Problem Set 3

Se Hyun Kim

Problem 1A

There are four conditions for the equation to be a valid estimation of the casual effect between the two variables.
1. The `COLLEGE` variable must be independent of ϵ.
2. The relationship between `COLLEGE` and `VOTE` variables must be linear.
3. There must be no confounding variables between `COLLEGE` and `VOTE` variables. 4. SUTVA With the exception of the second assumption, other assumptions are not likely to hold. The second assumption is likely to hold since `VOTE` and `COLLEGE` are both binary variables. The first condition is unlikely to hold since there are many factors that play into people's voting behavior. They might be a party supporter. They might feel the duty to vote strong. Or they might not vote because it rained on the voting day. The third condition is unlikely to hold. It could be that a person with greater political interest might both go to college and vote. The education level of a parent might influence an individual's likely chance of going to college and voting. SUTVA could be violated if those who went to college affect the voting behavior of those who didn't go to college.

Problem 1B

One could use the existence of college in a town as an instrumental variable. The model that employs an instrumental variable would be estimating the effect of the treatment on compliers. In this case, the compliers are those who live in towns with colleges and got a college education and those who live in towns without colleges and did get a college education.
There are five assumptions required for using instrumental variables. 1. SUTVA 2. Exogeneity of the instrument 3. Exclusion Restriction 4. Monotonicity 5. Relevance For a detailed discussion on these assumptions, refer to Problem 3B.

Problem 1C

The omission of 'political interest' will over-estimate the effect of `COLLEGE` on `VOTE` if the relationships between `political interest` and both `VOTE` and `COLLEGE` are positive. Using an instrumental variable would lower the estimate. Since using the instrumental variable functions as randomization, it will remedy the exogenous effect.

Problem 2A

Intention-to-treat (ITT) effect can be calculated as $\mathbb{E}[Y_i(1, D_i(1)) - Y_i(0, D_i(0))]$ which is $0.35 - 0.50$. The ITT is -0.15. For the ITT to be valid, it must satisfy two assumptions: SUTVA and exogeneity of the instrument. SUTVA is unlikely to hold since students living in dormitories are likely to socialize with one another. Those watching FOX news will affect those in the control group. Since we are only discussing ITT, the instrument only needs to be independent of the outcome. This is likely to hold since the assignment to treatment was randomized. It is unlikely that the approval of the control group units affected their encouragement.

Problem 2B

The average treatment effect for the compliers (LATE) is $\frac{E[Y_i|Z_i=1]-E[Y_i|Z_i=0]}{E[D_i|Z_i=1]-E[D_i|Z_i=0]}$. If we input the values into this equation, we get the following result. $LATE = \frac{0.35-0.50}{(148/200)-(0/200)}$

```
late2b <- (0.35-0.5)/(148/200 - 0/200)
```

Assuming SUTVA, this is a one-sided noncompliance (OSN) case: the control units have no access to the treatment. OSN guarantees monotonicity. Therefore, LATE, the causal effect of encouragement, ($\mathbb{E}[Y_i(1) - Y_i(0)|D_i(1) = 1]$ equals ATT ($\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]$). The causal effect of encouragement is -0.2027027. In other words, watching Fox news causes the Biden's approval rate to change -20.2702703%.

Problem 2C

```
# verifying 2A
v2a <- lm(treat ~ enc, approval)
summary(v2a)
```

```
##
## Call:
## lm(formula = treat ~ enc, data = approval)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -0.74   0.00   0.00   0.26   0.26
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.087e-15  2.199e-02     0.0        1
## enc          7.400e-01  3.109e-02    23.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3109 on 398 degrees of freedom
## Multiple R-squared:  0.5873, Adjusted R-squared:  0.5863
## F-statistic: 566.4 on 1 and 398 DF,  p-value: < 2.2e-16
```

```
# verifying 2B
v2b_a <- lm(treat ~ enc, approval)$fitted
v2b <- lm(approve ~ v2b_a, approval)
summary(v2b)
```

```
##
## Call:
## lm(formula = approve ~ v2b_a, data = approval)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5000 -0.3875 -0.3500  0.5000  0.6500
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.50000    0.03464  14.435  < 2e-16 ***
## v2b_a       -0.20270    0.06620  -3.062  0.00235 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4898 on 398 degrees of freedom
## Multiple R-squared:  0.02302,    Adjusted R-squared:  0.02056
## F-statistic: 9.377 on 1 and 398 DF,  p-value: 0.002346
```

