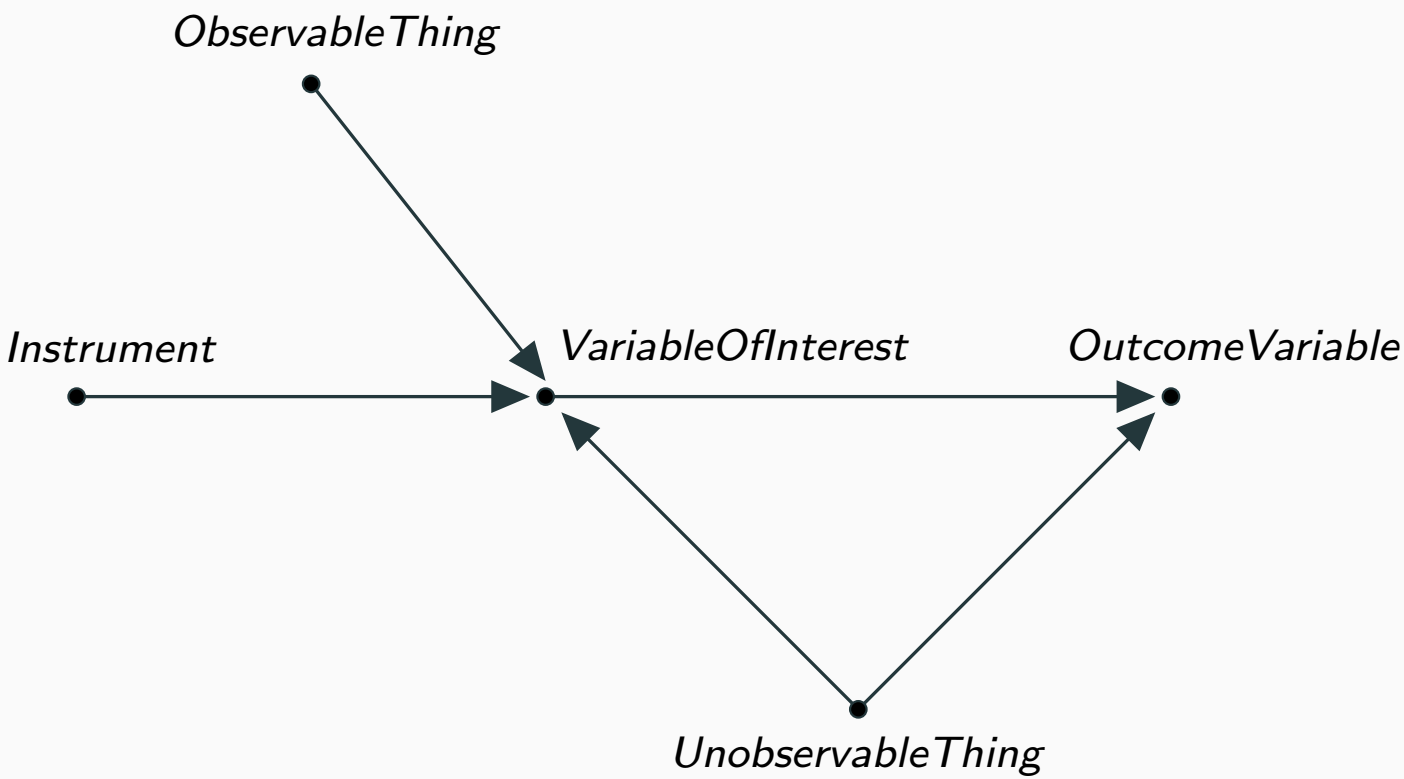
Question 2

Create a dataset with 50,000 observations. Create a variable called *Instrument*, which takes a value of 0 60% of the time and a value of 1 otherwise. Generate a variable *ObservableThing* $\sim N(0; 1)$ and a variable *UnobservableThing* which is also $\sim N(0; 1)$. Create the variable *VariableOfInterest*, which equals 1 if *ObservableThing* + *UnobservableThing* + *Instrument* ≥ 2.5 and which equals 0 otherwise.

Question 2

Create a dataset with 50,000 observations. Create a variable called *Instrument*, which takes a value of 0 60% of the time and a value of 1 otherwise. Generate a variable $ObservableThing \sim N(0; 1)$ and a variable *UnobservableThing* which is also $\sim N(0; 1)$. Create the variable *VariableOfInterest*, which equals 1 if $ObservableThing + UnobservableThing + Instrument \geq 2.5$ and which equals 0 otherwise. Finally, generate $OutcomeVariable = UnobservableThing + VariableofInterest + e$, where $e \sim N(0; 1)$.

Question 2 – Data Generating Process (causal diagram)



Question 2 – Data Generating Process (code)

In R:

```
Instrument <- rbinom(50000, size = 1, prob = 0.4)
ObservableThing <- rnorm(50000)
UnobservableThing <- rnorm(50000)

VariableOfInterest <-ifelse(ObservableThing +
    UnobservableThing + Instrument >= 2.5, 1, 0)

OutcomeVariable = UnobservableThing + 1 *
    VariableOfInterest + rnorm(50000)
```

Question 2 – Data Generating Process (code)

In Stata:

```
gen Instrument = rbinomial(1, 0.4)
gen ObservableThing = rnormal()
gen UnobservableThing = rnormal()

gen VariableOfInterest = (ObservableThing +
    UnobservableThing + Instrument >= 2.5)

gen OutcomeVariable = UnobservableThing + 1 *
    VariableOfInterest + rnormal()
```

Question 2 (a)

2. (a) (5 Marks) What is the causal effect of *VariableOfInterest* on *OutcomeVariable*?

Question 2 (a)

2. (a) (5 Marks) What is the causal effect of *VariableOfInterest* on *OutcomeVariable*?

Remember the equation that defines *OutcomeVariable*:

$$OutcomeVariable = UnobservableThing + VariableofInterest + e$$

Question 2 (a)

2. (a) (5 Marks) What is the causal effect of *VariableOfInterest* on *OutcomeVariable*?

Remember the equation that defines *OutcomeVariable*:

$$OutcomeVariable = UnobservableThing + VariableofInterest + e$$

The causal effect will have a size of 1!

Question 2 (b)

2. (b) (5 Marks) Regress *OutcomeVariable* on *VariableOfInterest* and interpret the coefficient. Why is this not close to the causal effect from part (a)?

Question 2 (b)

2. (b) (5 Marks) Regress *OutcomeVariable* on *VariableOfInterest* and interpret the coefficient. Why is this not close to the causal effect from part (a)?

Model:

Question 2 (b)

2. (b) (5 Marks) Regress *OutcomeVariable* on *VariableOfInterest* and interpret the coefficient. Why is this not close to the causal effect from part (a)?

Model: $\mathit{OutcomeVariable} = \alpha + \beta\,\mathit{VariableOfInterest} + u$

Question 2 (b)

2. (b) (5 Marks) Regress *OutcomeVariable* on *VariableOfInterest* and interpret the coefficient. Why is this not close to the causal effect from part (a)?

Model: $\mathit{OutcomeVariable} = \alpha + \beta\,\mathit{VariableOfInterest} + u$

The coefficient does not provide the causal effect of 1 from part (a) because we have *UnobservableThing* which is a confounder and:

Question 2 (b)

2. (b) (5 Marks) Regress *OutcomeVariable* on *VariableOfInterest* and interpret the coefficient. Why is this not close to the causal effect from part (a)?

Model: $\mathit{OutcomeVariable} = \alpha + \beta\,\mathit{VariableOfInterest} + u$

The coefficient does not provide the causal effect of 1 from part (a) because we have *UnobservableThing* which is a confounder and:

1. we are neither controlling for it

Question 2 (b)

2. (b) (5 Marks) Regress *OutcomeVariable* on *VariableOfInterest* and interpret the coefficient. Why is this not close to the causal effect from part (a)?

Model: $\mathit{OutcomeVariable} = \alpha + \beta\,\mathit{VariableOfInterest} + u$

The coefficient does not provide the causal effect of 1 from part (a) because we have *UnobservableThing* which is a confounder and:

1. we are neither controlling for it
2. nor instrumenting *VariableOfInterest*

Question 2 (b)

2. (b) (5 Marks) Regress *OutcomeVariable* on *VariableOfInterest* and interpret the coefficient. Why is this not close to the causal effect from part (a)?

