Exam Questions

STV4020B

Statistical Analysis, Fall 2021

Department of Political Science, University of Oslo

Candidate: 123

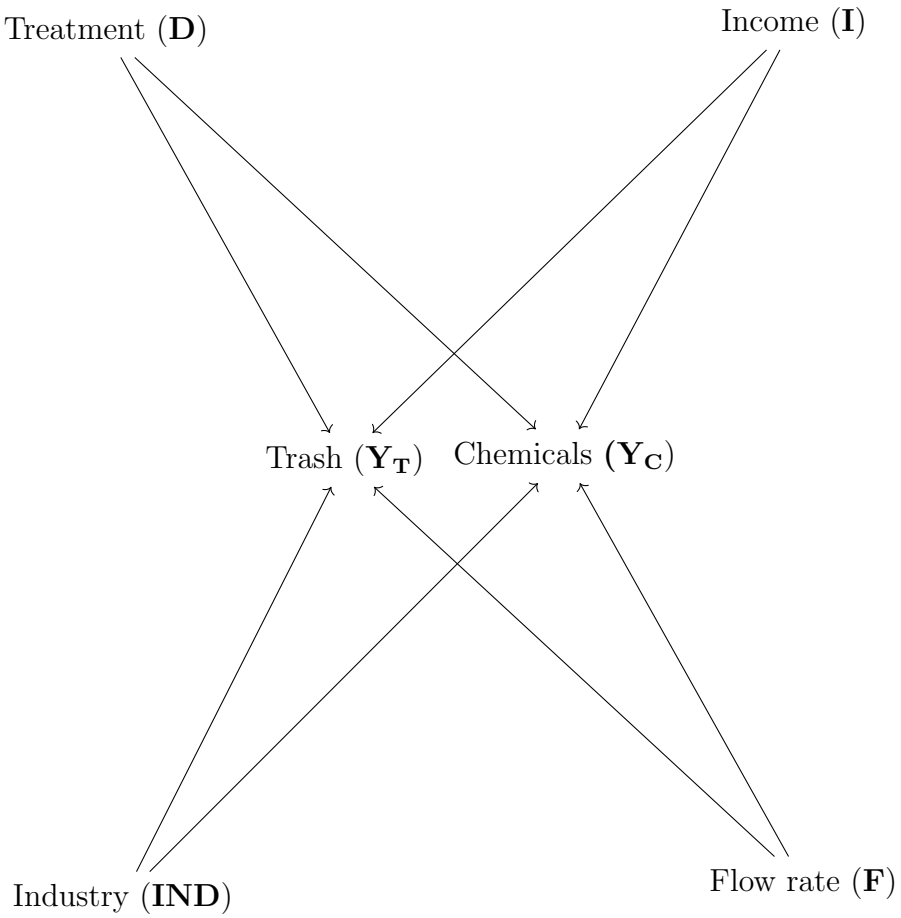
Word count: 4124

Contents

1 Week 1

1.1 Question 1

Answer to a)



$\mathrm{D} \longrightarrow Y_{\mathrm{T}}$

$\mathrm{D} \longrightarrow Y_{\mathrm{C}}$

Answer to b) There are not any backdoor paths, however the river's *flow rate*, *local income* as well as *local industry* all potentially cast spurious effects on our dependent variables, and should be controlled for to produce less biased estimates.

Answer to c i & ii) Our treatment is estimated to have a negative effect on water chemical levels, with the Average Treatment Effect (ATE) being -13.727, being significant on a p-value lower than 0.05. The Standard Error (SE) of the estimate is 6.190.

Answer to c ii) The SE of the treatment is slightly smaller when including the covariates for chemicals (from 6.190 to 6.010) and trash (from 1.278 to 1.256). This SE is slightly smaller than the SE of our previous bivariate model, suggesting that our estimates are slightly less uncertain. This happens as we are able to control for the effect of the covariates. We are taking them out of the error term and including them in the equation.

Table 1:

	Dependent variable:			
	chemicals	trash	chemicals	trash
	(1)	(2)	(3)	(4)
treatment	−13.727**	−8.468***	−12.087**	−8.370***
	(6.190)	(1.278)	(6.010)	(1.256)
industry			1.063***	0.039
			(0.143)	(0.030)
local_income			−0.052*	−0.036***
			(0.030)	(0.006)
flow_rate			−0.368***	−0.059**
			(0.121)	(0.025)
Constant	410.925***	100.174***	471.512***	139.570***
	(4.377)	(0.903)	(31.201)	(6.520)
Observations	1,000	1,000	1,000	1,000
R^2	0.005	0.042	0.067	0.080
Adjusted R^2	0.004	0.041	0.064	0.076
Residual Std. Error	97.874 (df = 998)	20.203 (df = 998)	94.893 (df = 995)	19.828 (df = 995)
F Statistic	4.917** (df = 1; 998)	43.923*** (df = 1; 998)	17.984*** (df = 4; 995)	21.666*** (df = 4; 995)

*p<0.1; **p<0.05; ***p<0.01

Answer to c iii) Adding trash on the right hand side heightens the standard error and lowers the estimated effect of the treatment (D) on chemical water pollution (Y_C). Its inclusion is not theoretically justified and introduces weak multicollinearity as trash (Y_T) is a post-treatment estimate being affected by the treatment (D). A cor.test in R shows that the correlation between treatment (D) and trash (Y_T) is -0.2053185.

Answer to d i) The causal inference assumption that is violated is the *Stable Unit Treatment Value Assumption* (SUTVA), implying that the treatment is received in homogeneous doses to all units, and that there are no externalities (spillovers to other treatment units). In this experiment, it means that the units (here, the different waterways in different communities) are no longer independent from each-other as the benefit of a treatment received on one waterway may reduces pollution in other waterways (both in the treatment group and the untreated group). Cunningham (2021) exemplifies this by illustrating that "[when] unit 1 receives the treatment, and there is some externality, then unit 2 will have a different Y_0 value than if unit 1 had not received the treatment" (Cunningham, 2021, p. 164). Therefore, when we encounter spillover, we have to use models that can account for STUVA violations, such as Goldsmith-Pinkham and Imbens (2013) (Cunningham, 2021, p. 164).

Answer to d ii) Yes, given that spillover violates SUTVA, the estimated effect of the treatment will be biased. It can be expected to be larger than the true effect as the benefit of the treatment on one waterway casts a potential benefit on other waterways. However, the extent of the spillover effect depends on the leverage of the spillover (how many units the spillover affects). However, it may be smaller if the benefit of one waterway affetcs

Answer to e) Yes, the treatment effect will be biased as the value of chemical and trash pollution in the untreated units will be superficially high, likely heightening our estimated treatment effect.

Answer to f i) Yes, as high-income neighborhoods are able to influence the randomization process and to have themselves put into the treatment group, it creates considerable *selection bias* which will be captured by the Simple Differences in Outcomes (SDO) value (SDO = ATE + Selection bias + Heterogeneous treatment effect bias), likely overestimating the effect of the ATE as the high-income communities already have more resources to implement community monitoring as well as less pollution in their waterways.

Answer to f ii) Yes, heterogeneous treatment effect bias will be present if we use SDO to estimate ATE, as the issued treatment will deliver what Cunningham (2021) calls "*different returns*" to the treatment on the different neighbourhoods multiplied by the share of the untreated population. Therefore, the SDO will overestimate the effect.

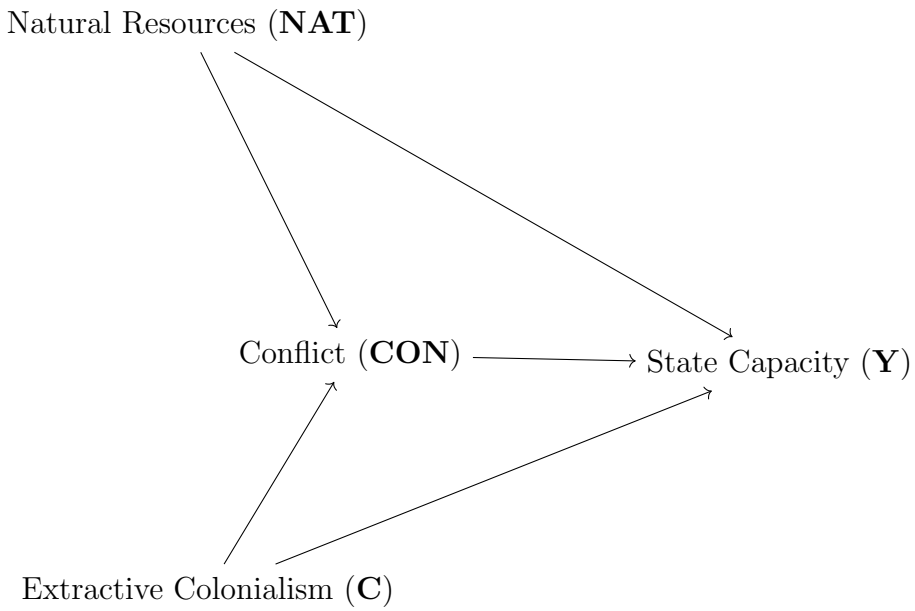
1.2 Question 2

The estimate is most likely biased. Several problems may occur: (1) The readers of the newspaper may be contingent on certain characteristics of the readers, creating selection bias and they (2) they have to have access to the internet/a computer (which may be contingent on income), (3) nation/state's school funding system (whether it is funded by taxpayers in smaller sub-state political units or whether it follows nation/state-wide standardization), (4) they have to have school-aged children, (5) and the survey design may create further bias as people who have school-aged children have to spend more time answering the survey, which could prove to be a disincentive from completing it.

The question states the the newspaper is a "major newspaper", meaning that it could have nation/state-wide coverage, resulting a large sample (closing in on the central limit theorem), which would still produce a biased, yet cursory estimate.

1.3 Question 3

Answer to a)



$\mathbf{C} \longrightarrow \mathbf{Y}$

$\mathbf{C} \longrightarrow \mathbf{CON} \longrightarrow \mathbf{Y}$

$\mathbf{C} \longrightarrow \mathbf{CON} \longleftarrow \mathbf{NAT} \longrightarrow \mathbf{Y}$

The variables *colonialism* (C) and *natural resource* (NAT) together create the collider *conflict* (CON).

Answer to b) The estimated effect of colonialism (C) on state capacity (Y) is - 3.828, significant with a p-value lower than 0.01***. See (1) in table 2.

Answer to c) In the countries with a conflict (CON) value higher than 1SD from the mean (N = 336), the estimated effect of colonialism (C) on state capacity (Y) is 0.286 (see mod 2, table 2). The effect is meaningfully different form the model (1) in task b as the effect is insignificant and we are

introducing selection bias by only keeping the countries with a high conflict value in our data set.

Answer to d) By regressing state capacity (Y) on conflict (CON) and colonialism (C) using the standard data set (N = 2000), the estimated effect of colonalism 1.958, significant with a p-value lower than 0.01***. See (3) in table 2.

Answer to e) When regressing state capacity (Y) on colonialism (CON), conflict (C), and natural resources (NAT) using the full data set, the estimated effect of colonialism is -1.086, significant with a p-value lower than 0.01***. See (4) in table 2.