Problem 2D

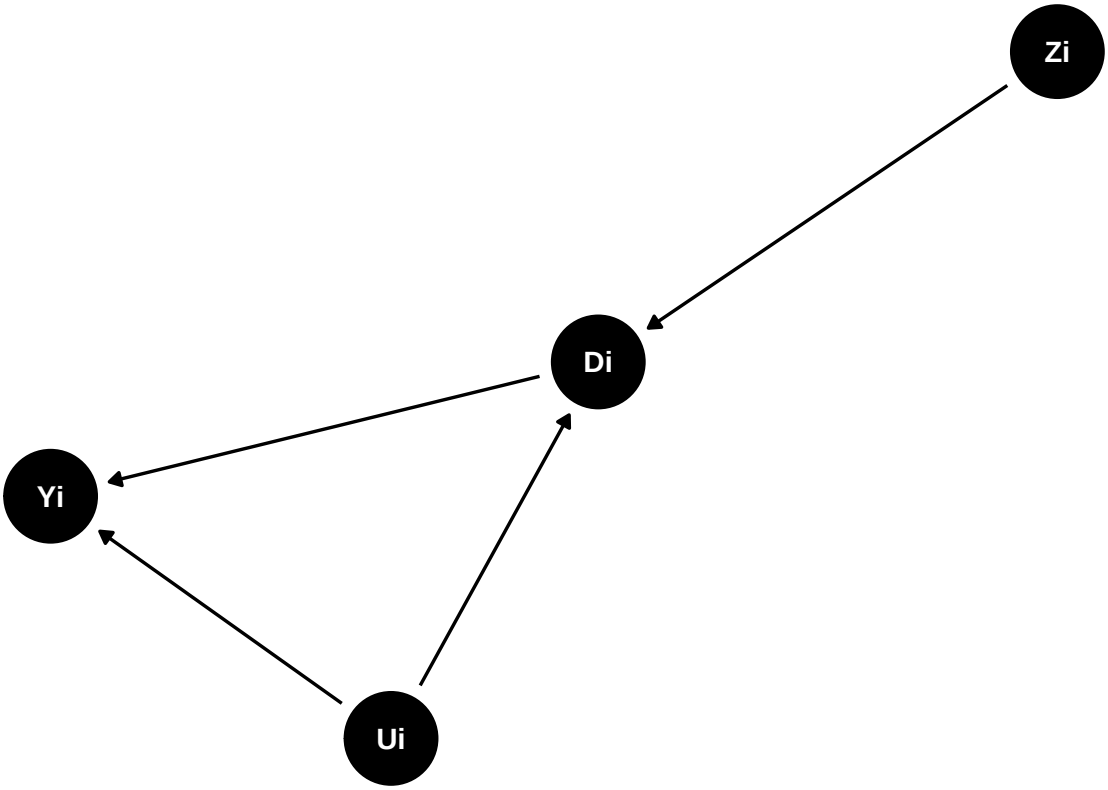
```
func2d <- function(x, y, z){
  # taking in X, Y, Z
  x <- as.matrix(cbind(1,x))
  y <- as.vector(y)
  z <- as.matrix(cbind(1,z))
  # compute IV estimator using one step
  ivcoef <- solve(t(z)%*%x)%*%t(z)%*%y
  # compute standard error
  u_hat <- y - x %*% ivcoef
  var <- t(u_hat)%*%u_hat/(nrow(x) - ncol(x))
  st_er <- sqrt(var*(diag(solve(t(x)%*%(z%*%solve(t(z)%*%z)%*%t(z))%*%x))))
  # compute t-value and p-value
  tval <- (ivcoef - 0)/st_er
  pval <- pt(abs(tval), nrow(x) - ncol(x), lower.tail = F)*2
  # making a table

  results <- cbind(ivcoef, st_er, tval, pval) %>%
    as_tibble()
  results %>%
    kbl(align=rep('c'),
        digits = 5,
        col.names = c("Coefficient",
                      "Standard Error",
                      "T-Statistic",
                      "P-Value")) %>%
  add_header_above(c("Local Average Treatment Effect" = 4)) %>%
  kable_minimal(full_width = F)
}
```