Model: $\textit{OutcomeVariable} = \alpha + \beta\,\textit{VariableOfInterest} + u$

The coefficient does not provide the causal effect of 1 from part (a) because we have *UnobservableThing* which is a confounder and:

1. we are neither controlling for it
2. nor instrumenting *VariableOfInterest*

(I'll show all results together at the end of Question 2)

Question 2 (c)

2. (c) (10 Marks) Use *Instrument* as the instrumental variable for *VariableOfInterest* in the regression of *OutcomeVariable* on *VariableOfInterest*. Calculate the causal effect of *VariableOfInterest* on *OutcomeVariable* both by hand and using the proper commands in Stata or R. Comment on what you observe.

Question 2 (c)

2. (c) (10 Marks) Use *Instrument* as the instrumental variable for *VariableOfInterest* in the regression of *OutcomeVariable* on *VariableOfInterest*. Calculate the causal effect of *VariableOfInterest* on *OutcomeVariable* both by hand and using the proper commands in Stata or R. Comment on what you observe.

2SLS model:

Question 2 (c)

2. (c) (10 Marks) Use *Instrument* as the instrumental variable for *VariableOfInterest* in the regression of *OutcomeVariable* on *VariableOfInterest*. Calculate the causal effect of *VariableOfInterest* on *OutcomeVariable* both by hand and using the proper commands in Stata or R. Comment on what you observe.

2SLS model:

$\widehat{VariableOfInterest} = \hat{\gamma} + \hat{\delta} Instrument$

Question 2 (c)

2. (c) (10 Marks) Use *Instrument* as the instrumental variable for *VariableOfInterest* in the regression of *OutcomeVariable* on *VariableOfInterest*. Calculate the causal effect of *VariableOfInterest* on *OutcomeVariable* both by hand and using the proper commands in Stata or R. Comment on what you observe.

2SLS model:

$$\widehat{VariableOfInterest} = \hat{\gamma} + \hat{\delta} Instrument$$

$$OutcomeVariable = \hat{\alpha} + \hat{\beta} \widehat{VariableOfInterest} + \hat{u}$$

Question 2 (c)

2. (c) (10 Marks) Use *Instrument* as the instrumental variable for *VariableOfInterest* in the regression of *OutcomeVariable* on *VariableOfInterest*. Calculate the causal effect of *VariableOfInterest* on *OutcomeVariable* both by hand and using the proper commands in Stata or R. Comment on what you observe.

2SLS model:

$$\widehat{VariableOfInterest} = \hat{\gamma} + \hat{\delta} Instrument$$

$$OutcomeVariable = \hat{\alpha} + \hat{\beta} \widehat{VariableOfInterest} + \hat{u}$$

Also remember that the instrumented unbiased effect of T_i is:

Question 2 (c)

2. (c) (10 Marks) Use *Instrument* as the instrumental variable for *VariableOfInterest* in the regression of *OutcomeVariable* on *VariableOfInterest*. Calculate the causal effect of *VariableOfInterest* on *OutcomeVariable* both by hand and using the proper commands in Stata or R. Comment on what you observe.

2SLS model:

$$\widehat{VariableOfInterest} = \hat{\gamma} + \hat{\delta} Instrument$$

$$OutcomeVariable = \hat{\alpha} + \hat{\beta} \widehat{VariableOfInterest} + \hat{u}$$

Also remember that the instrumented unbiased effect of T_i is:

$$d = \frac{Cov(Y_i, Z_i)}{Cov(T_i, Z_i)}$$

Question 2 (c)

Code in R:

```
# 1st and 2nd stage
first.st <- lm(VariableOfInterest ~ Instrument)
fitted.var <- first.st$fitted.values
second.st <- lm(OutcomeVariable ~ fitted.var)

model.iv <- ivreg(OutcomeVariable ~ VariableOfInterest
    | Instrument)
```

Question 2 (c)

Code in R:

```r
# 1st and 2nd stage
first.st <- lm(VariableOfInterest ~ Instrument)
fitted.var <- first.st$fitted.values
second.st <- lm(OutcomeVariable ~ fitted.var)

model.iv <- ivreg(OutcomeVariable ~ VariableOfInterest
    | Instrument)
```

Code in Stata:

```
reg VariableOfInterest Instrument
predict fitted_var
reg OutcomeVariable fitted_var

ivreg OutcomeVariable (VariableOfInterest=Instrument)
```

Question 2 (c) – Comparison of results

	Manually	`ivreg`
(Intercept)	0.02	0.02
	(0.01)	(0.01)
VariableOfInterest	0.91***	0.91***
	(0.13)	(0.13)
Adj. R^2	0.00	0.11
Num. obs.	50000	50000

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Question 2 (c) – Comparison of results

	Manually	`ivreg`
(Intercept)	0.02	0.02
	(0.01)	(0.01)
VariableOfInterest	0.91***	0.91***
	(0.13)	(0.13)
Adj. R^2	0.00	0.11
Num. obs.	50000	50000

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Two notes:

Question 2 (c) – Comparison of results

	Manually	`ivreg`
(Intercept)	0.02	0.02
	(0.01)	(0.01)
VariableOfInterest	0.91***	0.91***
	(0.13)	(0.13)
Adj. R^2	0.00	0.11
Num. obs.	50000	50000

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Two notes:

1. SEs will be slightly different because of the different vCov matrices, but parameters will be *identical*

Question 2 (c) – Comparison of results

	Manually	`ivreg`
(Intercept)	0.02	0.02
	(0.01)	(0.01)
VariableOfInterest	0.91***	0.91***
	(0.13)	(0.13)
Adj. R^2	0.00	0.11
Num. obs.	50000	50000

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Two notes:

1. SEs will be slightly different because of the different vCov matrices, but parameters will be *identical*
2. R^2 in the 1st model. We have a good instrument yet the variation we explain is low.

Question 2 (c) – Comparison of results

	Manually	`ivreg`
(Intercept)	0.02	0.02
	(0.01)	(0.01)
VariableOfInterest	0.91***	0.91***
	(0.13)	(0.13)
Adj. R^2	0.00	0.11
Num. obs.	50000	50000

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Two notes:

1. SEs will be slightly different because of the different vCov matrices, but parameters will be *identical*
2. R^2 in the 1st model. We have a good instrument yet the variation we explain is low. Don't over-trust the R^2!

Question 2 (d)

2. (d) (10 Marks) Change the setup from part (a) such that *Instrument* is only equal to 1 2% of the time. Then rerun the regression in part (c) (using Stata or R) and comment on what happens. What's wrong with this instrument?

Question 2 (d)

2. (d) (10 Marks) Change the setup from part (a) such that *Instrument* is only equal to 1 2% of the time. Then rerun the regression in part (c) (using Stata or R) and comment on what happens. What's wrong with this instrument?

If we only change the instrument and do not change the rest of the DGP, we will have a new instrument which will be very weakly correlated with the previous one and thus with *OutcomeVariable*, thus it will give us a bad estimate of the causal effect of *VariableOfInterest*

Question 2 (d)

Code in R:

```
Instrument2 <- rbinom(50000, size = 1, prob = 0.02)
cor(Instrument, Instrument2) # 0.0015: very very low!
cor(VariableOfInterest, Instrument2) # 0.0053: very
    low too

model.iv.weak <- ivreg(OutcomeVariable ~
    VariableOfInterest | Instrument2)
```

Question 2 (d)

Code in R:

```r
Instrument2 <- rbinom(50000, size = 1, prob = 0.02)
cor(Instrument, Instrument2) # 0.0015: very very low!
cor(VariableOfInterest, Instrument2) # 0.0053: very
    low too

model.iv.weak <- ivreg(OutcomeVariable ~
    VariableOfInterest | Instrument2)
```

Code in Stata:

```
gen Instrument2 = rbinomial(1, 0.02)
corr Instrument Instrument2
corr VariableOfInterest Instrument2

ivreg OutcomeVariable (VariableOfInterest =
    Instrument2)
```

Question 2 (e)

2. (e) (10 Marks) Change the setup from part a such that $VariableOfInterest = Obs.Thing + Unobs.Thing + 0.05 * Instr$. where *Instrument* still takes a value of 0 60% of the time and a value of 1 otherwise. Then rerun the regression in part (c) (using Stata or R) and comment on what happens.