Answer to f) If we assume that our theory is correct about the determinants of state capacity, model (4) gives the real causal effect as we are able to condition on both natural resources (NAT) and conflict (C). Model (3) gives us a biased estimate as we are conditioning on the collider, and model (2) introduces selection bias. The second best option would be to use model (1) where we are not conditioning on the collider, but biasing our estimates to the spurious effects of the covariates.

Table 2:

	Dependent variable:			
	statecapacity			
	(1)	(2)	(3)	(4)
conflict			−5.951***	−2.909***
			(0.024)	(0.216)
colonialism	−3.828***	0.286	1.958***	−1.086***
	(0.135)	(0.260)	(0.033)	(0.217)
resources				−3.083***
				(0.217)
Constant	−0.075	−10.652***	0.040*	0.039*
	(0.133)	(0.332)	(0.023)	(0.022)
Observations	2,000	336	2,000	2,000
R^2	0.288	0.004	0.978	0.980
Adjusted R^2	0.288	0.001	0.978	0.980
Residual Std. Error	5.959 (df = 1998)	3.610 (df = 334)	1.045 (df = 1997)	0.996 (df = 1996)
F Statistic	809.245*** (df = 1; 1998)	1.212 (df = 1; 334)	44,658.000*** (df = 2; 1997)	32,821.480*** (df = 3; 1996)

*p<0.1; **p<0.05; ***p<0.01

1.4 Script

```
### Exam Question 1 ci, A & B)

library(readr)
X4020b_2021_week1_Q1_waterways_sim_data <- read_csv("4020
   b_2021_week1_Q1_waterways_sim-data.csv")
View(X4020b_2021_week1_Q1_waterways_sim_data)

Mod1 <- lm(chemicals ~ treatment, data =
   X4020b_2021_week1_Q1_waterways_sim_data)
Mod2 <- lm(trash ~ treatment, data =
   X4020b_2021_week1_Q1_waterways_sim_data)
stargazer(Mod1, Mod2, type = "text")
stargazer(Mod1, Mod2)
# Our treatment is estimated to have a negative effect on water
    chemical levels,
# with the Average Treatment Effect (ATE) being -13.727 for
   chemicals, being significant on a **p<0.05 and
# a Standard Error(SE) of 6.190. The ATE for trash is -8.468,
   being significant on
# ***p<0.01 with a SE of 1.278.

### Exam Question 1 cii, A & B)

Mod3 <- lm(chemicals ~ treatment + industry + local_income +
   flow_rate, data = X4020b_2021_week1_Q1_waterways_sim_data)
Mod4 <- lm(trash ~ treatment + industry + local_income +
   flow_rate, data = X4020b_2021_week1_Q1_waterways_sim_data)
stargazer(Mod3, Mod4, type = "text")
# Our treatment is estimated to have a negative effect on water
    chemical levels,
# with the Average Treatment Effect (ATE) being -13.727, being
   significant on a p<0.05.
```

```
# The Standard Error(SE) of this estimate = 6.190.


### Exam Question 1 cii C)

stargazer(Mod1, Mod2, Mod3, Mod4, type = "text")
stargazer(Mod1, Mod2, Mod3, Mod4)
# Both ATE are slightly smaller when including the covariates,
# for chemicals (-13.727 -> )
# The SE of treatment is slightly smaller when including the
   covariates
# for chemicals (from 6.190 to 6.010) and trash (from 1.278 to
   1.256).


# This SE is slightly smaller than the SE of our previous
   bivariate model, suggesting
# that our estimates are slightly less uncertain. This happens
   as we are able to control for the effect
# of the covariates. We are taking them out of the error term
   and including them in the equation.


### Exam Question 1 c iii)

Mod6 <- lm(chemicals ~ treatment + trash, data =
   X4020b_2021_week1_Q1_waterways_sim_data)
stargazer(Mod6, type = "text")
Mod7 <- lm(chemicals ~ treatment, data =
   X4020b_2021_week1_Q1_waterways_sim_data)
stargazer(Mod7, type = "text")

stargazer(Mod6, Mod7, type = "text") #shows how our SE are
   effected by including trash on the right-hand side of the
```

```
   model.
# SE increases, which is worse meaning that now our estimates
   are more uncertain.

cor.test(X4020b_2021_week1_Q1_waterways_sim_data$treatment,
   X4020b_2021_week1_Q1_waterways_sim_data$trash)
# A simple cor.test shows that treatment affects trash
   (-0.2053185), introducing weak multicollinearity.

library(ggplot2)

Mod8 <- lm(chemicals ~ trash, data = X4020b_202)

stargazer(Mod7, Mo8, )
# Yes, doing so would introduce various biases. First, the
   treatment and the trash variable are related to each other
# which introduces multicolinearity which inflates the SE.
   Second, Mod8 shows that the variables chemcials and
# trash are related to each other (**p<0.05), meaning that it
   introduces more bias through endogeneity.

### Exam Question 3 a

install.packages("dagitty")
library(dagitty)

DAG_Q3 <- dagitty(dag{bb="0,0,1,1"
  "STATE CAPACITY" [outcome,pos="0.450,0.438"]
  COL [pos="0.182,0.555"]
  CON [pos="0.284,0.439"]
  NAT [pos="0.185,0.318"]
  COL -> "STATE CAPACITY"
  COL -> CON
  CON -> "STATE CAPACITY"
```

```
  NAT -> "STATE CAPACITY"
  NAT -> CON})
plot(DAG_Q3)


### Exam Question 3 b

X4020b_2021_week1_Q3_country_conflict_sim_data <- read_csv
   ("4020b_2021_week1_Q3_country_conflict_sim-data.csv")
View(X4020b_2021_week1_Q3_country_conflict_sim_data)

Mod31 <- lm(statecapacity ~ colonialism, data =
   X4020b_2021_week1_Q3_country_conflict_sim_data)
stargazer(Mod31, type = "text")
# The "effect" of colonialism on state capacity is -3.823,
   significant on ***p<0.01.

### Exam Question 3 c

conflict_data <- X4020b_2021_week1_Q3_country_conflict_sim_data
summary(conflict_data$conflict)

conflict_data2 <- conflict_data %>%
  filter(conflict > sd(conflict))
Mod32 <- lm(statecapacity ~ colonialism, data = conflict_data2)
stargazer(Mod31, Mod32, type = "text")
stargazer(Mod31, Mod32)
# This table shows that where conflict is SD >1, the effect of
   colonialism is
# 0.286 and not significant, and R2 very small. The effect is
   meaningfully different as we are introducing selection
# bias by keeping countries with conflict.

### Exam Question 3 d
```

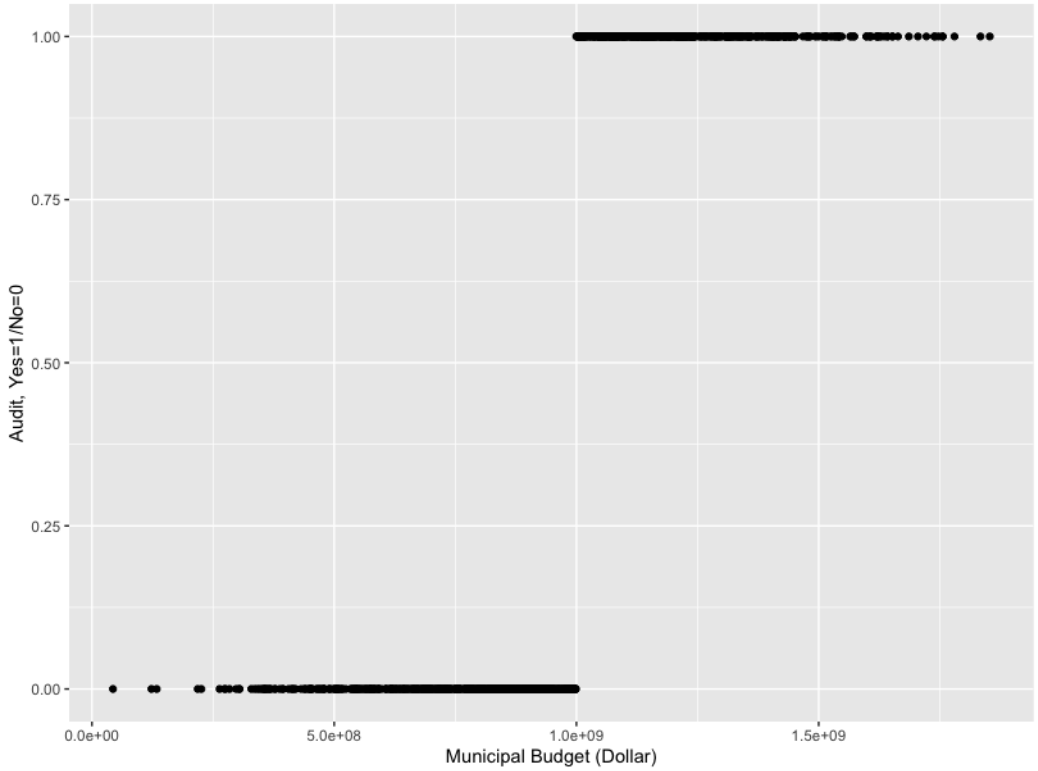
```
Mod33 <- lm(statecapacity ~ conflict + colonialism, data =
   conflict_data)
stargazer(Mod33, type = "text")
# In this regression table the "effect" of colonialism on
   statecapacity is positive (1.958)
# and significant on ***p<0.01. This is misleading given
   collider bias. Collider

### Exam Question 3 e

Mod34 <- lm(statecapacity ~ conflict + colonialism + resources,
    data = conflict_data)
stargazer(Mod34, type = "text")
# In this regression where we include resources table the "
   effect" of colonialism on statecapacity is negative
   (-1.086)
# and significant on ***p<0.01. This is also misleading given
   collider bias.

### Exam Question 3 f)
stargazer(Mod31, Mod32, Mod33, Mod34, type = "text")

stargazer(Mod31, Mod32, Mod33, Mod34)
```

2 Week 2

2.1 Question 1

Answer to 1)

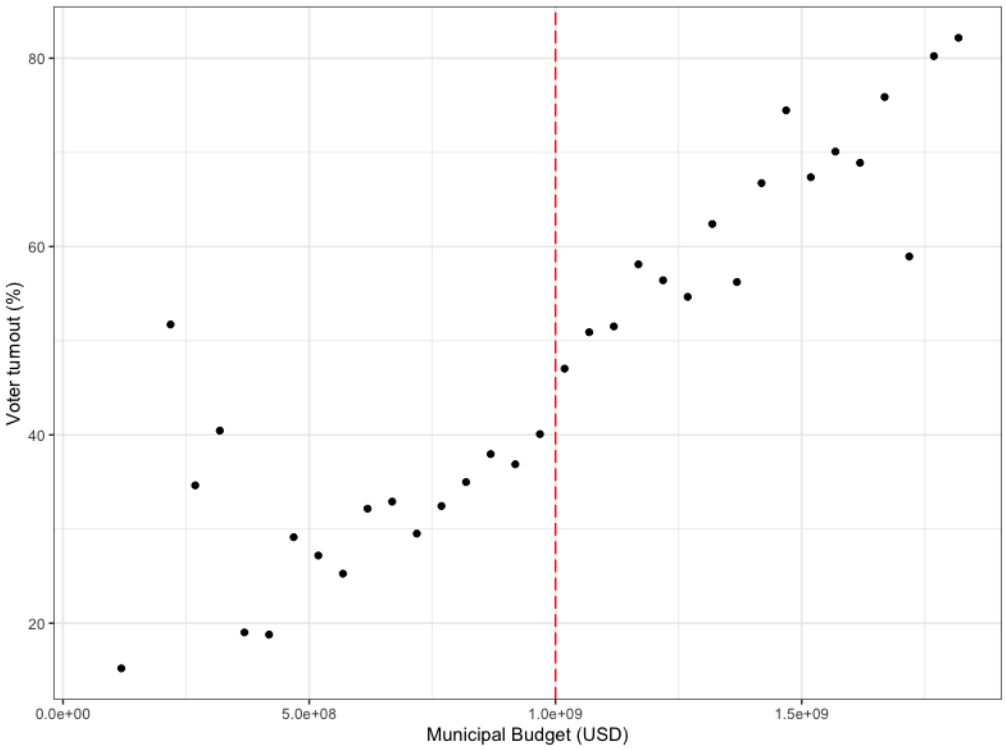
Figure 1:



The binary treatment (audit or no audit), is suitable for a sharp Regression Discontinuity Design (RDD). However, it may be hard to separate the treatment effect from endogenous variables that may influence voter turnout surrounding the $ 1 billion cutoff.

