Exam

Gergel Anastasia

7/2/2021

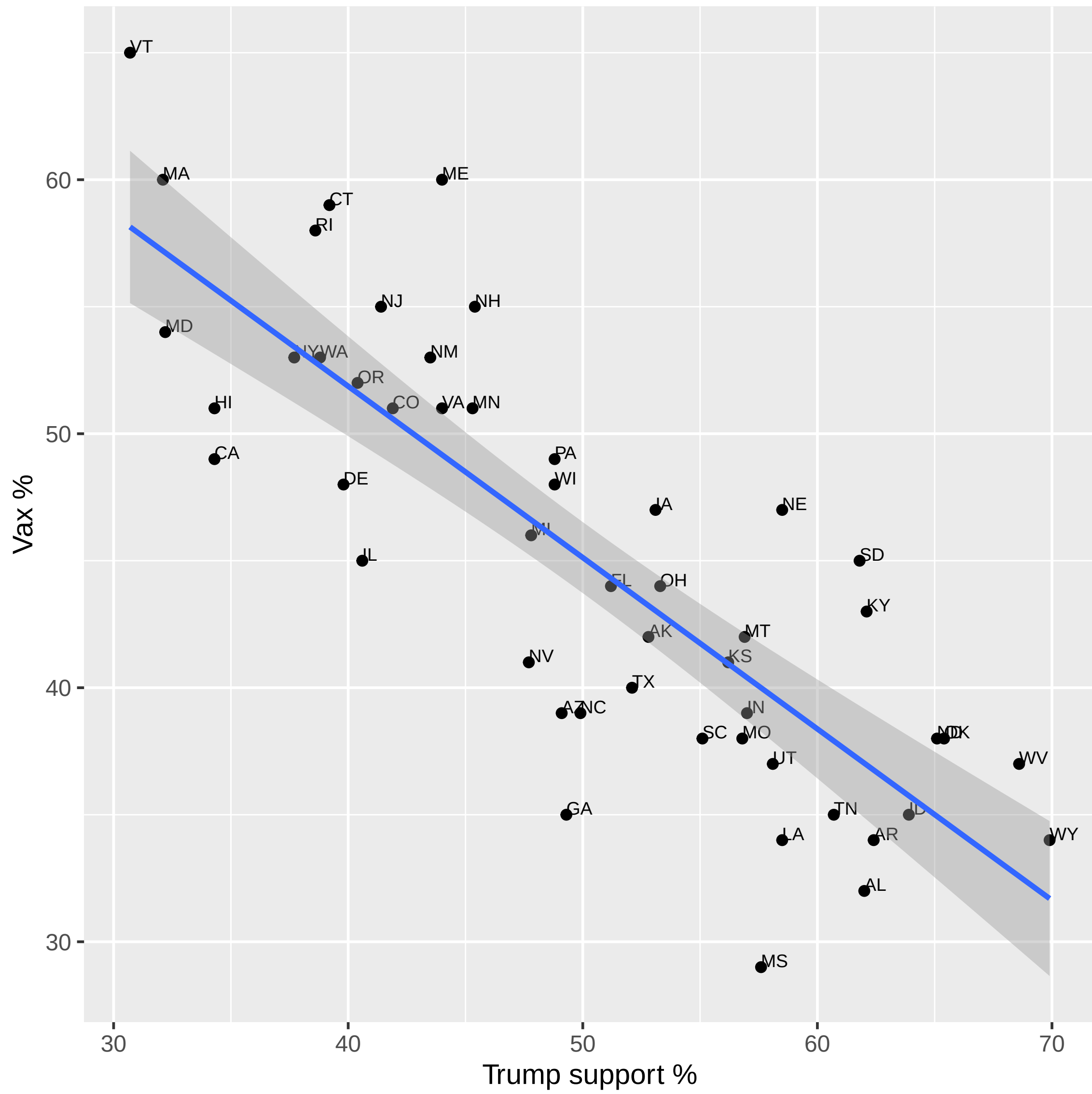
Question 1

(a) Model I.

```r
library(haven)
library(sandwich)

vax <- read_dta("/Users/herrhellana/Dropbox/_NYU studies/Quant I/exam/vax.dta")

# Model I
m1 <- lm(data=vax, vax_pct ~ trump_pct)
summary(m1)
```

```

Call:
lm(formula = vax_pct ~ trump_pct, data = vax)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9972  -3.5450   0.1903   3.2147  10.8296

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 78.84848    3.47841  22.668  < 2e-16 ***
trump_pct   -0.67450    0.06803  -9.914 3.36e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.92 on 48 degrees of freedom
Multiple R-squared:  0.6719,    Adjusted R-squared:  0.665
F-statistic: 98.29 on 1 and 48 DF,  p-value: 3.357e-13
```

```r
robust1 <- vcovHC(m1, type = "HC1")

library(ggplot2)
ggplot(data = vax, aes(x = trump_pct, y = vax_pct,
label=stateabb)) + geom_point() +
geom_text(aes(label=stateabb), size=2.5, hjust=0, vjust=0) +
geom_smooth(method = 'lm') +
labs(x = "Trump support %", y = "Vax %")
```



Trump support is strongly and positively associated with lower vaccination rates at the 99% confidence level. A one-percent change in Trump support lowers vacination rates by 0.99 percent (almost 1%, too). This is almost one-percent to one-percent negative association between the dependent and independent variables.

(b) Finding potential confounders.

```r
library(Hmisc)
library(dplyr)

rcorr(as.matrix(vax %>% select("vax_pct", "trump_pct", "over65_pct", "college_pct")))
```

```
            vax_pct trump_pct over65_pct college_pct
vax_pct        1.00     -0.82       0.19        0.79
trump_pct     -0.82      1.00       0.02       -0.77
over65_pct     0.19      0.02       1.00       -0.15
college_pct    0.79     -0.77      -0.15        1.00

n= 50
```

```
P
            vax_pct trump_pct over65_pct college_pct
vax_pct             0.0000    0.1825     0.0000
trump_pct   0.0000            0.9031     0.0000
over65_pct  0.1825  0.9031               0.3098
college_pct 0.0000  0.0000    0.3098
```

College education and vaccination rates are positively, strongly, and statistically significantly correlated with each other, while college education and Trump support are strongly negatively and statistically significantly correlated. Since college education is associated with both the dependent and independent variables, it would potentially confound the relationship between Trump support and vaccination rates. Older population is not statistically significantly correlated with any of the other variables.