Question 2 (e)

2. (e) (10 Marks) Change the setup from part a such that $\mathit{VariableOfInterest} = \mathit{Obs}.\mathit{Thing} + \mathit{Unobs}.\mathit{Thing} + 0.05 * \mathit{Instr}$. where *Instrument* still takes a value of 0 60% of the time and a value of 1 otherwise. Then rerun the regression in part (c) (using Stata or R) and comment on what happens.

The instrument will be a very weak determinant of *VariableOfInterest* (its causal effect is of size 0.05).

Question 2 (e)

2. (e) (10 Marks) Change the setup from part a such that $\mathit{VariableOfInterest} = \mathit{Obs}.\mathit{Thing} + \mathit{Unobs}.\mathit{Thing} + 0.05 * \mathit{Instr}$. where *Instrument* still takes a value of 0 60% of the time and a value of 1 otherwise. Then rerun the regression in part (c) (using Stata or R) and comment on what happens.

The instrument will be a very weak determinant of *VariableOfInterest* (its causal effect is of size 0.05). Thus it will give us a biased estimate of the causal effect of *VariableOfInterest* on *OutcomeVariable*

Question 2 (e)

Code in R:

```
VariableOfInterest = ObservableThing +
    UnobservableThing + 0.05*Instrument
model.iv2 <- ivreg(OutcomeVariable ~
    VariableOfInterest | Instrument)
```

Question 2 (e)

Code in R:

```
VariableOfInterest = ObservableThing +
    UnobservableThing + 0.05*Instrument
model.iv2 <- ivreg(OutcomeVariable ~
    VariableOfInterest | Instrument)
```

Code in Stata:

```
gen VariableOfInterest = ObservableThing +
    UnobservableThing + 0.05*Instrument
ivreg OutcomeVariable (VariableOfInterest = Instrument
    )
```

Question 2 (f)

2. (f) (10 Marks) Change the setup from part a such that $OutcomeVariable = Unobs.Thing + Var.ofInt. + 0.05 * Instr. + e$ where *Instrument* still takes a value of 0 60% of the time and a value of 1 otherwise. Then rerun the regression in part (c) (using Stata or R) and comment on what happens.

Question 2 (f)

2. (f) (10 Marks) Change the setup from part a such that $OutcomeVariable = Unobs.Thing + Var.ofInt. + 0.05 * Instr. + e$ where *Instrument* still takes a value of 0 60% of the time and a value of 1 otherwise. Then rerun the regression in part (c) (using Stata or R) and comment on what happens.

The instrument will be a determinant of *OutcomeVariable*, exclusion restriction is not met.

Question 2 (f)

2. (f) (10 Marks) Change the setup from part a such that $OutcomeVariable = Unobs.Thing + Var.ofInt. + 0.05 * Instr. + e$ where *Instrument* still takes a value of 0 60% of the time and a value of 1 otherwise. Then rerun the regression in part (c) (using Stata or R) and comment on what happens.

The instrument will be a determinant of *OutcomeVariable*, exclusion restriction is not met.

Even if its causal effect on *OutcomeVariable* will be weak (0.05), this will be enough to bias the estimate of the causal effect of *VariableOfInterest* when it is instrumented using *Instrument*.

Question 2 (e)

Code in R:

```
OutcomeVariable = UnobservableThing +
    VariableOfInterest + 0.05*Instrument + rnorm
    (50000)
model.iv3 <- ivreg(OutcomeVariable ~
    VariableOfInterest | Instrument)
```

Question 2 (e)

Code in R:

```
OutcomeVariable = UnobservableThing +
    VariableOfInterest + 0.05*Instrument + rnorm
    (50000)
model.iv3 <- ivreg(OutcomeVariable ~
    VariableOfInterest | Instrument)
```

Code in Stata:

```
gen OutcomeVariable = UnobservableThing +
    VariableOfInterest + 0.05*Instrument + rnormal()

ivreg OutcomeVariable (VariableOfInterest = Instrument
    )
```

Question 2 – comparison of results

	(b)	(c)	(d)	(e)	(f)
(Intercept)	-0.10^{***}	0.02	-0.53	-0.01	-0.02
	(0.01)	(0.01)	(0.49)	(0.04)	(0.02)
VariableOfInterest	2.34^{***}	0.91^{***}	7.70	3.15^{***}	2.47^{***}
	(0.02)	(0.13)	(6.20)	(1.18)	(0.57)
R^2	0.18	0.11	-0.75	-5.40	0.43
Adj. R^2	0.18	0.11	-0.75	-5.40	0.43
Num. obs.	50000	50000	50000	50000	50000

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Question 3

Question 3

`USStateLegislature.csv`:

- vote share of the female candidates in U.S. state legislature elections (`votemargin`) by electoral district

Question 3

`USStateLegislature.csv`:

- vote share of the female candidates in U.S. state legislature elections (`votemargin`) by electoral district
- turnout **among women in the NEXT election** by electoral district. (`turnout`).

Question 3

`USStateLegislature.csv`:

- vote share of the female candidates in U.S. state legislature elections (`votemargin`) by electoral district
- turnout **among women in the NEXT election** by electoral district. (`turnout`).
- vote share of the democratic presidential candidate in that electoral district and year (`democraticvoteshare_president`).

Question 3

`USStateLegislature.csv`:

- vote share of the female candidates in U.S. state legislature elections (`votemargin`) by electoral district
- turnout **among women in the NEXT election** by electoral district. (`turnout`).
- vote share of the democratic presidential candidate in that electoral district and year (`democraticvoteshare_president`).

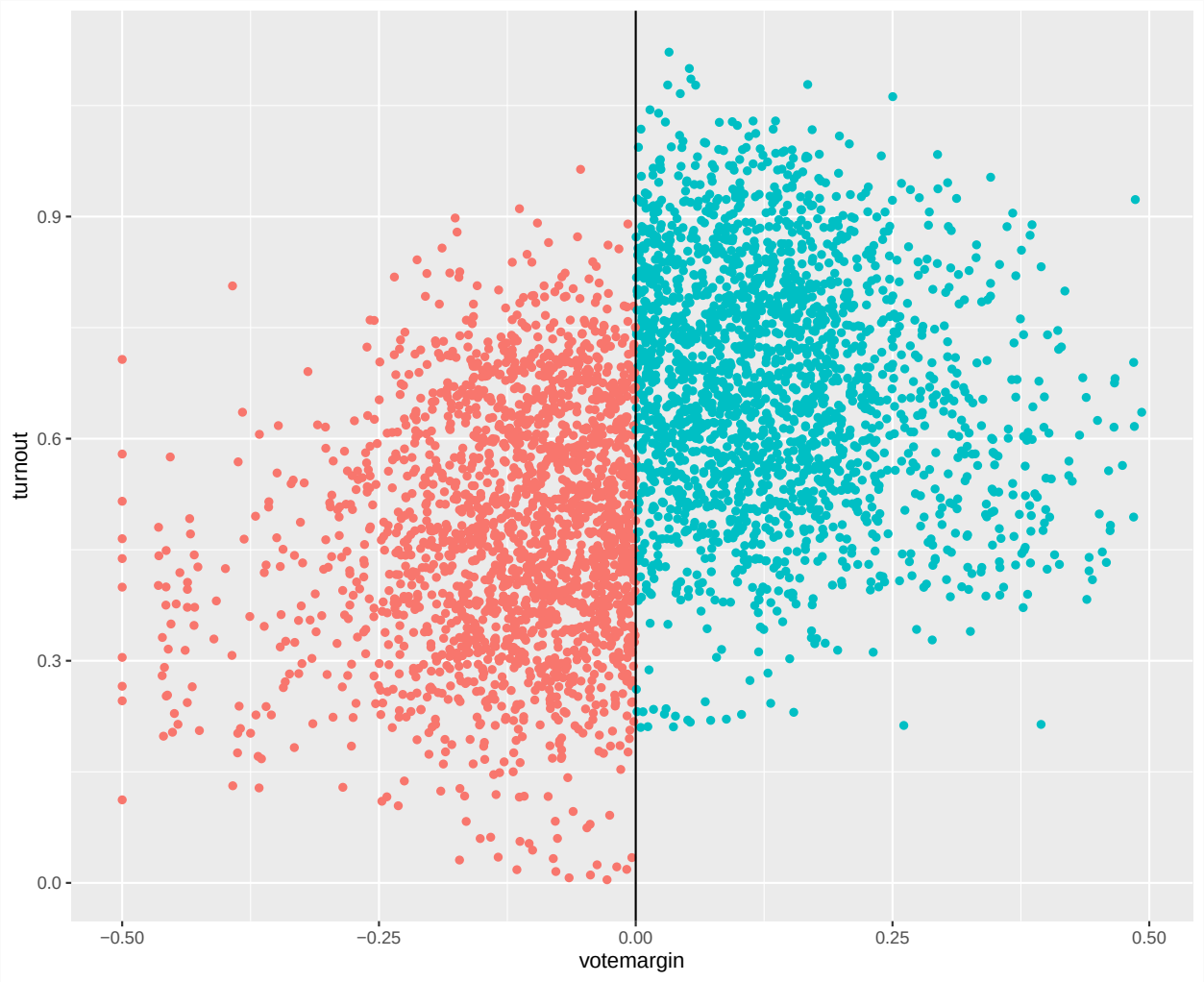
Res. Q: Is there an effect of having women enter elected office in one election cycle on women's turnout in the next election cycle?