Answer to 2)

Figure 2:



Eyeballing the plot suggests that there may be a slight discontinuity around the cutoff, from 40 to 47% voter turnout.

Answer to 3)

Figure 3:

```
Number of Obs.                  517          483
Eff. Number of Obs.             263          247
Order est. (p)                    1            1
Order bias  (q)                   2            2
BW est. (h)             207429709.253  207429709.253
BW bias (b)             330591844.930  330591844.930
rho (h/b)                       0.627        0.627
Unique Obs.                       517          483

=============================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
=============================================================================
  Conventional     6.723     3.101     2.168     0.030     [0.644 , 12.802]
        Robust         -         -     1.718     0.086    [-0.879 , 13.355]
=============================================================================
```

The estimated LATE coefficient is 6.732. The estimated standard error (SE) is 3.301. The p-value = 0.030, meaning that the estimate is significant on the 5% level. Therefore, the LATE here is substantially important, suggesting a 6.723% increase in voter turnout from the audit.

Answer to 4 a) In anticipation of the audit, corrupt officials just above the $ 1 billion cutoff are likely to be adjusting their budgets down to avoid being audited.

Answer to 4 b) This violates the *continuity assumption* as the units would anticipate the treatment in advance, and some are interested in adjusting their budgets (the running variable) as well as having the time to do so. Hence, there is no longer "as-if-random" treatment assignment.

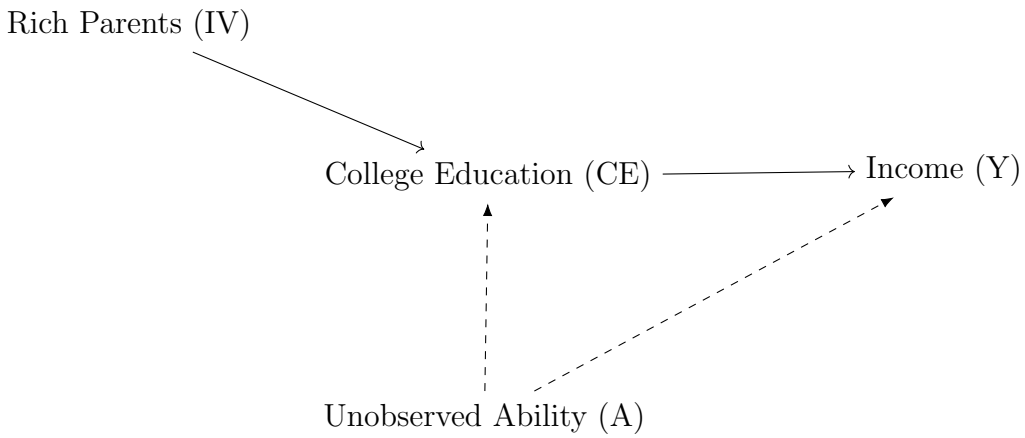
Answer to 4 c) The observed LATE would be smaller if the corrupt municipalities had the oppurtunity to adjust their budgets in anticipation of the audit, and larger if they did not. This is when we assume auditing corrupt municipal governments would inflate voter turnout.

Answer to 4 d) The *McCrary density test* (2008) allows us to check

whether units are sorting on the running variable (in this example, adjusting the municipal budget).

2.2 Question 2

Answer to 1)



Answer to 2) The estimated effect of college education (CE) on income (Y) is 26,445.100, significant with a p-value lower than 0.01. The estimated standard error (SE) is 1,028.488. See (1) in table 3. If the researcher's causal model is correct, then the above OLS estimate of the causal effect is biased due to omitted variable bias.

Answer to 3) Using the IV specification, the estimated effect of college education (CE) on income (Y) is 44,705.730, significant with a p-value lower than 0.01. The estimated SE is 3,971.551. See (2) in table 3.

Table 3:

	Dependent variable:	
	Income	
	OLS	*instrumental variable*
	(1)	(2)
College	26,445.100***	44,705.730***
	(1,028.488)	(3,971.551)
Constant	65,916.100***	58,118.810***
	(672.068)	(1,793.453)
Observations	1,000	1,000
R^2	0.398	0.208
Adjusted R^2	0.398	0.208
Residual Std. Error (df = 998)	16,087.570	18,454.240
F Statistic	661.138*** (df = 1; 998)	
Note:		*p<0.1; **p<0.05; ***p<0.01

Answer to 4) Having rich parents may be correlated with income through various unobserved variables such as geography (what neighbourhood/country/state you grew up in) or through the rich parents' contact network which could make it easier for the individuals in question to get higher paying jobs. If the exclusion restriction is violated, then the instrumental variable (IV) design no longer holds internal validity as the existence of the IV is supposed to identify or isolate the direct effect of the treatment (here CE) on income (Y) *independent* of other potential unobserved variables.

2.3 Script

```
install.packages("tidyverse")
library(tidyverse)
library(dplyr)

### 1 RDD



## Question 1

#Loading data set
exam_part_2_RDD <- readRDS("~/OneDrive - PRIO/Exam W2 STV4020B/
   exam_part_2_RDD.RDS")
govt_audit_data <- exam_part_2_RDD #renaming

#Making a plot with "b" = x and "audit" = y

library(ggplot2)
ggplot(govt_audit_data2, #making plot using ggplot2
       aes(y = audit,
           x = b)) +
  geom_point() +
  labs(title = "Municipal budget and Audit") +
```

```
  labs(x="Municipal Budget (Dollar)", y="Audit, Yes=1/No=0")
  theme_bw()

#Given the cutoff it initially seems like a good starting point
    for a harp RDD design,
# however it seems hard to separate confounding treatments at
   the 1B$ cutoff.

## Question 2

br=seq(from=min(govt_audit_data$b),to=max(govt_audit_data$b),by
   =5E7) #code for setting binwidt
midpoints=(br[-1]+br[-length(br)])/2 #code for setting
   midpoints
govt_audit_data$b=cut(govt_audit_data$b,breaks=br) # applying
   bins

B=data.frame(cbind(midpoints,by(govt_audit_data$voting,
   govt_audit_data$b,mean))) # new data frame, attached binned
    b by voting and running variable b.
names(B)=c('midpoints','audit') #renaming the axis

##making a plot
ggplot(B, aes(midpoints, audit)) +
  geom_point() +
  labs(title = "Municipal budget and Audit") +
  geom_vline(xintercept = 1E09, colour = "red", linetype = 5) +
  labs(x="Municipal Budget (Dollar)", y="Voter turnout (%)") +
  theme_bw()

#Eyeballing suggests that there may be a slight jump around the
    cutoff, from 40 to 46~8% voter turnout.

## Question 3
```

```
govt_audit_data2 <- readRDS("~/OneDrive - PRIO/Exam W2 STV4020B
   /exam_part_2_RDD.RDS")
govt_audit_data2$billion = govt_audit_data2$b - 1e9

install.packages("rdrobust")
library(rdrobust)

govt_audit_data2$billion = govt_audit_data2$b - 1E09 #adding
   new variable


Model = rdrobust(govt_audit_data2$voting,
   govt_audit_data2$billion)
summary(Model)
# The estimated LATE coefficient is 6.732.
# SE is 3.301. The p-value = 0.030,
# estimate is significant on the 5% level

### 2 IV

## Question 2

#load data and rename
coledu_income_data <- readRDS("~/OneDrive - PRIO/Exam W2
   STV4020B/4020b_2021_week2_Q2_education_sim-data.RDS")

#run bivariate regression
library(stargazer)
Mod1 <- lm(Income ~ College, data = coledu_income_data)
stargazer(Mod1, type = "text")
# Coef. = 26,445.100, significant on ***p<0.01.
# SE = 1,028.488
```

```
## Question 3
install.packages("ivreg")
library(ivreg)

#run IV regression
IV1 = ivreg(
  formula = Income ~ College | RichParent, data =
     coledu_income_data)
stargazer(IV1, type = "text")
summary(IV1)

#comparing
stargazer(Mod1, IV1, type = "text")
stargazer(Mod1, IV1)
# Coef. = 44,705.730, signficant on ***p<0.01
# SE = 3,971.551
```

3 Week 3

3.1 Question 1

Answer to 1)

Table 4:

	Dependent variable:	
	NA	
	(TWFE 1)	(TWFE 2)
regionDummy	-2.783^{***}	-2.783^{***}
	(0.120)	(0.124)
Time	0.636^{***}	0.951^{***}
	(0.120)	(0.124)
regionDummy:Time	3.738^{***}	-0.061
	(0.170)	(0.175)
Constant	11.000^{***}	11.000^{***}
	(0.085)	(0.087)
Observations	2,000	2,000
R^2	0.424	0.356
Adjusted R^2	0.423	0.355
Residual Std. Error (df = 1996)	1.899	1.954
F Statistic (df = 3; 1996)	490.067^{***}	367.740^{***}

*p<0.1; **p<0.05; ***p<0.01

The estimated treatment effect is 3.738, significant with a p-value lower than 0.01. The estimated SE is 0.236. Hence, in this specification (TWFE 1), the estimated effect of subsidies suggests an average *increase* of 3.738 million euro R&D spending in region A's firms. Given that the estimated effect of Time on R&D spending is only 0.638 in specification (1), the estimated treatment effect is substantively important.

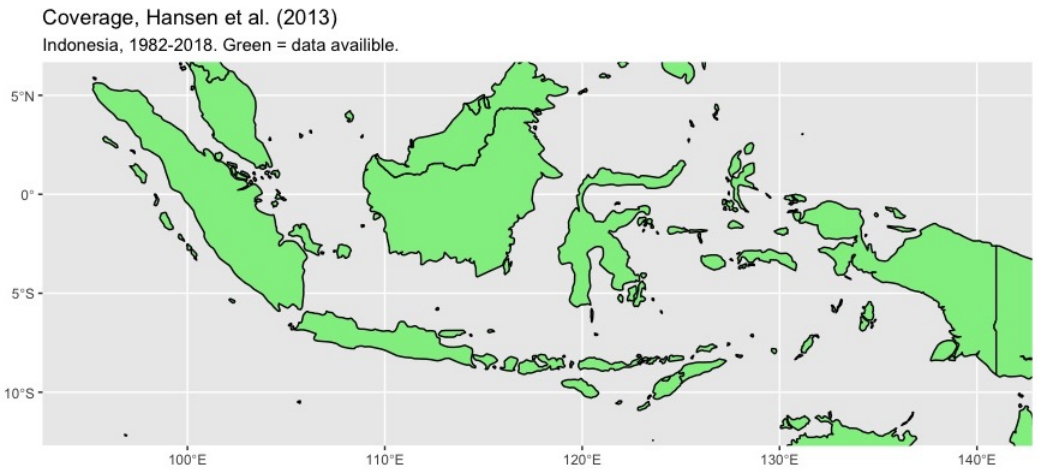
Answer to 2) The estimated treatment effect -0.061 (the DD coefficient), however it is not significant. The SE is (0.175). The results in this specification (TWFE 2) support the parallel trends assumption, as the DD coefficient in the the estimated pre-treatment period (-1 to 0) are statistically insignificant.