(c) Model II.

```
m2 <- lm(data=vax, vax_pct ~ trump_pct + college_pct + over65_pct)
robust2 <- vcovHC(m2, type = "HC1")

library(stargazer)
stargazer(m1, m2, type='text',
          dep.var.labels = "vax %",
          covariate.labels = c("trump support", "college education",
                               "age", "intercept"),
          se = list(sqrt(diag(robust1)), sqrt(diag(robust2))))
```

```
=================================================================
                                  Dependent variable:
                    ---------------------------------------------
                                         vax %
                              (1)                     (2)
-----------------------------------------------------------------
trump support              -0.675***               -0.380***
                            (0.065)                 (0.079)

college education                                   0.891***
                                                    (0.181)

age                                                 1.169***
                                                    (0.344)

intercept                  78.848***                16.237
                            (3.354)                 (12.070)

-----------------------------------------------------------------
Observations                  50                      50
R2                           0.672                   0.800
Adjusted R2                  0.665                   0.787
Residual Std. Error     4.920 (df = 48)         3.920 (df = 46)
F Statistic         98.290*** (df = 1; 48) 61.472*** (df = 3; 46)
=================================================================
Note:                                 *p<0.1; **p<0.05; ***p<0.01
```

College education was indeed a confounder: it changed the magnitude of the Trump support coefficient. Age is statistically significantly correlated with vaccination rates, and it has also changed the coefficient for Trump support, i.e. it confounds the $y - x$ relationship, too. This is easy to prove just by looking at the Trump support coefficient in the model without age but with college education:

```
# difference with Model II
lm(data=vax, vax_pct ~ trump_pct + college_pct)$coefficient[2]
```

```
 trump_pct
-0.4316893
```

However, neither of confounders have rendered the relationship between vaccination rates and Trump support spurious, it is still strong and significant.

Overall, the average effect of one-percent change in Trump support decreases vaccination rates by 0.38%, holding all other regressors constant.

(d) COVID deaths.

```
rcorr(as.matrix(vax %>% select("vax_pct", "death_rate", "diabetes_pct", "obese_pct")))
```

```
             vax_pct death_rate diabetes_pct obese_pct
vax_pct         1.00      -0.40        -0.48     -0.63
death_rate     -0.40       1.00         0.23      0.10
diabetes_pct   -0.48       0.23         1.00      0.74
obese_pct      -0.63       0.10         0.74      1.00

n= 50


P
             vax_pct death_rate diabetes_pct obese_pct
vax_pct              0.0039     0.0005       0.0000
death_rate   0.0039             0.1080       0.4752
diabetes_pct 0.0005  0.1080                  0.0000
obese_pct    0.0000  0.4752     0.0000
```

Both diabetes and obesity are statistically significantly correlated with vaccination rates but only the latter is statistically significantly correlated with the death rate.

```
death1 <- lm(data=vax, death_rate ~ vax_pct)
robust_death1 <- vcovHC(death1, type = "HC1")
death2 <- lm(data=vax, death_rate ~ vax_pct + diabetes_pct + obese_pct)
robust_death2 <- vcovHC(death2, type = "HC1")

stargazer(death1, death2, type='text',
          dep.var.labels = "recent COVID deaths",
          covariate.labels = c("vax %", "diabetes rate %",
                               "obesity rate %", "intercept"),
          se = list(sqrt(diag(robust_death1)), sqrt(diag(robust_death2))))
```

```
===============================================================
                               Dependent variable:
                    -------------------------------------------
                               recent COVID deaths
                            (1)                   (2)
---------------------------------------------------------------
vax %                    -0.002***             -0.003***
                          (0.001)               (0.001)
```

```
diabetes rate %                                0.009
                                              (0.005)

obesity rate %                               -0.006**
                                              (0.003)

intercept                  0.189***          0.331***
                           (0.026)            (0.093)

-----------------------------------------------------------------
Observations                  50                 50
R2                           0.161              0.247
Adjusted R2                  0.143              0.198
Residual Std. Error    0.043 (df = 48)     0.041 (df = 46)
F Statistic         9.181*** (df = 1; 48) 5.040*** (df = 3; 46)
=================================================================
Note:                                *p<0.1; **p<0.05; ***p<0.01
```

The obesity rate confounds the relationship between vaccination rates and recent COVID deaths, as it changes the magnitude of the vaccination rate coefficient, but it does not make the relationship spurious. The diabetes rate is not statistically significantly correlation with COVID deaths and it does not confound the $y - x$ relationship, again, see a restricted model coefficient:

```r
# identical to the baseline model
round(lm(data=vax, death_rate ~ vax_pct + diabetes_pct)$coefficient[2], 3)
```

```
vax_pct
 -0.002
```

On average, a 10 percent-point rise in vaccination rates is associated with a 0.003 decrease in the COVID death rate, i.e. this rise is associated with COVID-related deaths reduced by 300 in absolute values ($.003 \cdot 100,000$), holding other regressors constant.

(e) No, the answer in (d) is not causally identified. We still hold on on the assumption that $\mathbb{E}(u|x) = 0$, which can be still violated by non-inclusion of other potential confounders, besides the obesity rate (e.g. other chronic diseases like high blood pressure, comorbidities, rare allergic reactions that prevent the immunity build up), that biases the naive estimate of the vaccination rate's effect on covid deaths. Such an ommitted variable can open a back-door path between vaccination rates and COVID deaths. Moreover, the $y - x$ relationship may be confounded by unobservables such as beliefs that the pandemic is not real. These unobservables are obviously correlated with unwilligness to get vaccinated but also with COVID-related deaths as people with such beliefs are prone not to comply with the pandemic regulations and thus are more likely to get infected and/or hospitalized and/or die because of COVID. Figure 1 depicts the causal graph.

While confounding of the $y - x$ relationship by observables can be addressed by conditioning on observables (what we did in the second model here), we need to use other methods to address confounding by unobservables (like finding instruments for x) since we cannot bock the back-door path by conditioning on it.