Problem 3A

```
dag3a <- dagify(Yi ~ Di + Ui,
                Di ~ Zi + Ui)

ggdag(dag3a) + theme_dag()
```



It is important to include the hypothetical unobserved confounder because it accounts for possible backdoor paths. In a complex world, this is quite likely. For the case of AJR, the former colonies' access to seaports might be a confounder. It is more likely that regions that can be accessed conveniently by seaports are more likely to be ruled directly by the imperial government thus having a positive correlation with modern property rights institutions. Also, this access could affect the modern log GDP per capita since it allows for easier trade. Another confounder could be the level of average education based on the western social system. Those well versed in European education are more likely to push for modern property rights and institutions that guarantee it. Also, this education will allow the citizens to better communicate with the European states which could lead to better economic performance.

Problem 3B

The five assumptions required for employing instrumental variables are (1) SUTVA, (2) exogeneity of the instrument, (3) exclusion restriction, (4) monotonicity, and (5) relevance.

SUTVA: No Inference & No Hidden Variations of Treatments

SUTVA assumption must hold for $D_i(z)$ and $Y_i(z, d)$
For the AJR case, there are four ways in which SUTVA can be violated. SUTVA will be violated if former colonies with low mortality influence the possession of modern property institutions of former colonies with high mortality rate and vice versa. SUTA will also be violated if former colonies with modern property right effect the GDP per capita of former colonies without modern proerty right and vice vera. This assumption will have to be argued theoretically.
One way in which SUTVA could be violated is if two adjacent former colonies A and B, where A does have modern property institutions and B doesn't have modern property institutions, have high economic

relevance. That is, if A's wealth affects B in a way that B acquires high GDP without modern property institutions.
It is difficult to assess the validity of this assumption given that the interaction between different states become much more complex due to globalization.

Exogeneity of the Instrument: The IV must be independent from the encouragement and outcome.

$\{Y_i(1), Y_i(0), D_i(1), D_i(0)\} \perp\!\!\!\perp Z_i$ But for ITT (Intention to Treat effect), $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp Z_i$ is enough (however, $\{Y_i(1), Y_i(0)\} \not\perp\!\!\!\perp D_i | Z_i = z$).
For the AJR case, this assumption will be violated if the mortality rate was systematic, that is if the encouragement wasn't "as good as" random. This assumption will have to be argued theoretically.
Exogeneity will be violated if mortality rates were distributed in relation to geography. It could be that tropical regions will suffer from higher mortality rates.

Exclusion Restriction: instrument affects outcome only through treatment

$Y_i(1, d) = Y_i(0, d)$ for $d = 0, 1$
For the AJR case, the exclusion restriction will be violated if the morality rate (17, 18, and 19th century) effect the GDP per capita (1995) directly or indirectly other than through modern property institutions. This assumption will have to be argued theoretically.
Mortality rates could effect GDP per capita in another way other than through modern property institutions. For example, high mortality rate could reduce human resources which will lead to higher GDP per capita.