Answer to 3) The key assumption that allows me to separate the effect of the time from the effect of the treatment is the *parallel trends assumption*, which holds that the *effect of Time* (i.e. the natural growth of R&D spending) is fixed for all units (both A ans B). By definition this assumption is untestable, as we cannot observe the counterfactual conditional expected value of Time. My estimates of pre-treatment trends allows me to rebut some of the critics, as it shows that the parallel trends assumption holds for pre-treatment values (as checked in model TWFE 2), however there is always some possibility that the treatment is not strictly exogenous and that it is correlated with the structural error term.

Answer to 4) Yes, this would violate parallel trends as the the value for potential outcomes would be more more negative for region A than for region B. This would bias the estimated treatment effect downwards, as the relative decline in R&D spending in A would underestimate the effect of the treatment on R&D spending. However, the substantive conclusion would hold, but just downplay the effect.

3.2 Question 2

Answer to 1 to 6)



Does local government imposed measures to curb illegal logging affect tropical forest cover loss in North Sumatra, Indonesia? In my hypothetical research design I will employ a differences in differences (DiD) research design to address the effect of provincial measures against illegal logging implemented in 2013 (my treatment/independent variable) on levels of deforestation (my dependent variable).

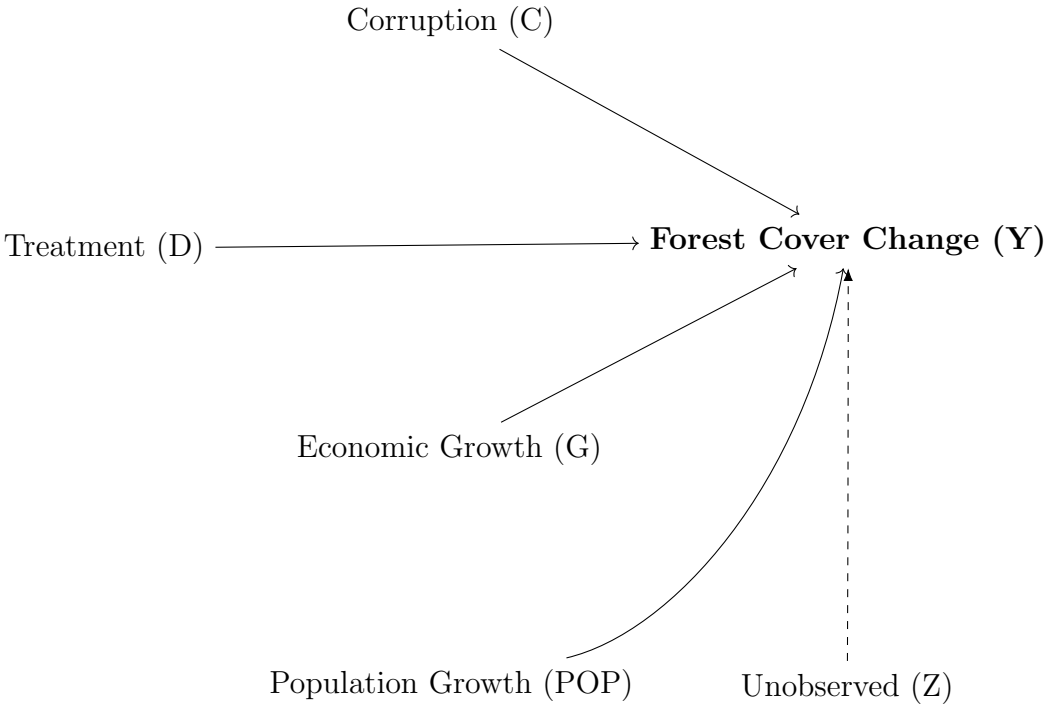
For data constituting my dependent variable (Y) I would use Landsat satellite imagery on forest cover change (in km^2) data from the Hansen et al. (2013) data set, offering continuous spatial, year-by-year coverage from 1982-2018 (see the above figure).

The data constituting my treatment variable (D) of interest would cover monitoring measures to limit illegal logging in North Sumatra in 2013. I will assume that North Sumatra was randomly selected as a result of the central government's efforts to trial new measures to curb illegal logging, and that the measures were implemented across all of North Sumarta's municipalities (*kabupaten*). Hence, North Sumatra is my treatment group k.[1]

Here, a simple 2 x 2 DiD design would prove suitable way to measure the effect of monitoring as I could compare the treated province to a neigh-

[1] This is a fictional treatment used for this assignment, and I assume that I have data covering implementation.

bouring province - West Sumatra - that has not implemented measures to limit illegal logging. West Sumatra will therefore serve as the control group U, as West Sumatra is located in a geographically and economically similar area of Indonesia, and experiences similar levels of deforestation in the years before 2013 - having similar pre-treatment trends on the dependent variable. Theoretically, I expect to find that the treatment (D) has a negative effect on forest cover change (Y) in treatment group k, i.e. causing less deforestation. A 2 x 2 DiD will isolate the average treatment effect on the treated (ATT) given that the *parallel trends assumption* holds.[2]



However, as seen in the above DAG, there may be several omitted variables affecting the dependent variable that need to be controlled for, such as levels of corruption, economic growth and higher population as these are all hypothesized variables that increase illegal logging.[3] For the parallel trends

[2] Also referred to as the DiD estimator.

[3] Data on population and economic growth will be sourced from the Indonesian Bureau of Statistics (*Badan Pusat Statistik*), https://www.bps.go.id/. For data on corruption I will use data on provincial corruption in Indonesia from Suhardjanto et al. (2018).

assumptions to be supported,[4] the two provinces need to have a similar effect of time on its dependent variable (forest cover change), as well as the control variables. To do this, one would conduct a check of the pre-treatment trends by comparing trend lines as well as running two-way fixed effects models in pre-treatment years to check whether the DiD estimator is significant.

There may be additional threats to causal inference, it is possible that some of the illegal loggers may have discovered the extra monitoring during the year of implementation (2013) in North Sumatra, and therefore shifted their logging efforts to the neighbouring province, causing a *spillover*. This is a possibility that needs to be taken into account as it can cause biased treatment effects.[5]

However, if the parallel trends assumption is supported and the ATT is valid, then this hypothetical study could yield weak external validity for purposed measures against illegal logging measures in other middle-income democracies suffering from tropical deforestation caused by illegal logging (for example the Philippines, Nigeria, Colombia or Brazil).

[4] Crucially, the parallel trends assumption is by definition *untestable* as we cannot observe the counterfactual conditional expectation.

[5] Natural environment changes may also be an additional unobserved variable that could bias the treatment effect.

3.3 Script

```
#### Exam W3 STV4020B
library(tidyverse)

### Part 1 - Differences in Differences

### Question 1

#load file
stv4020b_week3_exam_p1_input <-
  readRDS("~/OneDrive - PRIO/Exam W3 STV4020B/
     stv4020b_week3_exam_p1_input.RDS")
#renaming
tax_subsidyRD <- stv4020b_week3_exam_p1_input

tax_subsidyRDmod <- tax_subsidyRD

#creating dummy
tax_subsidyRDmod <- mutate(tax_subsidyRDmod,
         regionDummy = ifelse(Region == "A", 1, 0))

#subsetting data using which()
tax_subsidyRDmod1 <- tax_subsidyRDmod[which(
   tax_subsidyRDmod$Time != -1),]

#estimation DiD two-way fixed effects

Mod1 <- lm('ResDev ~ regionDummy + Time + regionDummy*Time',
   data=tax_subsidyRDmod1)
library(stargazer)
stargazer(Mod1, type = "text")
# The estimated treatment effect is 3.738 with a SE of 0.170.
```

```
# Firm subsidy increases R&D spending with 3.738 million euros
   on average.


## Question 2

#subsetting data using which()
tax_subsidyRDmod2 <- tax_subsidyRDmod[which(
   tax_subsidyRDmod$Time < 1),]

#Second model, DiD two-way fixed effects
Mod2 <- lm('ResDev ~ regionDummy + Time + regionDummy*Time',
   data=tax_subsidyRDmod2)
stargazer(Mod1, Mod2, type = "text")
stargazer(Mod1, Mod2)


### Part 2

#Drawing Indonesia
install.packages("sf")
install.packages("rnaturalearth")
install.packages("rnaturalearthdata")
install.packages("rgeos")
install.packages("cowplot")
install.packages("googleway")
install.packages("ggrepel")
install.packages("libwegeom")
library(sf)
library(ggplot2)
library(rnaturalearth)
library(rnaturalearthdata)
library(rgeos)
library(cowplot)
```

```
library(googleway)
library(ggrepel)
library(libwegeom)

world <- ne_countries(scale = "medium", returnclass = "sf")
class(world)

ggplot(data = world) +
  geom_sf(color = "black", fill = "lightgreen") +
  coord_sf(expand = FALSE) +
  coord_sf(xlim = c(92.66, 142.80), ylim = c(6.67, -12.71),
     expand = FALSE) + #coordinates for Indonesia
  ggtitle("Coverage, Hansen et al. (2013)", subtitle = "
     Indonesia, 1982-2018. Green = data availible.")
  theme_bw()
```

4 Week 4

4.1 Question 1

Figure 4:

```r
## a)

#Pr(false positive test | omicron) = 0.1
pr_falsepos <- 0.1 # unconditional prob
#Pr(true positive test | omicron) = 0.9
pr_truepos <- 0.9 # unconditional prob

#Pr(false negative test | omicron) = 0.1
pr_falseneg <- 0.1 # unconditional prob
#Pr(true negative test | omicron) = 0.9
pr_trueneg <- 0.9 # unconditional prob
#Pr(positive in population | omicron ) = 0.01
pr_popomi <- 0.01 # unconditional prob

#Pr(positive | omicron) = 0.08333333
Pr_positive <- pr_truepos * pr_positiveomi + pr_falsepos * (1 - pr_popomi)
Pr_positive # 0.108

## b)

#Pr(omicron | positive) = 0.08333333
Pr_Truepos_Trueomi <- (pr_truepos * pr_popomi) / Pr_positive
Pr_Truepos_Trueomi # 0.08333333
```

Answer to a) The probability of testing positive given an omicron infection is 0.108 (i.e. 10,8%). For calculations, see the above script (Figure 4).

Answer to b) The probability of having omicron given a positive test is 0.08333333 (i.e. 8.3%). For calculations, see the above script (Figure 4).

4.2 Question 2

Answer to a) The priors I have chosen for the alpha and beta are both weakly informative priors, the idea being that the prior that has a regularising effect, ruling out unreasonable parameter values but is not being so strong as to rule out values that might make sense - allowing the the "data to dominate the inference" (McElreath 2020, p. 321).