Question 2

```r
# (a) run a regression
gendereq <- read_dta("/Users/herrhellana/Dropbox/_NYU studies/Quant I/exam/gendereq.dta")

gendereq <- gendereq %>% mutate(gdp_log = log(gdp_k))
```

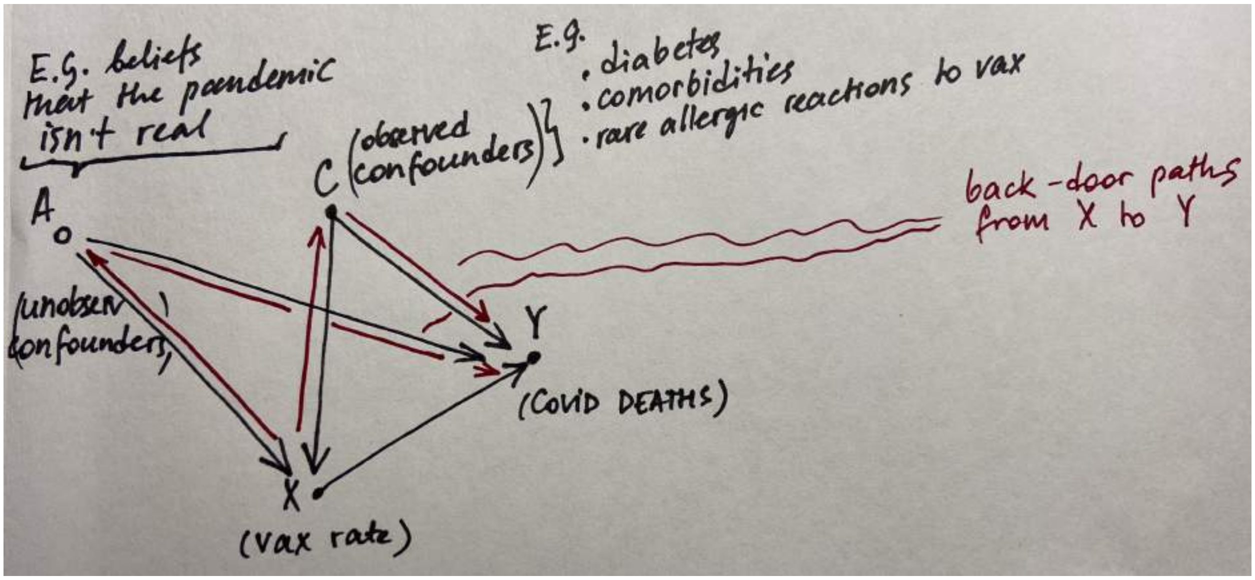


Figure 1: Causal Graph, COVID deaths and vaccination

```r
gender1 <- lm(data = gendereq, femjobs ~ relig + gdp_log)
v_gender1 <- vcovHC(gender1, type = 'HC1')

# (b) Partialling out (Frisch-Waugh-Lovell)
x <- gendereq %>% select(femjobs, relig, gdp_log)
x1star <- lm(data = gendereq, relig ~ gdp_log) %>% resid()
gender_FWL <- lm(x$femjobs ~ x1star)
v_gender_FWL <- vcovHC(gender_FWL, type = 'HC1')

# (c) table
stargazer(gender1, gender_FWL, type = 'text',
          dep.var.labels = c("support for gender equality","support for gender equality"),
          covariate.labels = c("religion", "GDPpc (log)", "religion (FWL)", "intercept"),
          se = list(sqrt(diag(v_gender1)), sqrt(diag(v_gender_FWL))))
```

```
===========================================================================
                                      Dependent variable:
                        ---------------------------------------------------
                        support for gender equality support for gender equality
                                    (1)                         (2)
---------------------------------------------------------------------------
religion                         -0.281***
                                  (0.076)

GDPpc (log)                      5.030***
                                  (1.793)

religion (FWL)                                               -0.281***
                                                              (0.087)

intercept                        44.994***                   41.554***
```

```
                              (6.370)                     (2.628)

---------------------------------------------------------------------------
Observations                     60                          60
R2                              0.361                       0.126
Adjusted R2                     0.338                       0.111
Residual Std. Error      17.564 (df = 57)            20.356 (df = 58)
F Statistic           16.072*** (df = 2; 57)       8.371*** (df = 1; 58)
===========================================================================
Note:                                           *p<0.1; **p<0.05; ***p<0.01
```

When we partial out the effect of religion on support for gender equality, we do not need to include any other variables to get the same effect as the marginal effect of religion in the full model. That is why these estimates of the religion effect are similar.

(d) MENA region bivariate regression

```r
gender_subset <- gendereq[gendereq$region == 1, ] # MENA region

mena_1 <- lm(data=gender_subset, femjobs ~ relig)
v_mena_1 <- vcovHC(mena_1, type = 'HC1')
stargazer(mena_1, type='text',
          dep.var.labels = "support for gender equality",
          covariate.labels = c("religion", "intercept"),
          se = list(sqrt(diag(v_mena_1))))
```

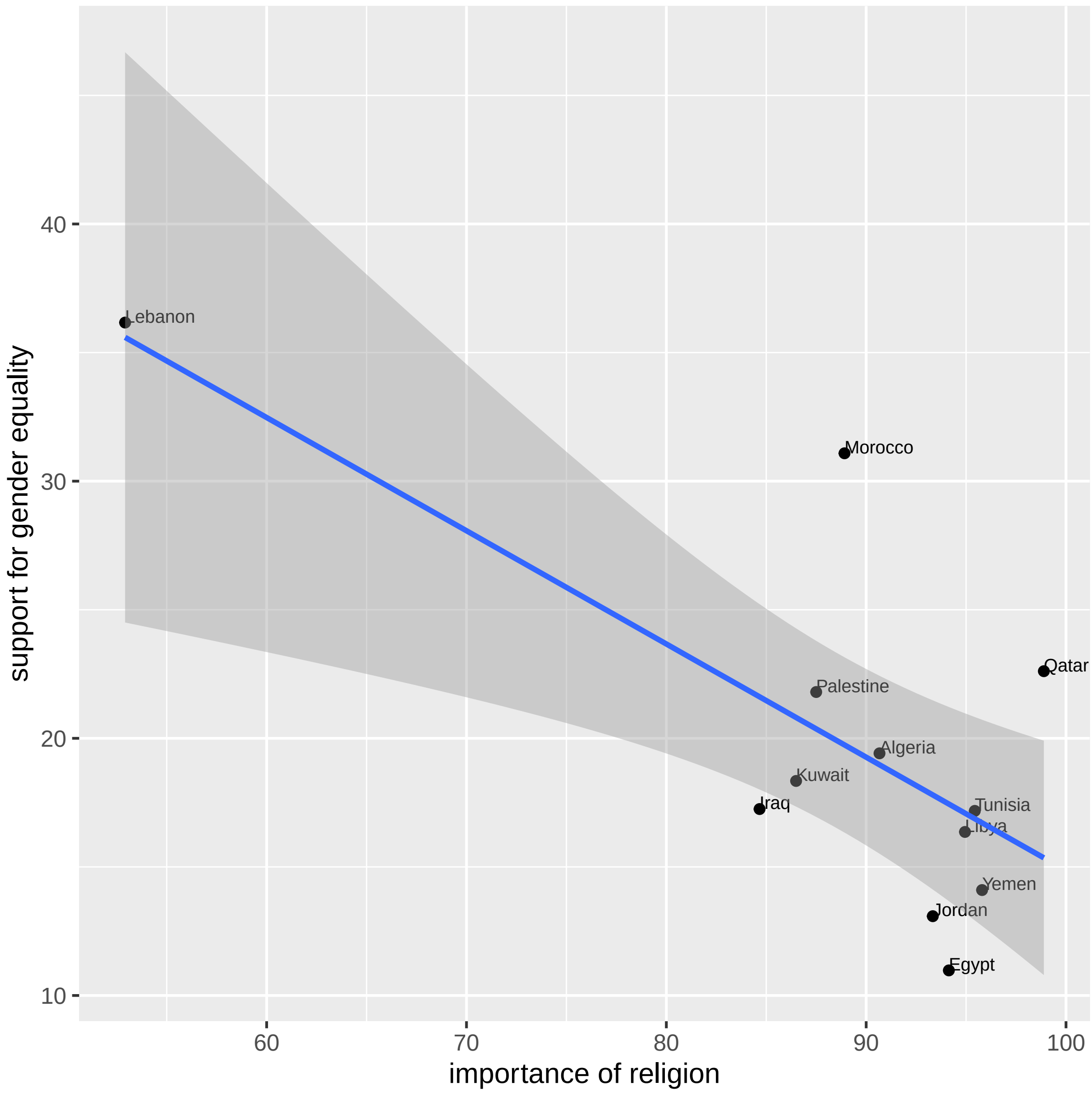
```
===============================================
                         Dependent variable:
                    ---------------------------
                    support for gender equality
-----------------------------------------------
religion                     -0.440***
                              (0.063)

intercept                    58.888***
                              (4.895)

-----------------------------------------------
Observations                     12
R2                              0.524
Adjusted R2                     0.477
Residual Std. Error        5.297 (df = 10)
F Statistic           11.030*** (df = 1; 10)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

The association between religion and support for gender equality is stronger in the MENA region than across all the regions. It is statistically significant at the 99% confidence interval. On average, a one-unit change in response about the importance of religion is associated with a 0.44 decrease in support for gender equality.

(e)