Monotonicity: no defiers

$D_i(1) \geq D_i(0)$ for all i
For the AJR case, monotonicity will be violated if former colonies have modern property institutions if they have high mortality rate and do not have modern property institutions if they have low mortality rate. This assumption will have to be argued theoretically.
This assumption is likely to hold since it seems quite unlikely that a former colony will adopt modern property institutions because it has high mortality and vice versa.

Relevance: non-zero average encouragement effect

$\mathbb{E}[D_i(1) - D_i(0)] \neq 0$
For the AJR case, relevance assumption will be violated if the morality rate has no relationship with modern property institutions. This assumption can be tested using regression (refer to 3F).
This assumption is likely to hold since it is likely that the more "white" settlers you have, the more likely the imperial government will have an interest and rule the region directly.

Problem 3C

```
c3a <- lm(logpgp95 ~ avexpr, ajr)
c3b <- lm(logpgp95 ~ avexpr + lat_abst + africa + asia + other, ajr)

c3a_rb <- coeftest(c3a, vcov = vcovHC(c3a, type = "HC2"))
c3b_rb <- coeftest(c3b, vcov = vcovHC(c3b, type = "HC2"))
```

```
screenreg(list(c3a_rb, c3b_rb),
          digits = 3,
          custom.header = list("Dependent Variable: 1995 GDP per capita (log) " = 1:2),
          custom.model.names = c("Without Covariates",
                                 "with Covariates"),
          custom.coef.names = c("Intercept",
                                "Protection Against Expropriation",
                                "Latitude of Capital",
                                "African Nation",
                                "Asian Nation",
                                "Other"))
```

```
##
## ============================================================================
##                                    Dependent Variable: 1995 GDP per ca
##                                    -----------------------------------
##                                    Without Covariates  with Covariates
## ----------------------------------------------------------------------------
## Intercept                           4.660 ***           5.737 ***
##                                    (0.322)             (0.396)
## Protection Against Expropriation    0.522 ***           0.401 ***
##                                    (0.050)             (0.066)
## Latitude of Capital                                     0.875
##                                                        (0.628)
## African Nation                                         -0.881 ***
##                                                        (0.154)
## Asian Nation                                           -0.577
##                                                        (0.307)
## Other                                                   0.107
##                                                        (0.251)
## ============================================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

The low p-values for both regression demonstrate that the independent variable has a plausible effect on the dependent variable: the effect is not by chance. When considered without the covariates, one unit increase in `avexpr` will lead to 68.5395071% increase of the log of GDP. The standard error is also relatively low. When considered with the covariates, one unit increase in `avexpr` will lead to 49.3317268% increase of the log of GDP. The estimate is relatively precise with standard error is also relatively low.
However, it would be difficult to argue that these results are a good estimate of the causal quantity of interest. For one, we can question the randomization process and two it is likely that there are omitted variables. The inability to guarantee randomization problematizes whether the model properly represents the population. This concerns the issues of identification. ## Problem 3D

```
d3a <- lm(logpgp95 ~ logem4, ajr)
d3b <- lm(logpgp95 ~ logem4 + lat_abst + africa + asia + other, ajr)

d3a_rb <- coeftest(d3a, vcov = vcovHC(d3a, type = "HC2"))
d3b_rb <- coeftest(d3b, vcov = vcovHC(d3b, type = "HC2"))

screenreg(list(d3a_rb, d3b_rb),
          digits = 3,
          custom.header = list("Dependent Variable: 1995 GDP per capita (log) " = 1:2),
          custom.model.names = c("Without Covariates",
```

```
                                    "with Covariates"),
          custom.coef.names = c("Intercept",
                                "Settler Mortality (log)",
                                "Latitude of Capital",
                                "African Nation",
                                "Asian Nation",
                                "Other"))
```