For the intercept parameter alpha I assigned a normally distributed prior, with a mean of 50 (i.e the intercept is around the 50% Obama voters) with a standard deviation (SD) of 20 assuming that with 2SD you can cover 80% of the observed values. For the beta coefficient prior, I assumed that it will be normally distributed with an mean of 0, and a standard deviation of 1, as I expect there to be a slight positive relationship between the two variables (i.e. the higher Obama voter count is related to higher population density). Lastly, I assigned an uninformative uniform prior to the parameter sigma with a mean of 0 and a SD of 20, also covering 80% of the assumed values.

Answer to b) Running the ulam package to run a Hamiltonian Monte Carlo (HMC) simulation of four chains, R supplies the following summary (see figure 5): The mean of alpha the is 50.48 and its standard deviation (SD) is 1.25, with 89% of the posterior probability lying in between 48.52 and 52.56. The mean of beta is 0.0 and a SD of 0, with 89% of the posterior probability lying in between 0.01 and 0.03. Lastly, sigma has a mean of 8.55 and a SD of 0.87, with 89% of the posterior probability lying in between 7.26 and 10.04. The n_eff value is a crude estimate of the number of independent samples acquired from the chains on each parameter.

Figure 5:

```
> precis(Ulam1)
       mean   sd  5.5% 94.5% n_eff Rhat4
a     50.48 1.25 48.52 52.56  1449     1
b1     0.02 0.00  0.01  0.03  1626     1
sigma  8.55 0.87  7.26 10.04  1215     1
>
```

Answer to c) Ulam produces a convergence diagnostic (the R-hat value) comparing the between- and within-chain estimates of the model parameter. If the R-hat value is above 1.05, this indicates that the chains have not mixed well. As seen in Figure 4, the R-hat value is 1 for alpha, beta as well as sigma, indicating that the different chains have converged evenly. This is reflected in the traceplot and trankplot produced in R (see Figure 6 & 7), where the chains stabilize around the mean value.

Figure 6:

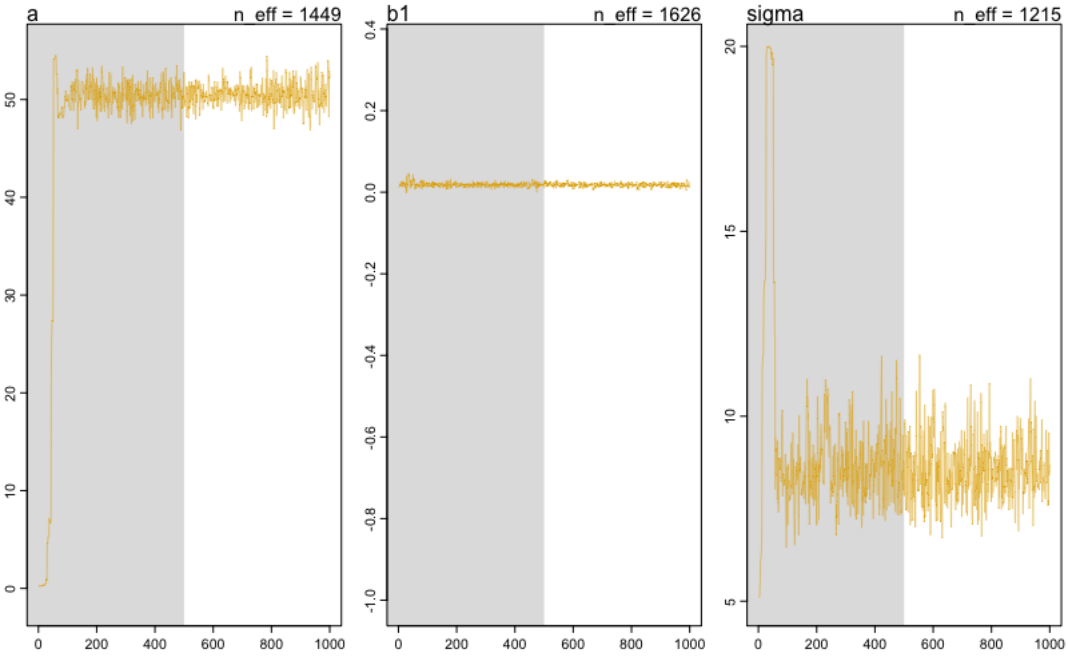
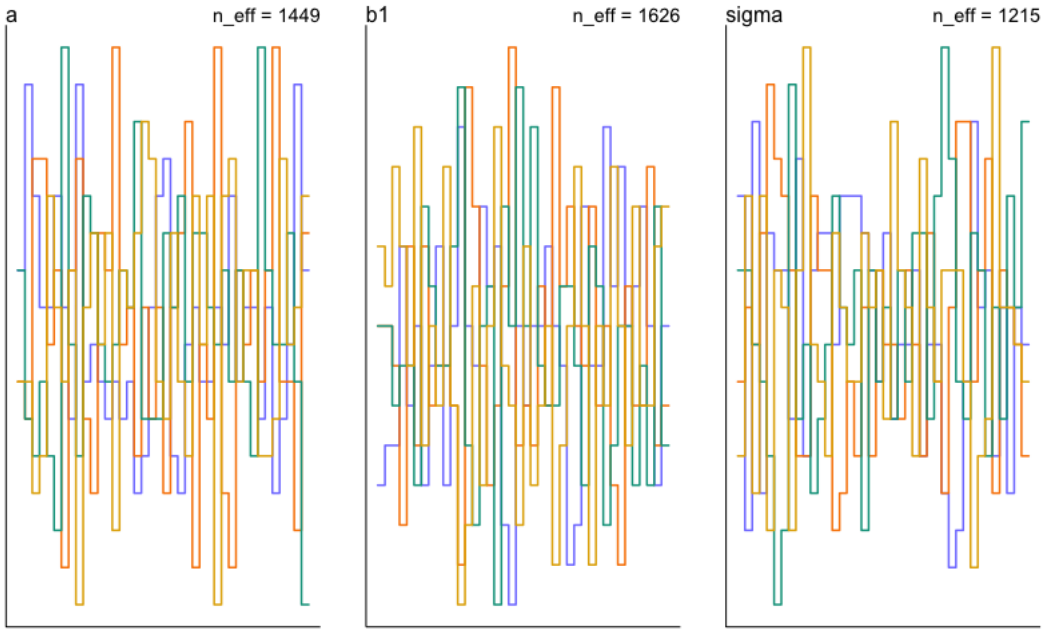
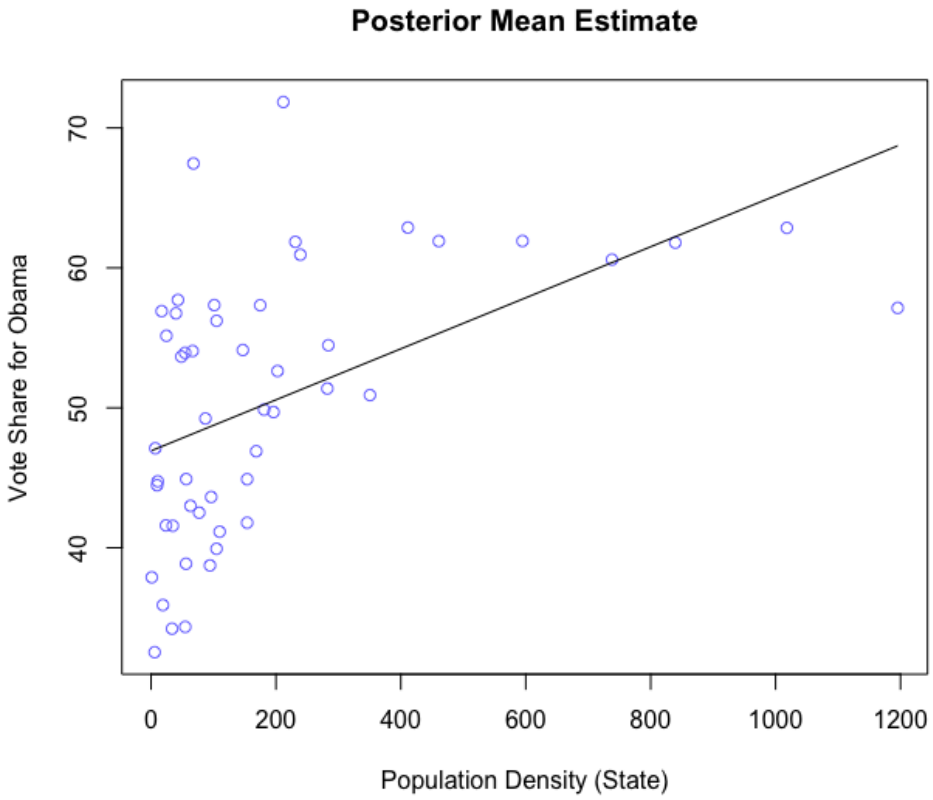


Figure 7:



Answer to d) The model seems reasonable in light of the plotted data, assuming that there is a positive relationship between population density and share of Obama voters in 2008 (Figure 8).

Figure 8:



Answer to e) I chose to logit transform the density variable. The effect of the logit transformation pulls out the ends of the distribution. Eyeballing Figure 10, the logged curve linear model suggests diminishing increases in share of Obama-voters following a relatively low state population density of 100∼200. The model with the logged density variable (Ulam1_log) is better, intuitively predicting share of Obama-voters better, which is also reflected in the values from the Widely Applicable Information Criterion (WAIC), with the logged model having a lower WAIC value, lower WAIC standard error and the highest Akaike weight of 0.96 (see Figure 11).