```r
ggplot(data = gender_subset, aes(x = relig, y = femjobs,
label=country)) + geom_point() +
geom_text(aes(label=country), size=2.5, hjust=0, vjust=0) +
geom_smooth(method = 'lm') +
```

```r
labs(x = "importance of religion", y = "support for gender equality")
```



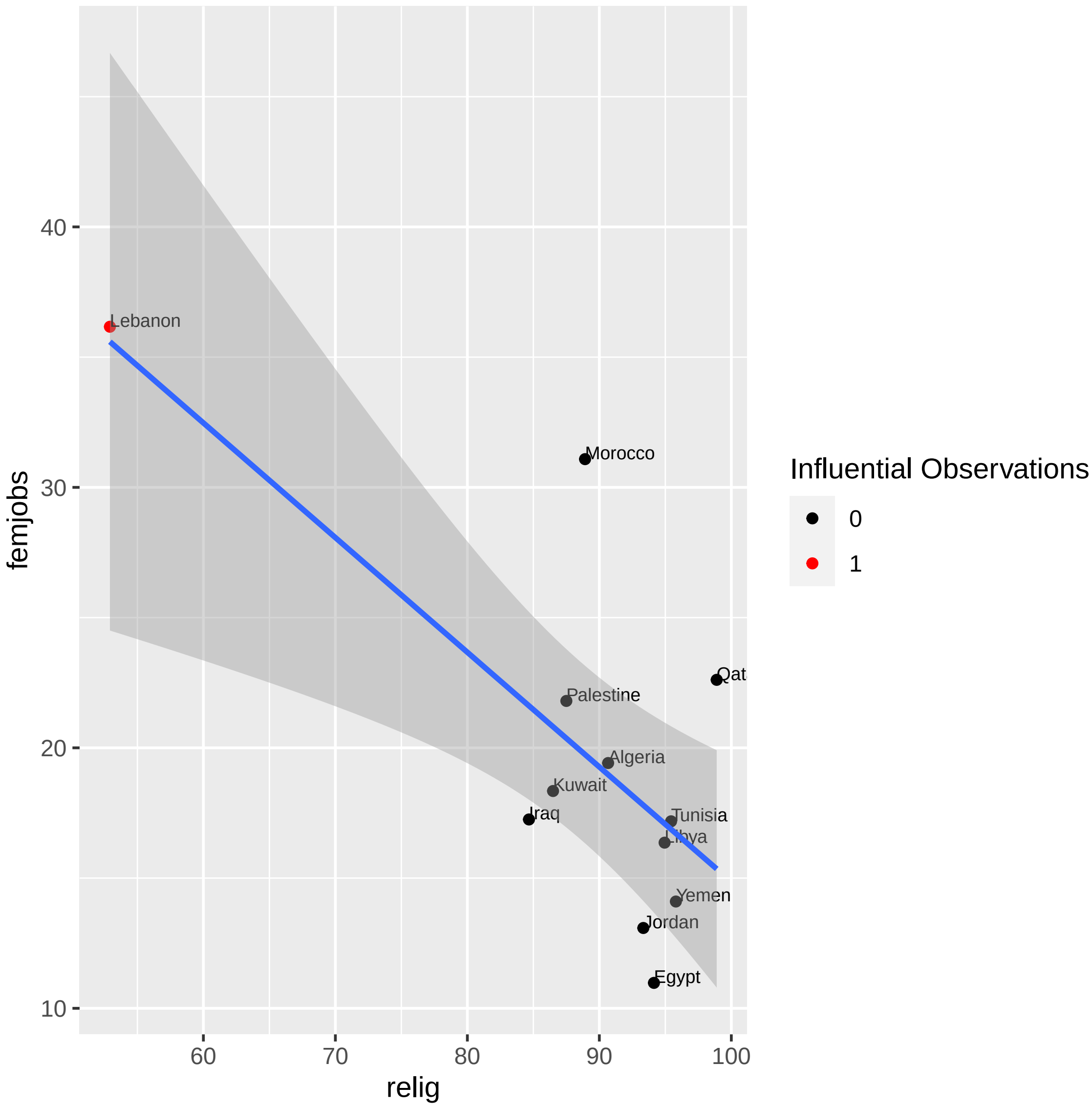
(f) The scatterplot shows that the data has an influential observation, i.e. Lebanon, which is far away from other observations that are clustered together. The regression line and thus its slope (the coefficient of the dependent variable of interest) can change substantially if we exclude the influential observation from the model.

(g) Cook's distance