```
##
## ============================================================
##                           Dependent Variable: 1995 GDP per ca
##                           ---------------------------------
##                           Without Covariates  with Covariates
## ------------------------------------------------------------
## Intercept                 10.731 ***           9.997 ***
##                           (0.385)             (0.767)
## Settler Mortality (log)   -0.573 ***          -0.377 *
##                           (0.074)             (0.145)
## Latitude of Capital                            1.046
##                                               (0.886)
## African Nation                                -0.723 **
##                                               (0.262)
## Asian Nation                                  -0.525
##                                               (0.382)
## Other                                          0.185
##                                               (0.257)
## ============================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

The low p-values for both regression demonstrate that the independent variable has a plausible effect on the dependent variable: the effect is not by chance. When considered without the covariates, one unit increase in `logem4` will lead to 77.3579818% decrease of the log of GDP. The estimate is relatively precise with standard error is also relatively low. When considered with the covariates, one unit increase in `logem4` will lead to 45.7904309% decrease of the log of GDP. The standard error is also relatively low.
These models estimate the ITT. One can interpret this as causal if SUTA and exogeneity of the instrument hold, and all participants are compliers. ## Problem 3E

```
func2d(ajr$avexpr, ajr$logpgp95, ajr$logem4)
```

Local Average Treatment Effect			
Coefficient	Standard Error	T-Statistic	P-Value
1.90967	1.02673	1.85995	0.06764
0.94428	0.15653	6.03275	0.00000

```
func2d(ajr[c("avexpr", "lat_abst","africa","asia","other")],ajr$logpgp95,ajr[c("logem4", "lat_abst","af
```

Local Average Treatment Effect			
Coefficient	Standard Error	T-Statistic	P-Value
1.44045	2.83959	0.50728	0.61389
1.10708	0.46357	2.38814	0.02021
-1.17818	1.75544	-0.67116	0.50478
-0.43727	0.42421	-1.03078	0.30692
-1.04709	0.52456	-1.99612	0.05062
-0.99040	0.99798	-0.99240	0.32512

The low p-values (0 and 0.02) in these two models demonstrate that the effect of the treatment variable on the outcome is not by chance. Without covariates, the model estimates that one unit increase in the average protection against expropriation risk will lead to 157.096162% increase in the log of 1995 GDP. With covariates, the model estimates that one unit increase in the average protection against expropriation risk will lead to 202.5510991% increase in the log of 1995 GDP. These effects are identified for the compliers, former colonies with low mortality rate having modern property institutions and former colonies with high mortality rate having no modern property institutions. However, the standard error is relatively large for the model with covariates (0.46357) and the precision of the estimate is low. ## Problem 3F

```
reg3f <- ivreg(logpgp95 ~ avexpr + lat_abst + africa + asia + other | logem4  + lat_abst + africa + asi

summary(reg3f, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = logpgp95 ~ avexpr + lat_abst + africa + asia +
##     other | logem4 + lat_abst + africa + asia + other, data = ajr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7155 -0.6381 -0.1535  0.8188  2.0714
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.4405     2.8396   0.507   0.6139
## avexpr        1.1071     0.4636   2.388   0.0202 *
## lat_abst     -1.1782     1.7554  -0.671   0.5048
## africa       -0.4373     0.4242  -1.031   0.3069
## asia         -1.0471     0.5246  -1.996   0.0506 .
## other        -0.9904     0.9980  -0.992   0.3251
##
## Diagnostic tests:
##                  df1 df2 statistic p-value
## Weak instruments   1  58     3.456 0.06811 .
## Wu-Hausman         1  57     9.777 0.00278 **
## Sargan             0  NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.082 on 58 degrees of freedom
## Multiple R-Squared: 0.01082, Adjusted R-squared: -0.07445
## Wald test: 6.847 on 5 and 58 DF,  p-value: 4.418e-05
```

The results for 3F presents the same results as the 3E model with covariates. The weak instrument test

gives an f-statistic of 3.456 and a p-value of 0.68. The result fails to reject the null hypothesis and we can conclude that the relevance assumption holds.