Figure 9:

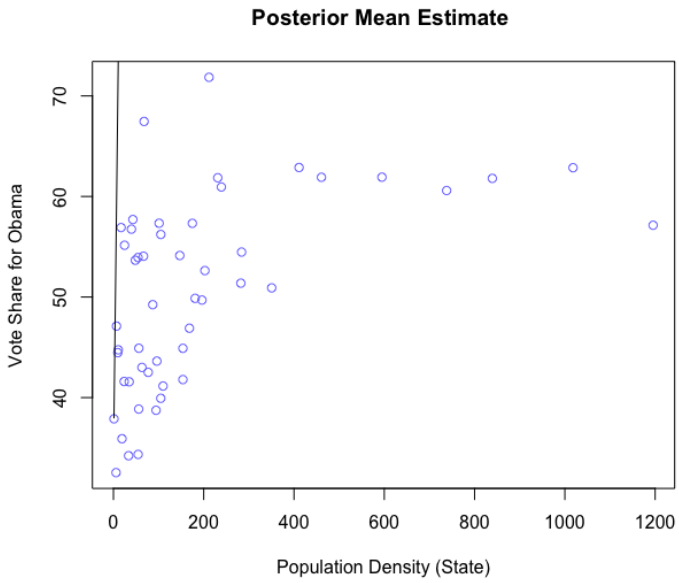


Figure 10:

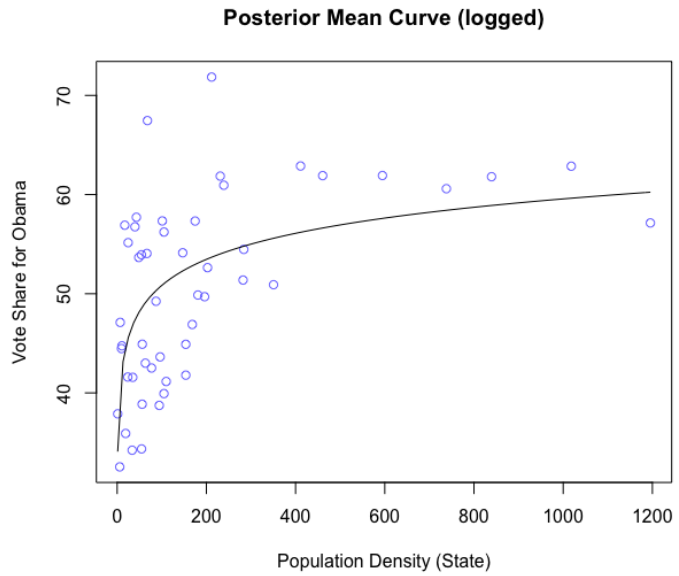


Figure 11:

```
> compare(Ulam1, Ulam1_log, func = WAIC)
             WAIC   SE dWAIC  dSE pWAIC weight
Ulam1_log 351.9 8.42   0.0   NA   2.2   0.96
Ulam1     358.2 9.19   6.3 4.86   2.7   0.04
```

4.3 Question 3

Answer to a)

Figure 12:

```
> precis(Ulam_inc, depth = 2)
        mean   sd  5.5% 94.5% n_eff Rhat4
a      50.42 1.04 48.70 52.10  2135     1
b_dens  2.47 0.83  1.16  3.85  1587     1
b1_inc  0.91 0.26  0.49  1.31  1784     1
sigma   7.33 0.79  6.18  8.68  1756     1
```

Figure 13:

```
> precis(Ulam_col, depth = 2)
        mean   sd  5.5% 94.5% n_eff Rhat4
a      50.48 1.00 48.87 52.03  2196     1
b_dens  2.70 0.77  1.45  3.91  1891     1
b2_col  0.95 0.24  0.56  1.34  2106     1
sigma   7.07 0.75  5.99  8.42  1362     1
```

Figure 14:

```
> precis(Ulam_col_inc, depth = 2)
        mean   sd  5.5% 94.5% n_eff Rhat4
a      50.45 1.04 48.81 52.09  1639  1.00
b_dens  2.52 0.82  1.21  3.81  1472  1.00
b1_inc  0.30 0.40 -0.34  0.94  1316  1.00
b2_col  0.72 0.38  0.15  1.32  1331  1.00
sigma   7.10 0.75  6.03  8.42  1350  1.01
```

The **first model** using income and population density as predictors of share of Obama voters (Ulam_inc, Figure 12), shows an alpha (intercept) mean of 50.42 and a SD of 1.04, with 89% of the posterior probability lies between 48.7 and 52.10. The beta (slope) for population density has a mean of 2.47 and a SD of 0.77, with 89% of the posterior probability lying between 1.15 and 3.85 - suggesting a positive relationship with share of Obama voters. The beta 1 (slope) for income has a mean of 0.91 and a SD of 0.26, with 89% of the posterior probability lying in between 0.49 and 1.31 - suggesting suggesting a weak positive relationship with share of Obama voters. Lastly, sigma has a mean of 7.33 and a SD of 0.79, with 89% of posterior probability

lying between 6.18 and 8.68 R-hat values of 1 suggests that all the chains have converged evenly.

The **second model** using college and population density as predictors of share of Obama voters (Ulam_col, Figure 13), shows an alpha (intercept) mean of 50.48 and a SD of 1.0, with 89% of the posterior probability lies between 48.87 and 52.03. The beta (slope) for population density has a mean of 2.52 and a SD of 0.77, with 89% of the posterior probability lies between 1.45 and 3.91 - suggesting a positive relationship with share of Obama voters. The beta 2 (slope) for college has a mean of 0.95 and a SD of 0.24, with 89% of the posterior probability lying in between 0.56 and 1.34 - suggesting suggesting a weak positive relationship with share of Obama voters. Lastly, sigma has a mean of 7.07 and a SD of 0.79, with 89% of posterior probability lying between 6.18 and 8.68 - suggesting a slightly narrower posterior than in the previous model. Again, R-hat values of 1 suggests that all the chains have converged evenly.

The **third model** using college, income and population density as predictors of share of Obama voters (Ulam_col_inc, Figure 14), shows an alpha (intercept) mean of 50.45 and a SD of 1.04, with 89% of the posterior probability lies between 48.81 and 52.09. The beta (slope) for population density has a mean of 2.52 and a SD of 0.82, with 89% of the posterior probability lying between 1.21 and 3.81 - still suggesting a positive relationship with share of Obama voters. The beta 1 (slope) for income has a mean of 0.30 and a SD of 0.40, with 89% of the posterior probability lying in between -0.34 and 0.94 - significantly different from the summary of the previous two models. The beta 2 (slope) for college has a mean of 0.72 and a SD of 0.38, with 89% of the posterior probability lying in between 0.15 and 1.32 - still suggesting suggesting a weak positive relationship with share of Obama voters. Lastly, sigma has a mean of 7.10 and a SD of 0.75, with 89% of posterior probability lying in between 6.03 and 8.42 - similar to the previous model. Again, R-hat values of 1 suggests that all the chains have converged evenly.

Answer to b and c)

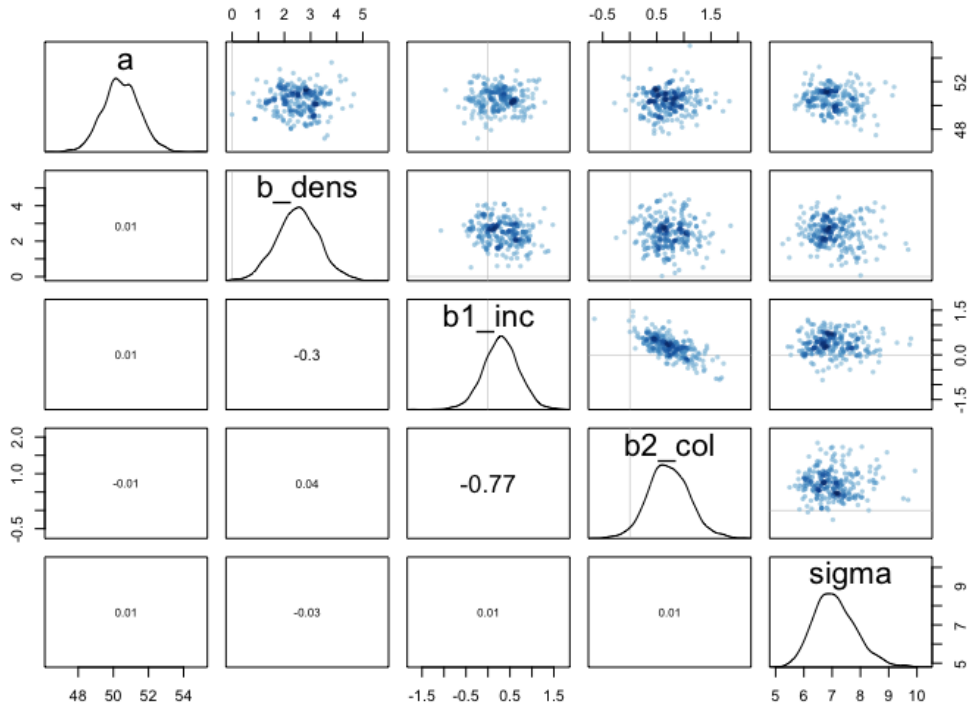
Figure 15:

```
> compare(Ulam_col, Ulam_inc , Ulam_col_inc , func = WAIC)
               WAIC    SE dWAIC  dSE pWAIC weight
Ulam_col      339.2  9.26   0.0   NA   3.2   0.73
Ulam_col_inc 342.0  9.83   2.8 2.06   5.0   0.18
Ulam_inc      343.4 11.09   4.3 6.84   3.8   0.09
```

The WAIC is the log-posterior-predictive-density, providing information about the out-of-sample predictive performance of a model. The *lower* the WAIC value, the lower out of sample deviance and the more accurate WAIC thinks that the model predicts. The SE is the approximate standard error for each WAIC over the data points, while dSE is the standard error of the difference in WAIC values and the top-ranked model. The dWAIC is the difference between each model's WAIC and the best WAIC in the set, while the pWAIC is the estimated effective number of parameters. Weight (or Akaike weight) is the relative support for each model (McElreath 2020, p. 225-9).

Hence, the results from WAIC suggests that the model with college and density as predictors (Ulam_col) gives more accurate predictions as it has a lower WAIC value and WAIC SE, as well as being the top-ranked dWAIC (with model Ulam_col_inc being 2.8 units of deviance smaller), and having the highest Akaike weight of 0.73. However, that does not mean that the college and density variables are causally related with Obama voter turnout, as there may be other omitted or unobserved variables creating bias.

Answer to d)



In light of the comparison in part (b) the above pairs plot og model Ulam_col_inc (with income, density and college as predictors) shows that two of the predictor variables income (b1_inc) and college (b2_col) are negatively correlated with one anther (-0.77), as well as income(b1_inc) being weakly negatively correlated with density (b_dens), suggesting that the model is suffering from multicollinearity when income is introduced to the model. The predictive effect of income on share of Obama voters could also be affected by various omitted variables, which is may be why WAIC suggests that Ulam_col produces overall better predictions compared to Ulam_inc and Ulam_col_inc.