```r
# top 5% influential observations
obs <- gender_subset %>% mutate(cd = cooks.distance(mena_1)) %>%
  select(country, femjobs, relig, cd)
obs <- obs %>% mutate(infl = as.numeric(cd >= quantile(cd, probs = 0.95)))

ggplot(data = obs, aes(x = relig, y = femjobs, label=country)) +
  geom_point(aes(color = as.factor(infl))) +
  geom_text(aes(label=country), size=2.5, hjust=0, vjust=0) +
  geom_smooth(method = 'lm', formula = str(mena_1$call)) +
  scale_color_manual(values = c('black', 'red')) +
```

```
  labs(color = 'Influential Observations')
```

```
## language lm(formula = femjobs ~ relig, data = gender_subset)
```

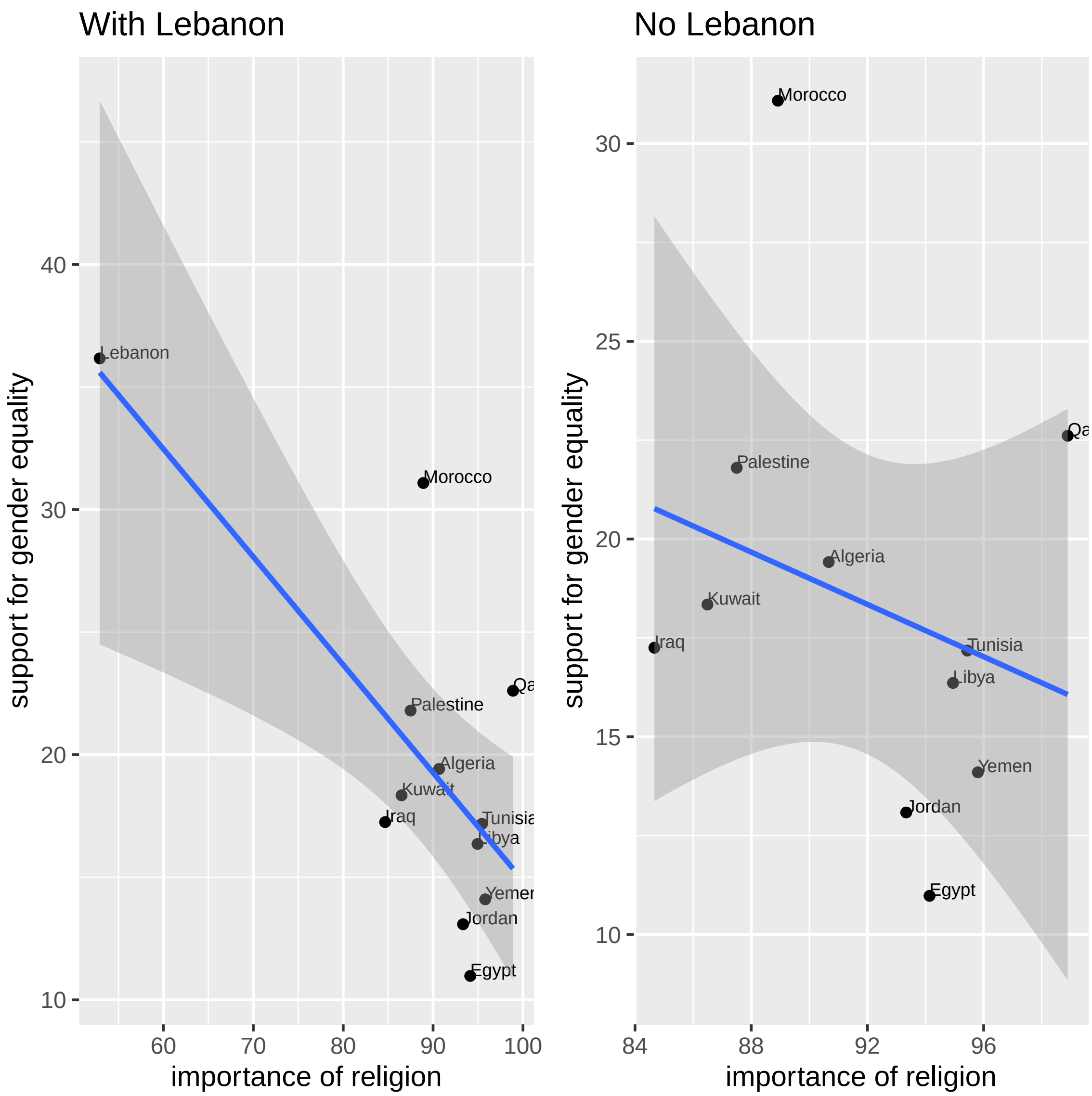


According to the Cook's distance estimate, Lebanon is indeed an influential observation.

We can also visualize how different are two fitted regression lines with and without the influential observation.

```
plot1 <- ggplot(data = gender_subset, aes(x = relig, y = femjobs, label=country)) +
        geom_point() + geom_text(aes(label=country), size=2.5, hjust=0, vjust=0) +
        geom_smooth(method = "lm") +
        ggtitle("With Lebanon") +
        labs(x = "importance of religion", y = "support for gender equality")
plot2 <- ggplot(data = gender_subset[1:11,], aes(x = relig, y = femjobs, label=country)) +
        geom_point() + geom_text(aes(label=country), size=2.5, hjust=0, vjust=0) +
        geom_smooth(method = "lm") +
        ggtitle("No Lebanon") +
        labs(x = "importance of religion", y = "support for gender equality")
```

```
gridExtra::grid.arrange(plot1, plot2, ncol=2)
```



We see that the slope of the regression line changes significantly. Hence, the finding from the regression (d) is not robust. We can also chek it by comparing two bivariate regressions estimated on the full MENA dataset and on its subset without Lebanon.

```
mena_2 <- lm(data=gender_subset[1:11,], femjobs ~ relig)
v_mena_2 <- vcovHC(mena_2, type = 'HC1')

stargazer(mena_1, mena_2, type='text',
          dep.var.labels = "support for gender equality",
          covariate.labels = c("religion", "intercept"),
          se = list(sqrt(diag(v_mena_1)), sqrt(diag(v_mena_2))))
```

```
============================================================
                              Dependent variable:
                    ----------------------------------------
```

```
                          support for gender equality
                             (1)                    (2)
------------------------------------------------------------
religion                  -0.440***              -0.331
                           (0.063)                (0.365)

intercept                 58.888***              48.755
                           (4.895)               (33.680)

------------------------------------------------------------
Observations                 12                     11
R2                          0.524                  0.074
Adjusted R2                 0.477                 -0.028
Residual Std. Error    5.297 (df = 10)       5.555 (df = 9)
F Statistic         11.030*** (df = 1; 10) 0.723 (df = 1; 9)
============================================================
Note:                            *p<0.1; **p<0.05; ***p<0.01
```

The table confirms that the finding from the first model is not robust: the magnitude of the coefficient for importance of religion changes, and the coefficient looses its statistical significance. Note that the coefficient's standard error in the second model substantially increases, too.

Question 3

(a) Replicate Model II from the Table 1 in Egan, Mullin (2012).