Question 3 (a)

3. (a) (3 marks) Generate a scatter plot `turnout` vs `votemargin` and add a vertical line at $votemargin = 0$. What do you see?

Question 3 (a)

3. (a) (3 marks) Generate a scatter plot `turnout` vs `votemargin` and add a vertical line at *votemargin* $= 0$. What do you see?



Question 3 (a) – coding

In R (I use the `indicator` variable from point b):

```
data %>% ggplot(aes(y = turnout, x = votemargin)) +
    geom_point(aes(colour = indicator)) +
  geom_vline(xintercept = 0) + theme(legend.position =
    "none") # this way we get rid of the legend
```

Question 3 (a) – coding

In R (I use the `indicator` variable from point b):

```
data %>% ggplot(aes(y = turnout, x = votemargin)) +
    geom_point(aes(colour = indicator)) +
  geom_vline(xintercept = 0) + theme(legend.position =
    "none") # this way we get rid of the legend
```

In Stata:

```
twoway ///
  (scatter turnout votemargin if votemargin >= 0) ///
  (scatter turnout votemargin if votemargin < 0),
    xline(0) leg(off)
```

Question 3 (a) – coding

In R (I use the `indicator` variable from point b):

```
data %>% ggplot(aes(y = turnout, x = votemargin)) +
    geom_point(aes(colour = indicator)) +
  geom_vline(xintercept = 0) + theme(legend.position =
    "none") # this way we get rid of the legend
```

In Stata:

```
twoway ///
  (scatter turnout votemargin if votemargin >= 0) ///
  (scatter turnout votemargin if votemargin < 0),
    xline(0) leg(off)
```

Districts where female candidates barely lost have a generally lower female turnout in the subsequent election than those were female candidates barely won

Question 3 (b)

3. (b) (7 marks) Draw lines through the data cloud below and above $votemargin = 0$. Lay a locally smoothed curve, a linear fit line, or a quadratic fit line over the scatter plot. What line approximates the data best (just eyeball)?

Question 3 (b) – coding in R

Various types of fit lines:

```
# indicator:
data$indicator[data$votemargin >= 0] <- 1
data$indicator[data$votemargin < 0] <- 0

# first save plot:
p <- data %>% ggplot(aes(y = turnout, x = votemargin))
    + geom_point(aes(colour = indicator)) +
  geom_vline(xintercept = 0) + theme(legend.position =
    "none")

# different fits:
p + geom_smooth(aes(group = indicator))
p + geom_smooth(aes(group = indicator), method = "lm")
p + geom_smooth(aes(group = indicator), method = "lm",
    formula = y ~ x + I(x^2))
```

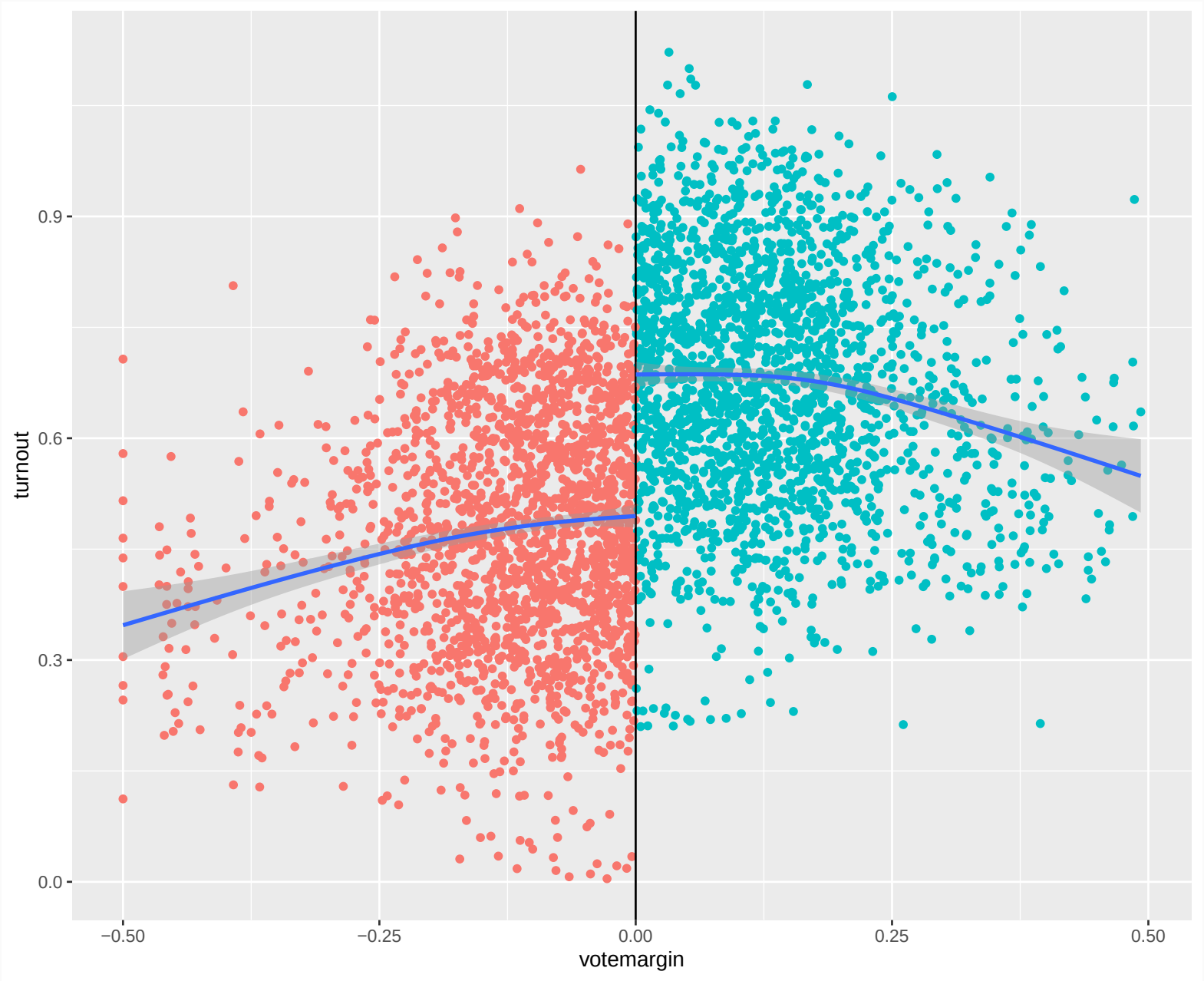
Question 3 (b) – coding in Stata

Various types of fit lines:

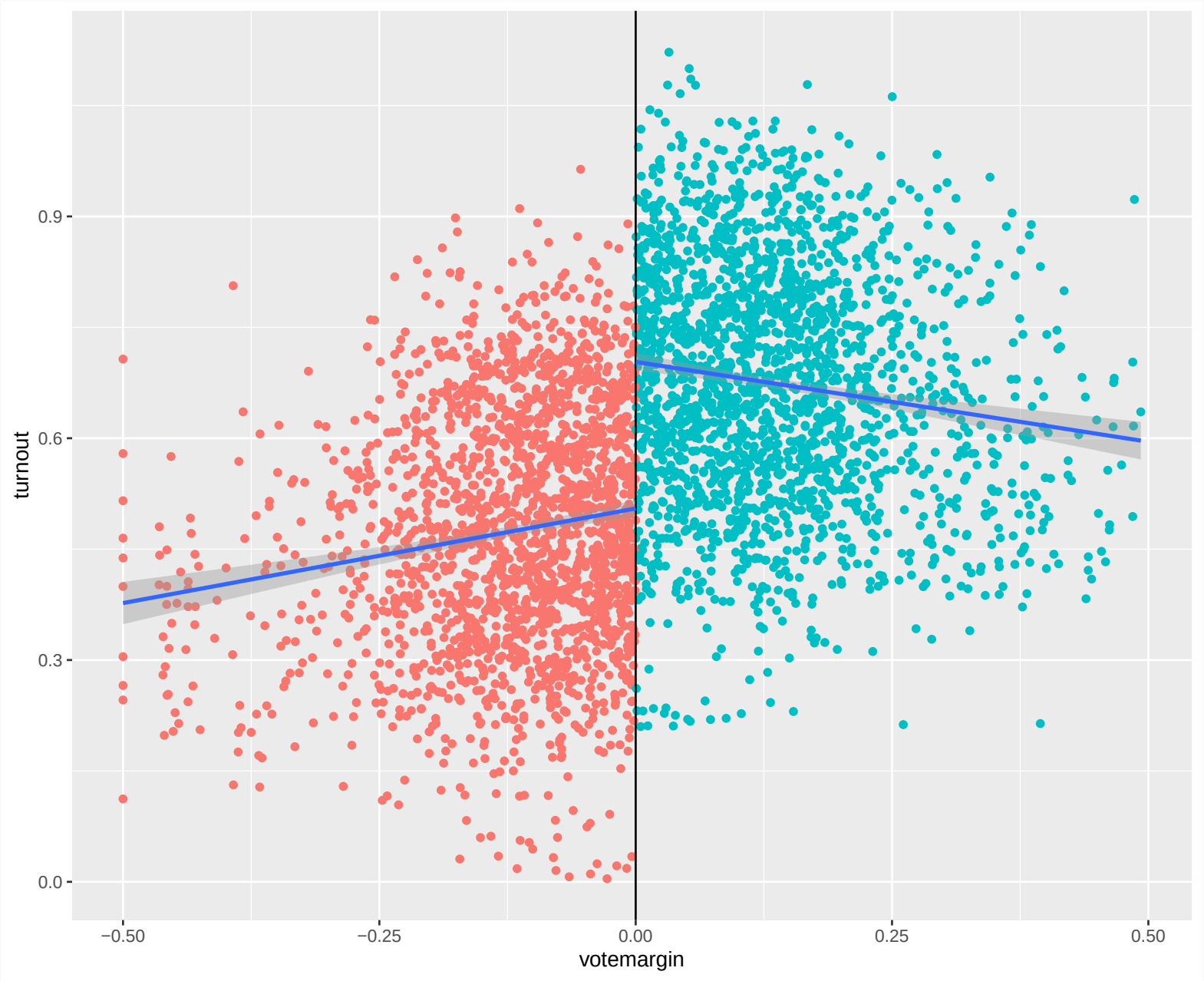
```
* linear model
twoway (scatter turnout votemargin if votemargin >= 0)
    ///
  (scatter turnout votemargin if votemargin < 0) ///
  (lfit turnout votemargin if votemargin >= 0) ///
  (lfit turnout votemargin if votemargin < 0), xline
    (0) leg(off)

* quadratic model
twoway (scatter turnout votemargin if votemargin >= 0)
    ///
  (scatter turnout votemargin if votemargin < 0) ///
  (qfit turnout votemargin if votemargin >= 0) ///
  (qfit turnout votemargin if votemargin < 0), xline
    (0) leg(off)
```

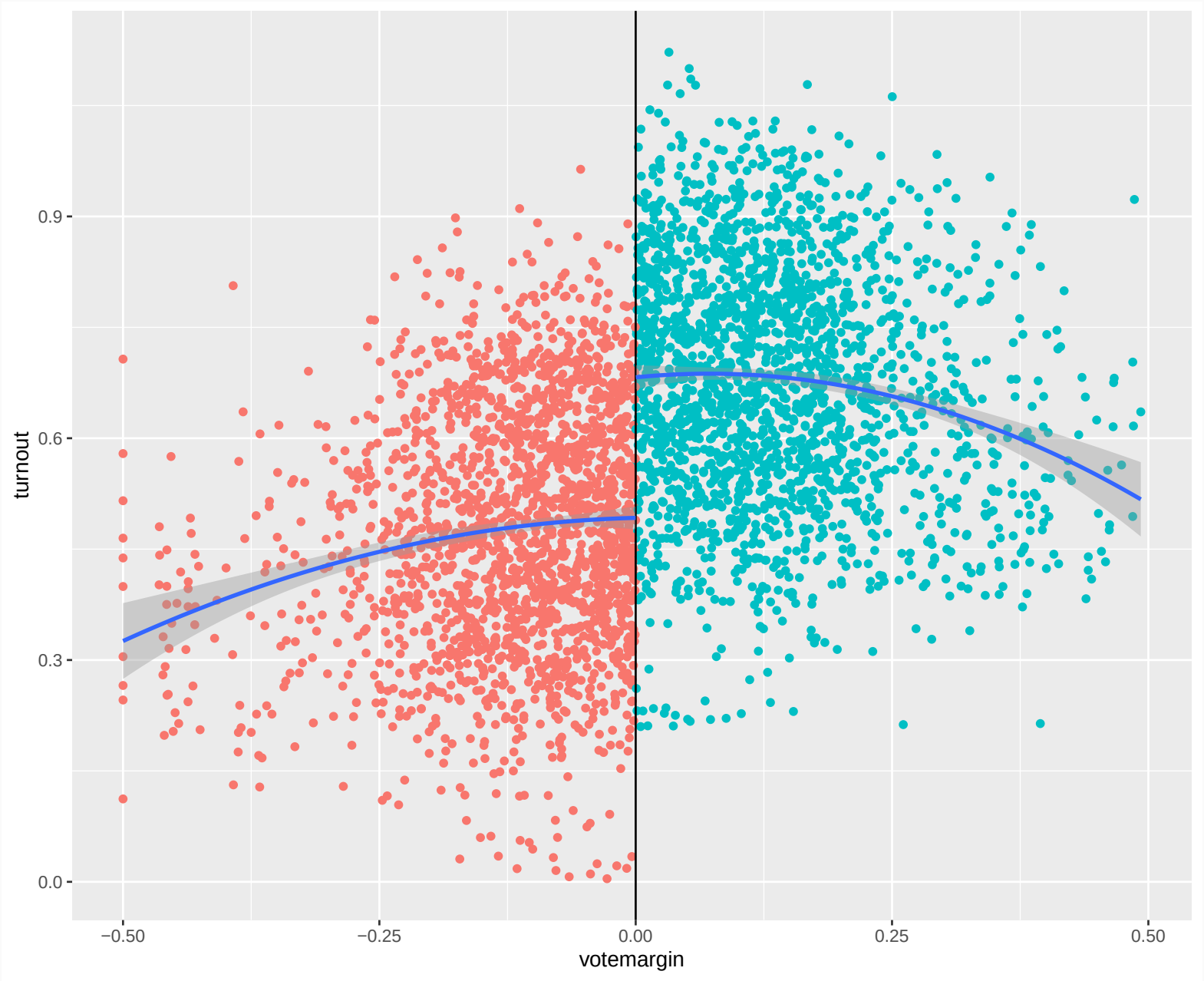
Question 3 (b) – Locally smoothed fit



Question 3 (b) – Linear fit



Question 3 (b) – Quadratic fit



Question 3 (c)

3. (c) (15 marks) Estimate the effect of having a woman enter elected office on female turnout.