4.4 Script

```
#### Exam Week 4
library(rethinking)
library(rstan)

### Question 1

## a)

#Pr(false positive test | omicron) = 0.1
pr_falsepos <- 0.1 # unconditional prob
#Pr(true positive test | omicron) = 0.9
pr_truepos <- 0.9 # unconditional prob

#Pr(false negative test | omicron) = 0.1
pr_falseneg <- 0.1 # unconditional prob
#Pr(true negative test | omicron) = 0.9
pr_trueneg <- 0.9 # unconditional prob
#Pr(positive in population | omicron ) = 0.01
pr_popomi <- 0.01 # unconditional prob

#Pr(positive | omicron) = 0.08333333
Pr_positive <- pr_truepos * pr_positiveomi + pr_falsepos * (1 -
    pr_popomi)
Pr_positive # 0.108

## b)

#Pr(omicron | positive) = 0.08333333
Pr_Truepos_Trueomi <- (pr_truepos * pr_popomi) / Pr_positive
Pr_Truepos_Trueomi # 0.08333333
```

```
###Question 2

#loding states2
load("~/OneDrive - PRIO/Bayes/states2.Rda")
precis(states2) #exploring

## a)

precis(states2$density) # exploring the data
dens(states2$density) # exploring the data
states2$density_centered <- (states2$density - mean(
   states2$density)) # making a new centered variable
precis(states2$density_centered)

xbar <- mean(states2$density)


#Making a linear model with quap w/weakly informative priors
Mod1 <- quap(
  alist(
    obama08 ~ dnorm(mu , sigma) ,
    mu <- a + b1*(density - xbar) ,
    a ~ dnorm( 50 , 20 ) ,
    b1 ~ dnorm( 0 , 10 ) ,
    sigma ~ dunif( 0 , 20 )
  ) , data=states2 )


## b and c)

##Using Hamiltonian Monte Carlo (HMC)

# generating data for priors
set.seed(1234)
```

```
N <- 100
a <- rnorm( N , 50, 20)
b1 <- rnorm( N , 0 , 10)


Ulam1 <- ulam(
  alist(
    obama08 ~ dnorm(mu , sigma) ,
    mu <- a + b1*(density - 194.962) ,
    a ~ dnorm( 50 , 20 ) ,
    b1 ~ dnorm( 0 , 10 ) ,
    sigma ~ dunif( 0 , 20 )
  ) , data=states2 , chains=4, log_lik = TRUE
)

#generating summary for first model 1
precis(Ulam1, depth = 2)
traceplot(Ulam1, chains=4)
trankplot(Ulam1)

## d)

#renaming variable
Vote_Obama <- states2$obama08

set.seed(1234)
N <- 100
a <- rnorm( N , 50, 20)
b <- rnorm( N , 0 , 10)

#Plotting postrior mean estimate
plot( Vote_Obama ~ density , data=states2, col=rangi2 ,
xlab="Population Density (State)" , ylab="Vote Share for Obama
   ",
```

```
main = "Posterior Mean Estimate" )
post <- extract.samples( Ulam1 ) #extracting samples from
   previous model
a_map <- mean(post$a)
b_map <- mean(post$b)
curve( a_map + b_map*(x - xbar) , add=TRUE ) # adding a line
   for the posterior mean estimate

## e)



#Making a logged model

#log transforming variables
states2$density_logged <- log(states2$density)
states2$obama08_logged <- log(states2$obama08)
xbar_logged <- log(xbar)

# generating data for priors
set.seed(1234)
N <- 100
a <- rnorm( N , 50, 20)
b_log <- rnorm( N , 0 , 10)

#Specifying logged model
Ulam1_log <- ulam(
  alist(
    obama08 ~ dnorm(mu , sigma) ,
    mu <- a + b_log*(density_logged - 5.272805) ,
    a ~ dnorm( 50 , 20 ) ,
    b_log ~ dnorm( 0 , 10 ) ,
    sigma ~ dunif( 0 , 20 )
  ) , data=states2 , chains=4, log_lik = TRUE
)
```

```
###Making scatterplot with raw values
plot( Vote_Obama ~ density , data=states2, col=rangi2,
xlab="Population Density (State)" , ylab="Vote Share for Obama
   ",
main = "Posterior Scatterplot")
post <- extract.samples( Ulam1_log )
a_map <- mean(post$a)
b_map <- mean(post$b)

#Adding posterior mean estimate on logged model
plot( Vote_Obama ~ density , data=states2, col=rangi2,
     xlab="Population Density (State)" , ylab="Vote Share for
        Obama",
     main = "Posterior Mean Estimate ")
post <- extract.samples( Ulam1_log )
a_map <- mean(post$a)
b_map <- mean(post$b)
curve( a_map + b_map*(x - 5.272805) , add=TRUE ) # adding a
   line for the posterior mean estimate

##Fitting a logged model on raw values
plot( Vote_Obama ~ density , data=states2, col=rangi2,
xlab="Population Density (State)" , ylab="Vote Share for Obama
   ",
main = "Posterior Mean Curve (logged)")
post <- extract.samples( Ulam1_log )
a_map <- mean(post$a)
b_map <- mean(post$b)
curve( a_map + b_map*(log(x) - 5.272805) , add=TRUE )
# The logged curvelinear estimates suggests diminishing returns
    on the population
# density of the state on voter share for Obama.
```

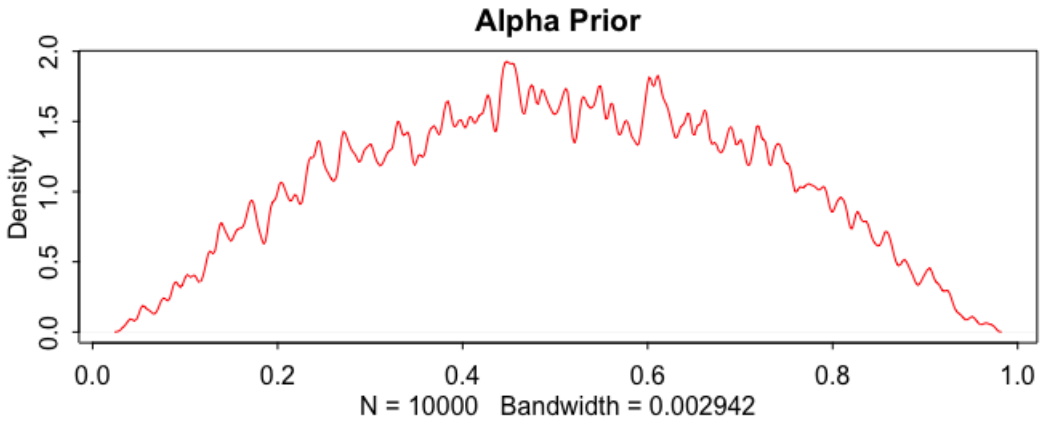
```
compare(Ulam1, Ulam1_log, func = WAIC)
s# WIAC comparison suggests that the logged model produces
   better estimates,
# having a lower WAIC value and standard error.


### Question 3

## a)

# Making slimmer dataset with relevant variables
dslim <- list(obama08 = states2$obama08,
              density = states2$density,
              college = states2$college,
              prcapinc = states2$prcapinc/1000
)

#adding logged density to build on 2.e
dslim$density_logged <- log(dslim$density)


#transforming income variable
xbardensity <- mean(dslim$density)
xbarcollege <- mean(dslim$college)
xbarprcapinc <- mean(dslim$prcapinc)
xbardensity_logged <- mean(dslim$density_logged)

# choosing priors
set.seed(1234)
N <- 100
a <- rnorm(N, 50, 10)
b_dens <- rnorm(N, 0, 5)
```

```
b1_inc <- rnorm(N, 0, 10)
b2_col <- rnorm(N, 0, 10)

#Model 1 - Including income as predictor for share of Obama
   voters
Ulam_inc <- ulam(
  alist(
    obama08 ~ dnorm(mu , sigma) ,
    mu <- a + b1_inc*(prcapinc - 31.9511) +
    b_dens*(density_logged - 4.493501),
    a ~ dnorm( 50 , 10 ) ,
    b_dens ~ dnorm( 0 , 5 ) ,
    b1_inc ~ dnorm( 0 , 10 ) ,
    sigma ~ dunif( 0 , 20 )
  ) , data=dslim , chains=4, log_lik = TRUE
)

precis(Ulam_inc, depth = 2)
summary(Ulam_inc)

#Model 2 - Including college as predictor for share of Obama
   voters
Ulam_col <- ulam(
  alist(
    obama08 ~ dnorm(mu , sigma) ,
    mu <- a + b2_col*(college - 25.846) +
      b_dens*(density_logged - 4.493501),
    a ~ dnorm( 50 , 10 ) ,
    b_dens ~ dnorm( 0 , 5 ) ,
    b2_col ~ dnorm( 0 , 10 ) ,
    sigma ~ dunif( 0 , 20 )
  ) , data=dslim , chains=4, log_lik = TRUE
)
```

```
precis(Ulam_col, depth = 2)

# Model 3 - Including both college and income as predictors

Ulam_col_inc <- ulam(
  alist(
    obama08 ~ dnorm(mu , sigma) ,
    mu <- a + b1_inc*(prcapinc - 31.9511) +
      b2_col*(college - 25.846) +
      b_dens*(density_logged - 4.493501),
    a ~ dnorm( 50 , 10 ) ,
    b_dens ~ dnorm( 0 , 5 ) ,
    b1_inc ~ dnorm( 0 , 10 ) ,
    b2_col ~ dnorm( 0 , 10 ) ,
    sigma ~ dunif( 0 , 20 )
  ) , data=dslim , chains=4, log_lik = TRUE
)

precis(Ulam_col_inc, depth = 2)

## b & c)

compare(Ulam_col, Ulam_inc , Ulam_col_inc , func = WAIC)
compare(Ulam_col, Ulam_inc , Ulam_col_inc , func = WAIC)@dSE

# DOUBLE BAM! the model with college and density as predictors
   has the lowest
# WAIC value.

## d)

pairs(Ulam_col_inc) # producing a pairs plot
```

5 Week 5

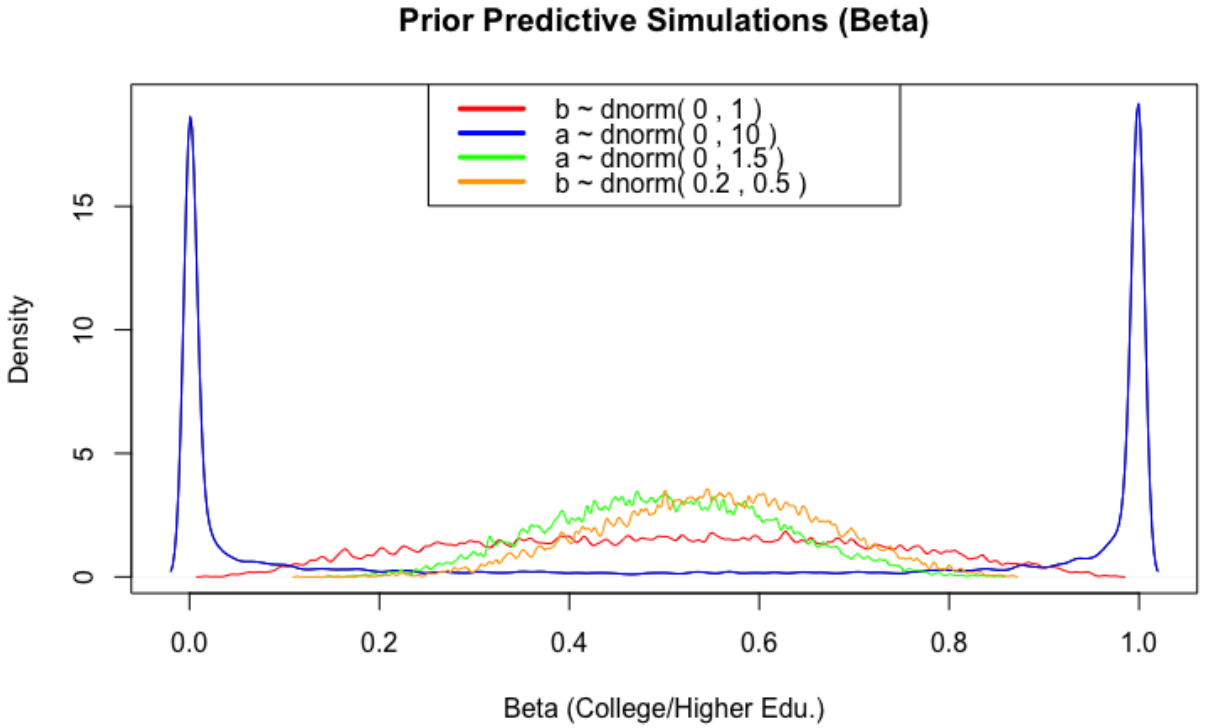
5.1 Question 1



For my constant, the alpha, I chose a generic weakly informative prior with a mean of 0 and a standard deviation of 1 ($\mathbf{a} \sim \mathbf{(\ 0\ ,\ 1\ )}$) producing a truncated normal distribution. I chose this prior as the alpha covers most of the outcome scale with heightened densities for the intercepts closer to the center, and less density to the extreme probability values closer to 0 and 1, hence having a regularising effect by attributing less density to the improbable outcomes.[6]

[6] This prior is displayed on the inverse-logit scale.