```r
warming <- read_dta("/Users/herrhellana/Dropbox/_NYU studies/Quant I/exam/warming.dta")
warming <- warming %>% mutate(getwarmord = as.factor(getwarmord))

# ordinal probit
library(MASS)
warm_II <- polr(data = warming, getwarmord ~ ddt_week + as.factor(doi) +
                  as.factor(statenum) + as.factor(wbnid_num),
                  method = 'probit')

vp_II <- vcovCL(warm_II, cluster = warming$statenum, type = 'HC1')

stargazer(warm_II, type="text",
          dep.var.labels = "opinion on global warming",
          covariate.labels = "departure from normal local
                              temperature (F) in week prior to survey",
          omit = c("doi", "statenum", "wbnid_num"),
          add.lines = list(c("fixed effects included",
                          "yes")),
          se = list(sqrt(diag(vp_II))))
```

```
=====================================================================
                                              Dependent variable:
                                          ---------------------------
                                           opinion on global warming
---------------------------------------------------------------------
temperature (F) in week prior to survey          0.013***
                                                  (0.005)
```

```
-------------------------------------------------------------------
fixed effects included                                    yes
Observations                                             6,726
===================================================================
Note:                                   *p<0.1; **p<0.05; ***p<0.01
```

```
# create a new binary variable
warming <- warming %>% mutate(getwarm01 = as.numeric(getwarmord == 3),
                              educf = as.factor(educ))
```

(b) Probit

```
library(clubSandwich)
warming_01 <- glm(data = warming, getwarm01 ~ ddt_week + educf +
                    ddt_week:educf, family = binomial(link = 'probit'))
vp_01 <- vcovCR(warming_01, cluster = warming$statenum, type = 'CR1')

stargazer(warming_01, type="text",
          dep.var.labels = "opinion on global warming",
          covariate.labels = c("departure from normal local temperature (F)",
          "some college", "college", "post-grad",
          "temp departure:some college", "temp departure:college",
          "temp departure:post-grad", "intercept"),
          se = list(sqrt(diag(vp_01))))
```

```
=======================================================================
                                                 Dependent variable:
                                             ---------------------------
                                              opinion on global warming
-----------------------------------------------------------------------
departure from normal local temperature (F)           0.019***
                                                      (0.006)

some college                                          -0.037
                                                      (0.054)

college                                               -0.007
                                                      (0.069)

post-grad                                             0.109*
                                                      (0.066)

temp departure:some college                           -0.010
                                                      (0.008)

temp departure:college                               -0.018**
                                                      (0.009)

temp departure:post-grad                             -0.016**
                                                      (0.008)

intercept                                            0.587***
                                                      (0.039)
```

```
----------------------------------------------------------------------
Observations                                              6,716
Log Likelihood                                          -3,873.523
Akaike Inf. Crit.                                        7,763.047
======================================================================
Note:                                      *p<0.1; **p<0.05; ***p<0.01
```

(c.1) $Pr(\widehat{getwarm}01 = 1) = \Phi(0.587 + 0.019 \cdot (-5)) = 0.689$, for a person with a high-school education when `ddt_week`=-5.

```r
pnorm(0.587 + 0.019*(-5))
```

```
[1] 0.6886403
```

(c.1) $Pr(\widehat{getwarm}01 = 1) = \Phi(0.587 + 0.019 \cdot 5) = 0.752$, for a person with a high-school education when `ddt_week`=+5.

```r
pnorm(0.587 + 0.019*5)
```

```
[1] 0.7523805
```

(d.1) $Pr(\widehat{getwarm}01 = 1) = \Phi(0.587 + 0.019 \cdot (-5) + 0.109 - 0.016 \cdot (-5)) = 0.752$, for a person with a post-grad degree when `ddt_week`=-5.

```r
pnorm(0.587 + 0.019*(-5) + 0.109 - 0.016*(-5))
```

```
[1] 0.7520643
```

(d.2) $Pr(\widehat{getwarm}01 = 1) = \Phi(0.587 + 0.019 \cdot 5 + 0.109 - 0.016 \cdot 5) = 0.762$, for a person with a post-grad degree when `ddt_week`=+5.

```r
pnorm(0.587 + 0.019*5 + 0.109 - 0.016*5)
```

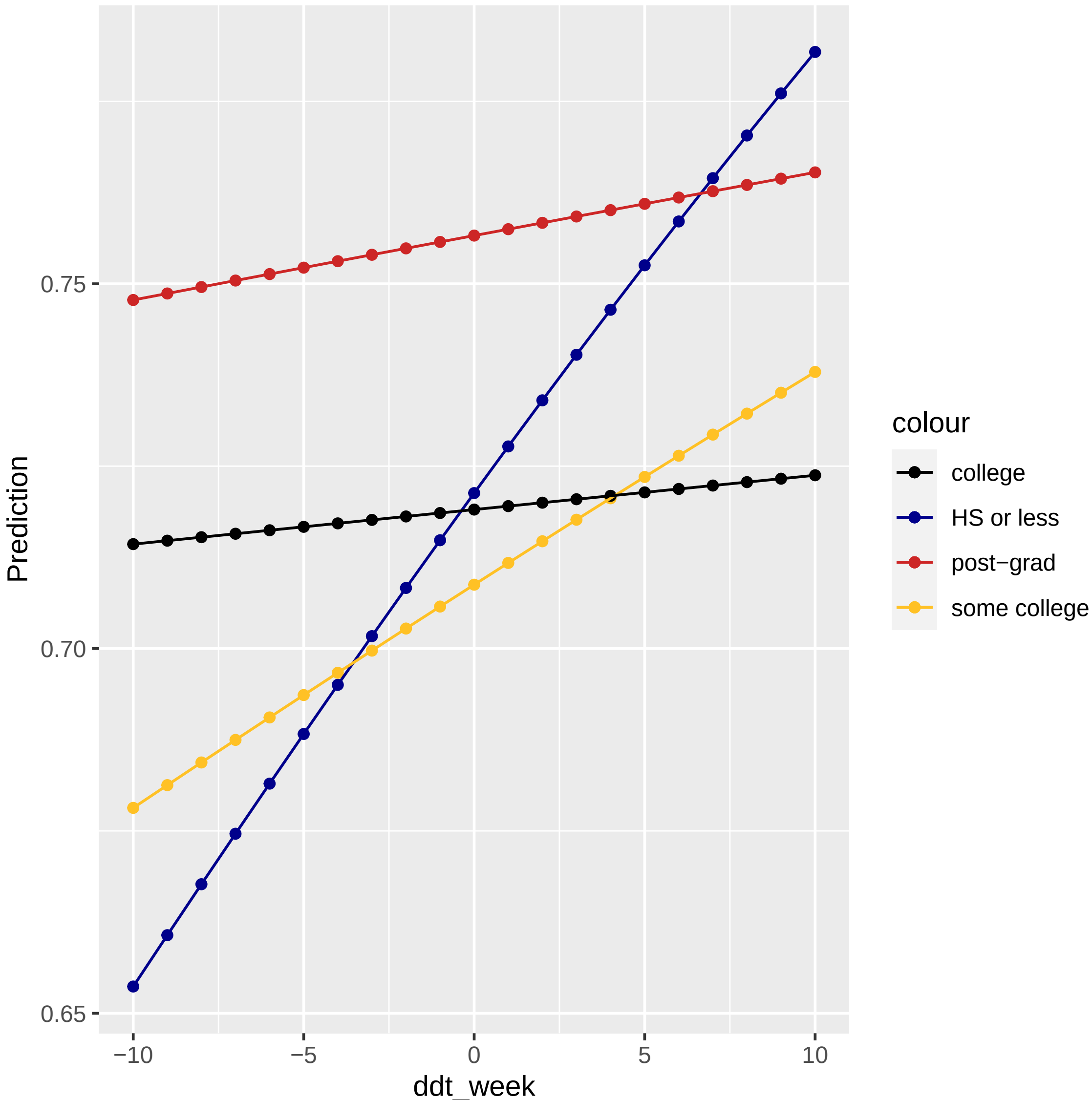
```
[1] 0.7614579
```

(e) The relative magnitude of the effect of weather on attitudes varies greater among respondents with high school education than among respondents with post-grad education, i.e. attitudes of respondents with high school education are more sensitive to changes in temperature in contrast to attitudes of those with post-grad degrees.

(f)