Question 3 (c)

3. (c) (15 marks) Estimate the effect of having a woman enter elected office on female turnout.
    - For R, use the `RDestimate` command from the `rdd` package.

Question 3 (c)

3. (c) (15 marks) Estimate the effect of having a woman enter elected office on female turnout.
    - For R, use the `RDestimate` command from the `rdd` package.
    - In Stata use `rdrobust` command (type `findit rdrobust`)

Question 3 (c)

3. (c) (15 marks) Estimate the effect of having a woman enter elected office on female turnout.
   - For R, use the `RDestimate` command from the `rdd` package.
   - In Stata use `rdrobust` command (type `findit rdrobust`)

   Set a proper cut-point for the forcing variable `votemargin`.

Question 3 (c)

3. (c) (15 marks) Estimate the effect of having a woman enter elected office on female turnout.
   - For R, use the `RDestimate` command from the `rdd` package.
   - In Stata use `rdrobust` command (type `findit rdrobust`)

   Set a proper cut-point for the forcing variable `votemargin`.
   - What is the estimated causal effect of a woman elected to the legislature on `turnout`?

Question 3 (c)

3. (c) (15 marks) Estimate the effect of having a woman enter elected office on female turnout.
   - For R, use the `RDestimate` command from the `rdd` package.
   - In Stata use `rdrobust` command (type `findit rdrobust`)

   Set a proper cut-point for the forcing variable `votemargin`.
   - What is the estimated causal effect of a woman elected to the legislature on `turnout`?
   - What type of average treatment effect are you estimating here?

Question 3 (c)

3. (c) (15 marks) Estimate the effect of having a woman enter elected office on female turnout.
   - For R, use the `RDestimate` command from the `rdd` package.
   - In Stata use `rdrobust` command (type `findit rdrobust`)

   Set a proper cut-point for the forcing variable `votemargin`.
   - What is the estimated causal effect of a woman elected to the legislature on `turnout`?
   - What type of average treatment effect are you estimating here?

   In R and Stata you can specify different bandwidth manually or choose between different bandwidth selection mechanism for estimation.

Question 3 (c)

3. (c) (15 marks) Estimate the effect of having a woman enter elected office on female turnout.
   - For R, use the `RDestimate` command from the `rdd` package.
   - In Stata use `rdrobust` command (type `findit rdrobust`)

   Set a proper cut-point for the forcing variable `votemargin`.
   - What is the estimated causal effect of a woman elected to the legislature on `turnout`?
   - What type of average treatment effect are you estimating here?

   In R and Stata you can specify different bandwidth manually or choose between different bandwidth selection mechanism for estimation.
   - What does bandwidth selection have to do with the trade-off between variance and bias of the estimate of the causal effect?

Question 3 (c) – coding

In R:

```
model <- RDestimate(turnout ~ votemargin, data = data,
    cutpoint = 0, se.type = "HC1")

# specify bandwidth:
model2 <- RDestimate(turnout ~ votemargin, data = data
   , cutpoint = 0, bw = c(0.10), se.type = "HC1")
```

Question 3 (c) – coding

In R:

```r
model <- RDestimate(turnout ~ votemargin, data = data,
     cutpoint = 0, se.type = "HC1")

# specify bandwidth:
model2 <- RDestimate(turnout ~ votemargin, data = data
    , cutpoint = 0, bw = c(0.10), se.type = "HC1")
```

In Stata:

```
rdrobust turnout votemargin, c(0)  vce(hc1)

* specify bandwidth:
rdrobust turnout votemargin, c(0) b(-.2 .2)  vce(hc1)
```

Question 3 (c) – results

```
           Bd.w.   Obs.  Estimates   SE.
LATE       0.2038  3296  0.1926 ***  0.011916
Half-BW    0.1019  1830  0.2008 ***  0.016068
Double-BW  0.4076  3999  0.1951 ***  0.009649
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

F-statistics:
           F      Num. DoF  Denom. DoF  p
LATE       393.7  3         3292        0
Half-BW    212.8  3         1826        0
Double-BW  502.4  3         3995        0
```

Question 3 (c) – results

```
           Bd.w.    Obs.  Estimates    SE.
LATE       0.2038   3296  0.1926 ***   0.011916
Half-BW    0.1019   1830  0.2008 ***   0.016068
Double-BW  0.4076   3999  0.1951 ***   0.009649
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

F-statistics:
           F       Num. DoF  Denom. DoF  p
LATE       393.7   3         3292        0
Half-BW    212.8   3         1826        0
Double-BW  502.4   3         3995        0
```

Here we are estimating a **Local** Average Treatment Effect

Question 3 (c) – results

```
           Bd.w.   Obs.  Estimates   SE.
LATE       0.2038  3296  0.1926 ***  0.011916
Half-BW    0.1019  1830  0.2008 ***  0.016068
Double-BW  0.4076  3999  0.1951 ***  0.009649
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

F-statistics:
           F      Num. DoF  Denom. DoF  p
LATE       393.7  3         3292        0
Half-BW    212.8  3         1826        0
Double-BW  502.4  3         3995        0
```

Here we are estimating a **Local** Average Treatment Effect

Trade-off: enlarging the bandwith gives us more observations (smaller variance), but the unobservable confounders get stronger

Question 3 (d)

3. (d) (15 marks) For assessing the validity of the RDD estimate of the causal effect it is crucial to check whether the assumptions allowing RDD to make causal claims are met. The three most important assumptions are

Question 3 (d)

3. (d) (15 marks) For assessing the validity of the RDD estimate of the causal effect it is crucial to check whether the assumptions allowing RDD to make causal claims are met. The three most important assumptions are
   (1) The assignment of T_i by the forcing variable x_i is not manipulable.
   (2) Treatment effect only occurs at cut-off x_0 not at other cut-offs.
   (3) T_i only affects the outcome variable Y_i but not any other non-outcome covariates.

Question 3 (d)

3. (d) (15 marks) For assessing the validity of the RDD estimate of the causal effect it is crucial to check whether the assumptions allowing RDD to make causal claims are met. The three most important assumptions are
   (1) The assignment of T_i by the forcing variable x_i is not manipulable.
   (2) Treatment effect only occurs at cut-off x_0 not at other cut-offs.
   (3) T_i only affects the outcome variable Y_i but not any other non-outcome covariates.

   Here are your tasks:
   i. Explain the meaning of each assumption in your own words in 2-3 sentences.
   ii. Which of these three assumptions can be empirical tested?
   iii. Provide empirical tests for those assumptions you identified in (ii.) that can be tested using `USStateLegislature.csv`.

Question 3 (d) – i and ii

(1) The assignment of T_i by the forcing variable x_i is not manipulable.

Question 3 (d) – i and ii

(1) The assignment of T_i by the forcing variable x_i is not manipulable. No observation i can self-select in either being "treated" or not (begin below or above the threshold).

Question 3 (d) – i and ii

(1) The assignment of T_i by the forcing variable x_i is not manipulable. No observation i can self-select in either being "treated" or not (begin below or above the threshold). Otherwise unobservable factors accounting for the self-selection should be modelled!

Question 3 (d) – i and ii

(1) The assignment of T_i by the forcing variable x_i is not manipulable. No observation i can self-select in either being "treated" or not (begin below or above the threshold). Otherwise unobservable factors accounting for the self-selection should be modelled! **We cannot test it**

Question 3 (d) – i and ii

(1) The assignment of T_i by the forcing variable x_i is not manipulable. No observation i can self-select in either being "treated" or not (begin below or above the threshold). Otherwise unobservable factors accounting for the self-selection should be modelled! **We cannot test it**

(2) Treatment effect only occurs at cut-off x_0 not at other cut-offs.