For the simulations of the beta (coefficient) prior I ran 5 different prior predictive simulations, in which I for explanatory reasons are showing 4 different possible distributions below.[7]

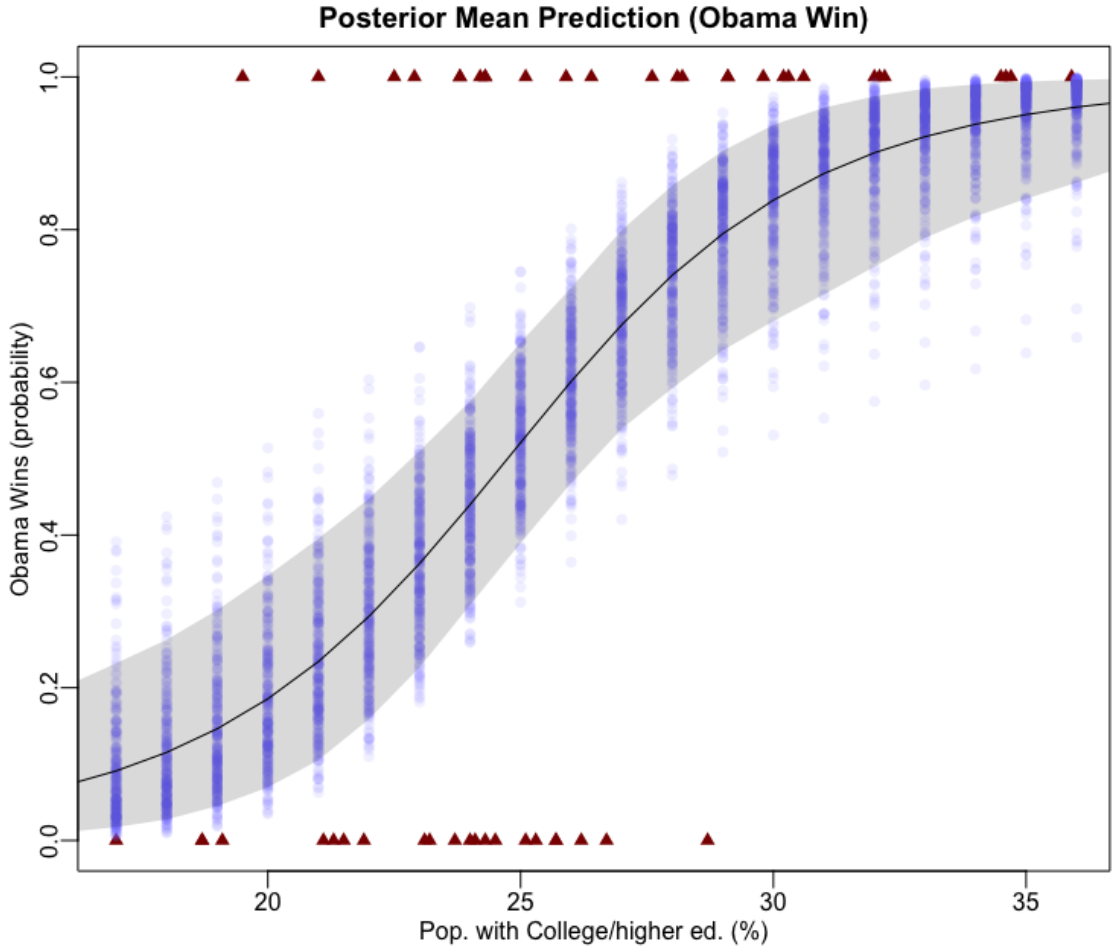


According to the principle of maximum entropy, if *nothing* is known about a distribution except that it belongs to a certain class of distributions, then the distribution with the largest entropy should be chosen as the least-informative default (in this case, the red distribution $\mathbf{b \sim (\ 0\ ,\ 1\ )}$). However, I will assume that the college parameter refers to something that we have background information about; that voters with college or higher education tend to be more prone to vote for liberal candidates (such as Obama). It is therefore theoretically probable that the beta of college is located on the end of the outcome scale. I therefore chose the orange beta prior $\mathbf{b \sim (\ 0.2\ ,\ 0.5\ )}$, which is another weakly informative prior (to avoid biasing the result)

[7] These priors are also displayed on the inverse-logit scale. The blue prior, $\mathbf{b \sim (0\ ,\ 10)}$, is an example of a wrongly scaled flat prior for the logit space, which is not a flat prior on the outcome probability space. All the density (probability mass) is piled on the area around 0 and 1.

with a slight skew towards to higher end of the inverse logit scale.

5.2 Question 2



The above plot shows a visualization of the posterior mean prediction of Obama winning in 2008.[8] Eyeballing the plot suggests that percentage of population with college educated/higher educated is a good predictor of which states Obama won in 2008. The red triangular points are the actual values of college in different states and their actual outcomes in 2008,[9]

[8] The span of the college variable starts at a minimum of 17% college educated population and all the way up to maximum of 35.9% college educated population. I present a sequence of values from 15 to 40, an approximation of the above.

[9] 1 = Obama winning, 0 = Obama not winning.

while the shaded area is the credible interval of the posterior mean prediction. With an assigned 89% credible interval, the predicted posterior mean estimates fall within 89% of the shaded area, and 11% outside the shaded area. The blue dots represent 250 simulated posterior means, adding another visual representation of the full posterior interval.

5.3 Script

```
#### Exam Week 5
library(rstan)
library(rethinking)



## Question 1

#loading data
load("~/OneDrive - PRIO/Exam W5 STV4020B/states2.Rda")

states2$college_centered <- states2$college - mean(
   states2$college)
mean(states2$college_centered) # checking centered college
   variable
colcen <- mean(states2$college)

# Prior predictive distribution of alpha (intercept) on the
   outcome scale

Alpha_prior <- quap(
  alist(
    obama_win08 ~ dbinom( 1 , p ) ,
    logit(p) <- a ,
    a ~ dnorm( 0 , 1.5 )
  ) , data=states2 )
```

```
set.seed(1234)
prior <- extract.prior( Alpha_prior , n=1e4 )

p <- inv_logit( prior$a )
dens( p , adj=0.1 , col="red", main = "Alpha Prior")


# Prior predictive simulations of beta on the outcome scale

Prior_simulation1 <- quap(
  alist(
    obama_win08 ~ dbinom( 1 , p ) ,
    logit(p) <- a + b*(college-25.8) ,
    a ~ dnorm( 0 , 1.5 ) ,
    b ~ dnorm( 0 , 1) #
  ) , data=states2 )


Prior_simulation2 <- quap(
  alist(
    obama_win08 ~ dbinom( 1 , p ) ,
    logit(p) <- a + b*(college-25.8) ,
    a ~ dnorm( 0 , 1 ) ,
    b ~ dnorm( 0 , 10) #
  ) , data=states2 )

Prior_simulation3 <- quap(
  alist(
    obama_win08 ~ dbinom( 1 , p ) ,
    logit(p) <- a + b*(college-25.8) ,
    a ~ dnorm( 0 , 1 ) ,
    b ~ dnorm( 0 , 0.5) #
  ) , data=states2 )
```

```
Prior_simulation4 <- quap(
  alist(
    obama_win08 ~ dbinom( 1 , p ) ,
    logit(p) <- a + b*(college-25.8) ,
    a ~ dnorm( 0 , 1 ) ,
    b ~ dnorm( 0 , 2) #
  ) , data=states2 )

Prior_simulation5 <- quap(
  alist(
    obama_win08 ~ dbinom( 1 , p ) ,
    logit(p) <- a + b*(college-25.8) ,
    a ~ dnorm( 0 , 1 ) ,
    b ~ dnorm(0.2 , 0.5) #
  ) , data=states2 )



set.seed(1234) # setting seed
prior_sim <- extract.prior( Prior_simulation , n=1e4 ) #
   extracting priors
prior_sim2 <- extract.prior( Prior_simulation2 , n=1e4)
prior_sim3 <- extract.prior( Prior_simulation3 , n=1e4)
prior_sim4 <- extract.prior( Prior_simulation4 , n=1e4)
prior_sim5 <- extract.prior( Prior_simulation5 , n=1e4)

p10 <- inv_logit(prior_sim$b) #only simulating beta, hence the
   (prior_sim$b)
p11 <- inv_logit(prior_sim2$b)
p12 <- inv_logit(prior_sim3$b)
#p13 <- inv_logit(prior_sim4$b)
p14 <- inv_logit(prior_sim5$b)
```

```
plot(density(p11, adj=0.1 ), main="Prior Predictive Simulations
    (Beta)", xlab="Beta (College/Higher Edu.)")
lines(density(p10 , adj=0.1), col="red") # b ~ dnorm( 0 , 1)
lines(density(p11 , adj=0.1), col="blue") # a ~ dnorm( 0 , 10 )
lines(density(p12 , adj=0.1), col="green") # a ~ dnorm( 0 , 1.5
    )
#lines(density(p13 , adj=0.1), col="purple") # b ~ dnorm( 0 ,
   2)
lines(density(p14 , adj=0.1), col="orange") # b ~ dnorm( 0.2 ,
   0.5 )

legend("top", legend = c("b ~ dnorm( 0 , 1 )", "a ~ dnorm( 0 ,
   10 )", "a ~ dnorm( 0 , 1.5 )", "b ~ dnorm( 0.2 , 0.5 )"), #
   plotting explanations
 col = c("red", "blue", "green", "orange"), lwd=3)



## Question 2
summary(states2$college) # Min value is 17.00, Max is 35.9
# Min. 1st Qu. Median Mean 3rd Qu. Max.
# 17.00 23.12 25.10 25.85 29.00 35.90
college.seq <- seq(from=15, to=40, by=1 )
mu <- link(Prior_simulation5, data=data.frame(college=college.
   seq)) # college made into seqence and attached to mu
str(mu)
plot(states2$college, states2$obama_win08, pch = 17, col="
   darkred", xlab="Pop. with College/higher ed. (%)" , ylab="
   Obama Wins (probability)" ,
     main = "Posterior Mean Prediction (Obama Win)") #making
        plot
for (i in 1:250)
  points(college.seq , mu[i,], pch=16, col=col.alpha(rangi2,
     0.1), # simulating blue dots
```

```
  )
mu.mean <- apply( mu, 2 , mean) #applying mean
mu.PI <- apply (mu, 2, PI, prob= 0.89) # applying posterior
   interval (PI).
lines(college.seq, mu.mean) # adding line graph to show mean.
shade(mu.PI, college.seq) # adding shade on graph to show PI.
```

Bibliography

Cunningham, S. (2021). *Causal Inference - The Mixtape*. Yale University Press.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., and Townshend, J. R. G. (2013). High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science*, 342(6160):850–853. Publisher: American Association for the Advancement of Science.

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*. CRC Press.

Sarma, A. and Kay, M. (2020). Prior Setting in Practice: Strategies and Rationales Used in Choosing Prior Distributions for Bayesian Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–12, New York, NY, USA. Association for Computing Machinery.

Suhardjanto, D., Syafruddin, M., Andini, R. P., and Rahmatika, M. W. (2018). Accountability and Corruption Level of Provincial Government in Indonesia. *Review of Integrative Business and Economics Research*, 7(3):37.