```r
library(margins)
grrr <- prediction(warming_01, data = warming,
                   at = list(educf = na.omit(unique(warming$educf)),
                                     ddt_week = c(-10:10)),
                   vcov = vp_01) %>% summary()

colnames(grrr)[1:2] <- c("educf", "ddt_week")

ggplot() +
  geom_point(data = grrr %>% filter(educf == 1),
             aes(x = ddt_week, y = Prediction, color = 'HS or less')) +
  geom_point(data = grrr %>% filter(educf == 2),
             aes(x = ddt_week, y = Prediction, color = 'some college')) +
  geom_point(data = grrr %>% filter(educf == 3),
            aes(x = ddt_week, y = Prediction, color = 'college')) +
  geom_point(data = grrr %>% filter(educf == 4),
            aes(x = ddt_week, y = Prediction, color = 'post-grad')) +
  geom_line(data = grrr %>% filter(educf == 1),
```

```
           aes(x = ddt_week, y = Prediction, color = 'HS or less')) +
geom_line(data = grrr %>% filter(educf == 2),
           aes(x = ddt_week, y = Prediction, color = 'some college')) +
geom_line(data = grrr %>% filter(educf == 3),
           aes(x = ddt_week, y = Prediction, color = 'college')) +
geom_line(data = grrr %>% filter(educf == 4),
           aes(x = ddt_week, y = Prediction, color = 'post-grad')) +
scale_colour_manual(values=c("black", "blue4",
                             "firebrick3", "goldenrod1"))
```



Question 4

(a) Causal graph: see Figure 2. Note that C includes other demographic variables as well.

(b) Goldstein and You's research question: does lobbying by local government (X) causes any difference in federal resource allocation (Y)?

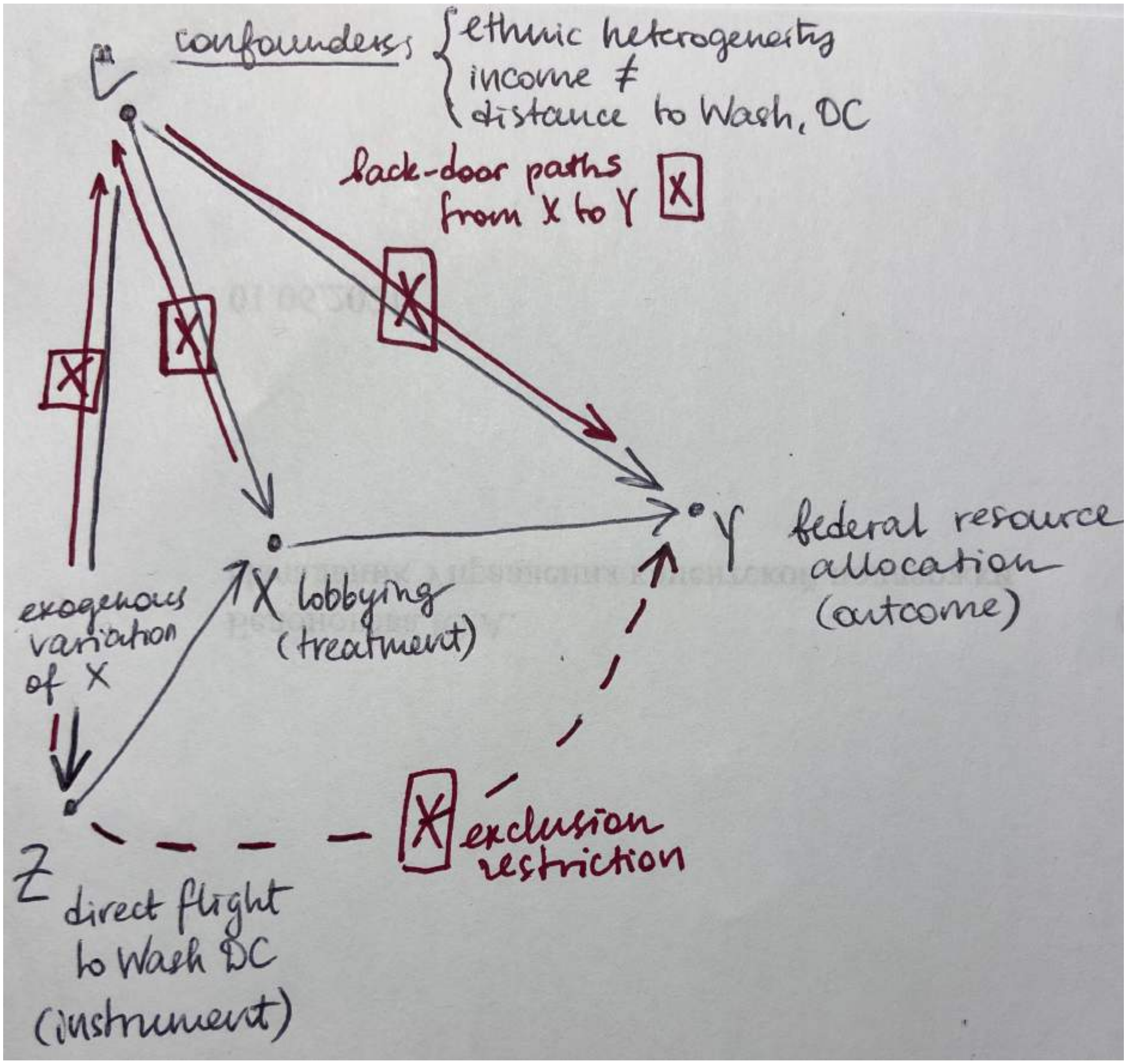


Figure 2: Causal Graph, Goldstein and You (2017)

- The main challenge to inference posed by the question concerns endogeneity: whether there are reverse causality and join determination between treatment and outcome. The research design addresses this issue by introducing a valid instrumental variable Z (direct flight from a relevant city to Washington, DC) that is correlated with the dependent variable but is not associated with the independent variable. The instrument thus guaranteed the exogenous variation of X. The instrument does not have any direct influence on the outcome Y, meeting the exclusion restriction. The authors also condition on observed confounders C to block the back-door path between the dependent and independent variables.
- The author's main finding from the instrumental variable regression is that a 10% increase in lobbying spending increases the amount of earmarks and Recovery Act grants (i.e. federal resource transfers to a city) by 10.2% and 4.7%, respectively. So the association between lobbying and federal resource transfers is positive, causal, large, and statistically significant.

(c) The instrument is valid when it meets (1) the **relevance** criterion (i.e. the instrument has a substantial impact on the level of X) and satisfies the (2) **exclusion restriction** (i.e. the instrument has no impact on Y except throught its effect on X). The instrument in Goldstein and You (2017) is relevant because the existence of a direct flight to Washington, DC, from a city is a strong and statistically significant predictor of this city's lobbying expenditures. The instrument in the paper also meets the exclusion restriction, as it is not associated with the city's previous years' lobbying spending or political affiliation of their federal representatives.

(d) Replication of the 2SLS estimation in Table 4, Model 4.