Question 3 (d) – i and ii

(1) The assignment of T_i by the forcing variable x_i is not manipulable. No observation i can self-select in either being "treated" or not (begin below or above the threshold). Otherwise unobservable factors accounting for the self-selection should be modelled! **We cannot test it**

(2) Treatment effect only occurs at cut-off x_0 not at other cut-offs. There are no other cut-off points that could be used to distinguish between treatment and control groups, and the cut-off is the same for all observations.

Question 3 (d) – i and ii

(1) The assignment of T_i by the forcing variable x_i is not manipulable. No observation i can self-select in either being "treated" or not (begin below or above the threshold). Otherwise unobservable factors accounting for the self-selection should be modelled! **We cannot test it**

(2) Treatment effect only occurs at cut-off x_0 not at other cut-offs. There are no other cut-off points that could be used to distinguish between treatment and control groups, and the cut-off is the same for all observations. **We can test it**

Question 3 (d) – i and ii

(1) The assignment of T_i by the forcing variable x_i is not manipulable. No observation i can self-select in either being "treated" or not (begin below or above the threshold). Otherwise unobservable factors accounting for the self-selection should be modelled! **We cannot test it**

(2) Treatment effect only occurs at cut-off x_0 not at other cut-offs. There are no other cut-off points that could be used to distinguish between treatment and control groups, and the cut-off is the same for all observations. **We can test it**

(3) T_i only affects the outcome variable Y_i but not any other non-outcome covariates.

Question 3 (d) – i and ii

(1) The assignment of T_i by the forcing variable x_i is not manipulable. No observation i can self-select in either being "treated" or not (begin below or above the threshold). Otherwise unobservable factors accounting for the self-selection should be modelled! **We cannot test it**

(2) Treatment effect only occurs at cut-off x_0 not at other cut-offs. There are no other cut-off points that could be used to distinguish between treatment and control groups, and the cut-off is the same for all observations. **We can test it**

(3) T_i only affects the outcome variable Y_i but not any other non-outcome covariates. The effect isolated by the threshold is uniquely determined by the treatment variable and impacts only the outcome variable.

Question 3 (d) – i and ii

(1) The assignment of T_i by the forcing variable x_i is not manipulable. No observation i can self-select in either being "treated" or not (begin below or above the threshold). Otherwise unobservable factors accounting for the self-selection should be modelled! **We cannot test it**

(2) Treatment effect only occurs at cut-off x_0 not at other cut-offs. There are no other cut-off points that could be used to distinguish between treatment and control groups, and the cut-off is the same for all observations. **We can test it**

(3) T_i only affects the outcome variable Y_i but not any other non-outcome covariates. The effect isolated by the threshold is uniquely determined by the treatment variable and impacts only the outcome variable. All other covariates remain unchanged.

Question 3 (d) – i and ii

(1) The assignment of T_i by the forcing variable x_i is not manipulable. No observation i can self-select in either being "treated" or not (begin below or above the threshold). Otherwise unobservable factors accounting for the self-selection should be modelled! **We cannot test it**

(2) Treatment effect only occurs at cut-off x_0 not at other cut-offs. There are no other cut-off points that could be used to distinguish between treatment and control groups, and the cut-off is the same for all observations. **We can test it**

(3) T_i only affects the outcome variable Y_i but not any other non-outcome covariates. The effect isolated by the threshold is uniquely determined by the treatment variable and impacts only the outcome variable. All other covariates remain unchanged. **We can test it so far as we have information on the covariates**

Question 3 (d) – iii – Test assumption 2

Question 3 (d) – iii – Test assumption 2

We can run different models testing different cut-off points and expect to have **no** significant result

Question 3 (d) – iii – Test assumption 2

We can run different models testing different cut-off points and expect to have **no** significant result

In R:

```
RDestimate(turnout~votemargin,cutpoint=-.1,data=data,
    se.type = "HC1") # not significant
RDestimate(turnout~votemargin,cutpoint=.1,data=data,
    se.type = "HC1") # not significant
```

Question 3 (d) – iii – Test assumption 2

We can run different models testing different cut-off points and expect to have **no** significant result

In R:

```
RDestimate(turnout~votemargin,cutpoint=-.1,data=data,
    se.type = "HC1") # not significant
RDestimate(turnout~votemargin,cutpoint=.1,data=data,
    se.type = "HC1") # not significant
```

In Stata:

```
rdrobust turnout votemargin, c(-.1) vce(hc1)
rdrobust turnout votemargin, c(.1) vce(hc1)
```

Question 3 (d) – iii – Test assumption 3

Question 3 (d) – iii – Test assumption 3

Re-run all the analysis using covariates as dependent variables (hopefully no significant results).

Question 3 (d) – iii – Test assumption 3

Re-run all the analysis using covariates as dependent variables (hopefully no significant results).

In R:

```
data %>% ggplot(aes(y=democraticvoteshare_president,x=
    votemargin)) + geom_point(aes(col = indicator)) +
  geom_vline(xintercept=0)
RDestimate(democraticvoteshare_president ~ votemargin,
    cutpoint = 0, data = data, se.type = "HC1")
```

Question 3 (d) – iii – Test assumption 3

Re-run all the analysis using covariates as dependent variables (hopefully no significant results).

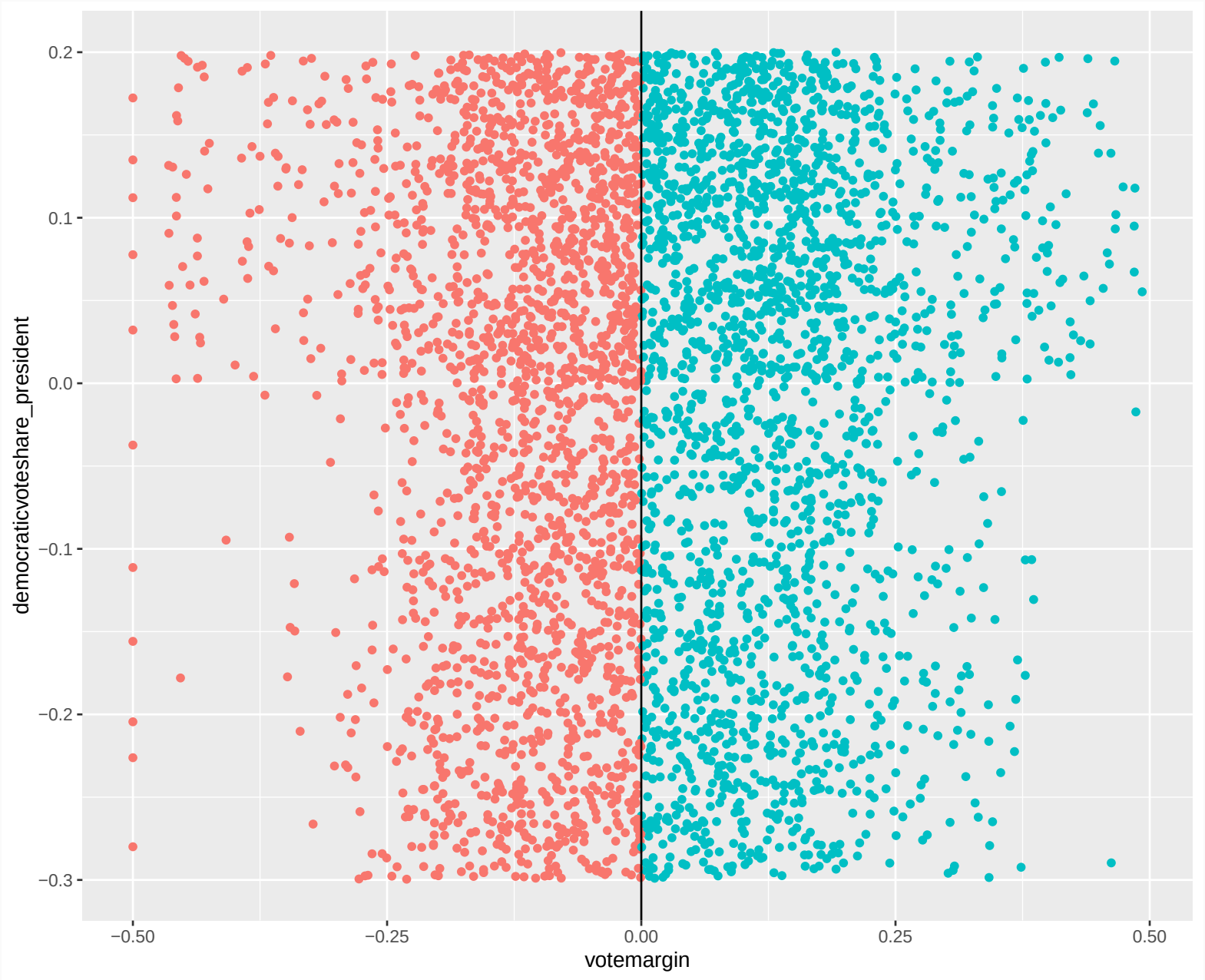
In R:

```
data %>% ggplot(aes(y=democraticvoteshare_president,x=
    votemargin)) + geom_point(aes(col = indicator)) +
  geom_vline(xintercept=0)
RDestimate(democraticvoteshare_president ~ votemargin,
     cutpoint = 0, data = data, se.type = "HC1")
```

In Stata:

```
twoway (scatter democraticvoteshare_president
    votemargin if votemargin >= 0) ///
  (scatter democraticvoteshare_president votemargin if
     votemargin < 0), xline(0) leg(off)
rdrobust democraticvoteshare_president votemargin, c
    (0)  vce(hc1)
```

Question 3 (d) – iii – Test assumption 3



Question 3 (d) – iii – Test assumption 1?

Assumption 1:

Question 3 (d) – iii – Test assumption 1?

Assumption 1:

- In principle not testable

Question 3 (d) – iii – Test assumption 1?

Assumption 1:

- In principle not testable
- Yet, we can get some **evidence** that could speak to it

Question 3 (d) – iii – Test assumption 1?

Assumption 1:

- In principle not testable
- Yet, we can get some **evidence** that could speak to it
- The assumption tells us that manipulation/sorting into T and C is not possible: evaluate by density plot.

Question 3 (d) – iii – Test assumption 1?

Assumption 1:

- In principle not testable
- Yet, we can get some **evidence** that could speak to it
- The assumption tells us that manipulation/sorting into T and C is not possible: evaluate by density plot.
- If there is no increased density around the cut-off, we probably do not observe manipulation.

Question 3 (d) – iii – Test assumption 1?

Assumption 1:

- In principle not testable
- Yet, we can get some **evidence** that could speak to it
- The assumption tells us that manipulation/sorting into T and C is not possible: evaluate by density plot.
- If there is no increased density around the cut-off, we probably do not observe manipulation.

In R:

Question 3 (d) – iii – Test assumption 1?

Assumption 1:

- In principle not testable
- Yet, we can get some **evidence** that could speak to it
- The assumption tells us that manipulation/sorting into T and C is not possible: evaluate by density plot.
- If there is no increased density around the cut-off, we probably do not observe manipulation.

In R:

```
DCdensity(data$votemargin,0, verbose=T, plot=T)
```

Question 3 (d) – iii – Test assumption 1?

Assumption 1:

- In principle not testable
- Yet, we can get some **evidence** that could speak to it
- The assumption tells us that manipulation/sorting into T and C is not possible: evaluate by density plot.
- If there is no increased density around the cut-off, we probably do not observe manipulation.

In R:

```
DCdensity(data$votemargin,0, verbose=T, plot=T)
```

The command comes with a test of the null-hypothesis that there is no difference in the densities before and after the discontinuity. We do not reject the null ($p - value = 0.17$)

Question 3 (d) – iii – Test assumption 1?

Assumption 1:

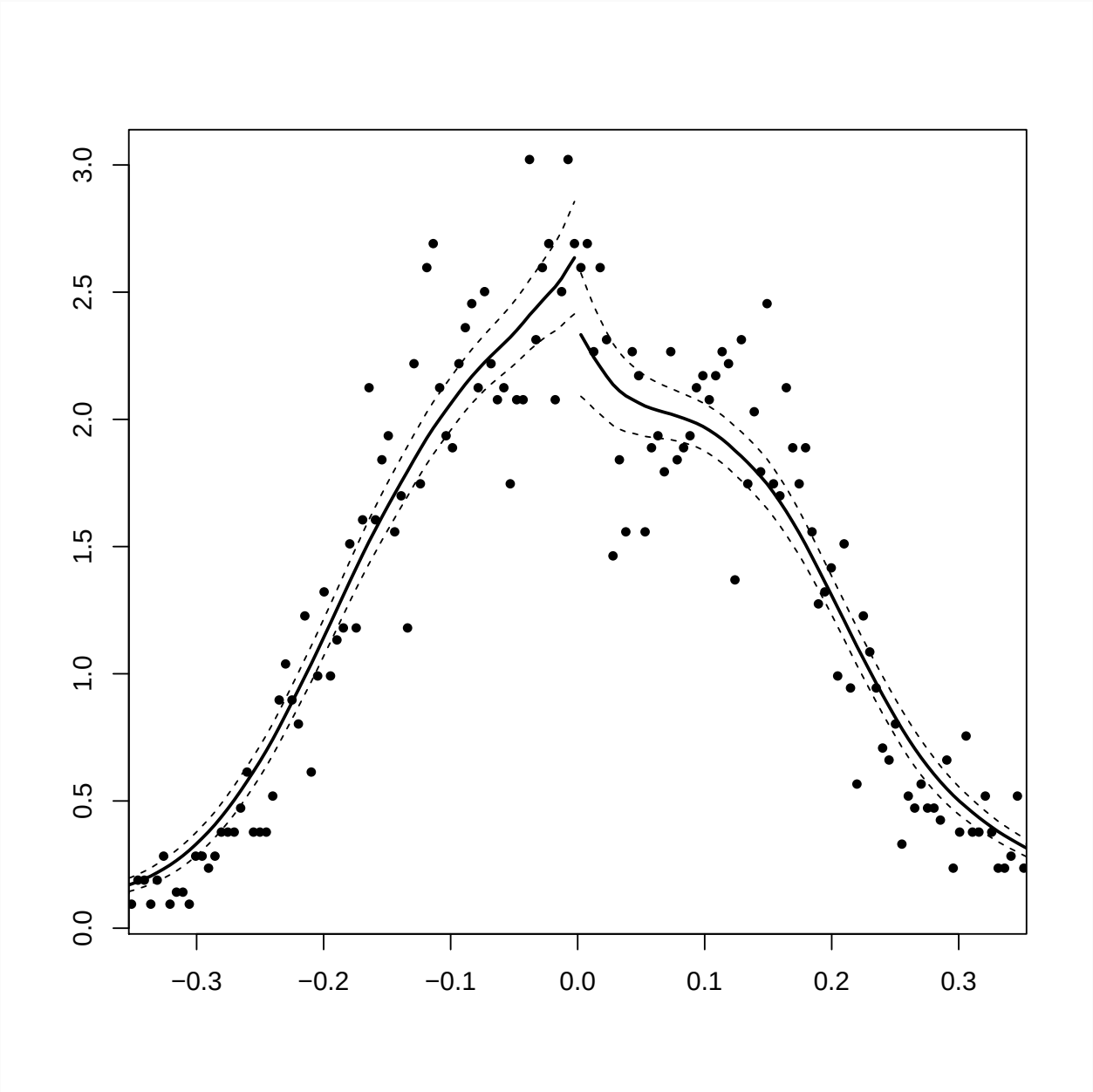
- In principle not testable
- Yet, we can get some **evidence** that could speak to it
- The assumption tells us that manipulation/sorting into T and C is not possible: evaluate by density plot.
- If there is no increased density around the cut-off, we probably do not observe manipulation.

In R:

```
DCdensity(data$votemargin,0, verbose=T, plot=T)
```

The command comes with a test of the null-hypothesis that there is no difference in the densities before and after the discontinuity. We do not reject the null ($p - value = 0.17$) Unfortunately I could not find any equivalent command in Stata

Question 3 (d) – iii – Test assumption 1?



Question 4

Question 4

Your research project:

Question 4

Your research project: Always feel free to ask me specific questions if you have them, or pass by my office during academic support hour.

Conclusion

All clear? More questions?

Thanks and see you next week!