```r
cities <- read_dta("/Users/herrhellana/Dropbox/_NYU studies/Quant I/exam/cities.dta")

# first stage
fs <- lm(data = cities, ln_citylob ~ direct_flight_dc +
           diverge2_r + pop_r + land_r + water_r + senior_r +
           student_r + ethnic_r + mincome_r + unemp_r +
           poverty_r + gini_r + city_propertytaxshare_r +
           city_intgovrevenueshare_r + city_airexp_r + houdem_r +
           ln_countylob + as.factor(state2))

# second stage
d <- cities %>% mutate(xhat = predict(fs, newdata = cities))

ss <- lm(data = d, ln_recovery ~ xhat +
           diverge2_r + pop_r + land_r + water_r + senior_r +
           student_r + ethnic_r + mincome_r + unemp_r +
           poverty_r + gini_r + city_propertytaxshare_r +
           city_intgovrevenueshare_r + city_airexp_r + houdem_r +
           ln_countylob + as.factor(state2))

stargazer(fs, ss, type="text",
          digits = 2,
          dep.var.caption = "DV: (ln) recovery grant",
          dep.var.labels = c("first stage (lobbying)", "second stage"),
          covariate.labels = c("direct flight to DC",
                               "(ln) city lobbying spending"),
          omit.stat = c("rsq", "adj.rsq", "ser"),
          omit = c("diverge2_r", "pop_r", "land_r", "water_r",
                   "senior_r", "student_r", "ethnic_r", "mincome_r",
                   "unemp_r", "poverty_r", "gini_r", "city_propertytaxshare_r",
                   "city_intgovrevenueshare_r", "city_airexp_r",
                   "houdem_r", "ln_countylob", "state2", "Constant"),
          add.lines = list(c("state fixed effects", "yes", "yes"),
```

```
                        c("controls", "yes", "yes")))
```

```
================================================================
                                   DV: (ln) recovery grant
                             -----------------------------------
                             first stage (lobbying) second stage
                                      (1)                (2)
----------------------------------------------------------------
direct flight to DC                 2.67***
                                    (0.60)

(ln) city lobbying spending                           0.47***
                                                      (0.17)

----------------------------------------------------------------
state fixed effects                   yes                yes
controls                              yes                yes
Observations                         1,262              1,262
F Statistic (df = 66; 1195)         8.69***            6.31***
================================================================
Note:                              *p<0.1; **p<0.05; ***p<0.01
```

Question 5

(a) Causal graph: see Figure 3.

(b) The research question in Hall (2015) is whether the nomination of an extremist candidate for the general election (X) causes changes in general-election outcomes and legislative behavior in the U.S. House (Y).

- The research design is represented in Figure 3. It shows that the as-if random assignment of the extremist candidate guaranteed by the running variable Z (the extremist candidate's vote-share winning margin) blocks the back-door path from X to Y. The $f(Z)$ represents how Hall fits the regression line within the bandwidth. Hall also considers that the incumbency advantage C can confound the y-x relationship, but confounding is not a concern here because of the as-if random assignment.
- The main challenge to inference posed by the question is that meaningful covariates could be not smooth at the discontinuity (i.e. that districts where the relatively moderate primary candidate barely wins are in the limit comparable to those in which the relative moderate barely loses). Hall addresses this challenge of "sorting" by using ideology balance tests. Another threat for inference is when the RDD estimate is sensitive to the bandwidth's size. Hall addresses this issue by replicating the main analysis at a large variety of bandwidths and specifications. The final threat Hall address is the existence of multiple treatments around the cutoff by exploring the overall heterogeneity (all other differences) between the two types of candidates. The author also prevents the emergence of post-treatment bias by scaling the candidates on the basis of their primary-election campaign receipts instead of general-election contributions.
- The author finds that the "as-if" random nomination of the extremist candidate causes a substantial decrease in the party's vote share and probability of victory in the general election. When an extremist goes from barely losing the primary to barely winning it, the party's general-election vote share decreases noticeably. And these decreases are large enough to produce a reversal in observed roll-call voting for the district in the next Congress, i.e. when a more extreme Democrat is nominated, the district's roll-call voting in the next Congress becomes more conservative, and vice versa.

(c) The degree of polynomial is either $p = 1$ (local linear) or $p = 3$ (cubic), as shown in Table 2.

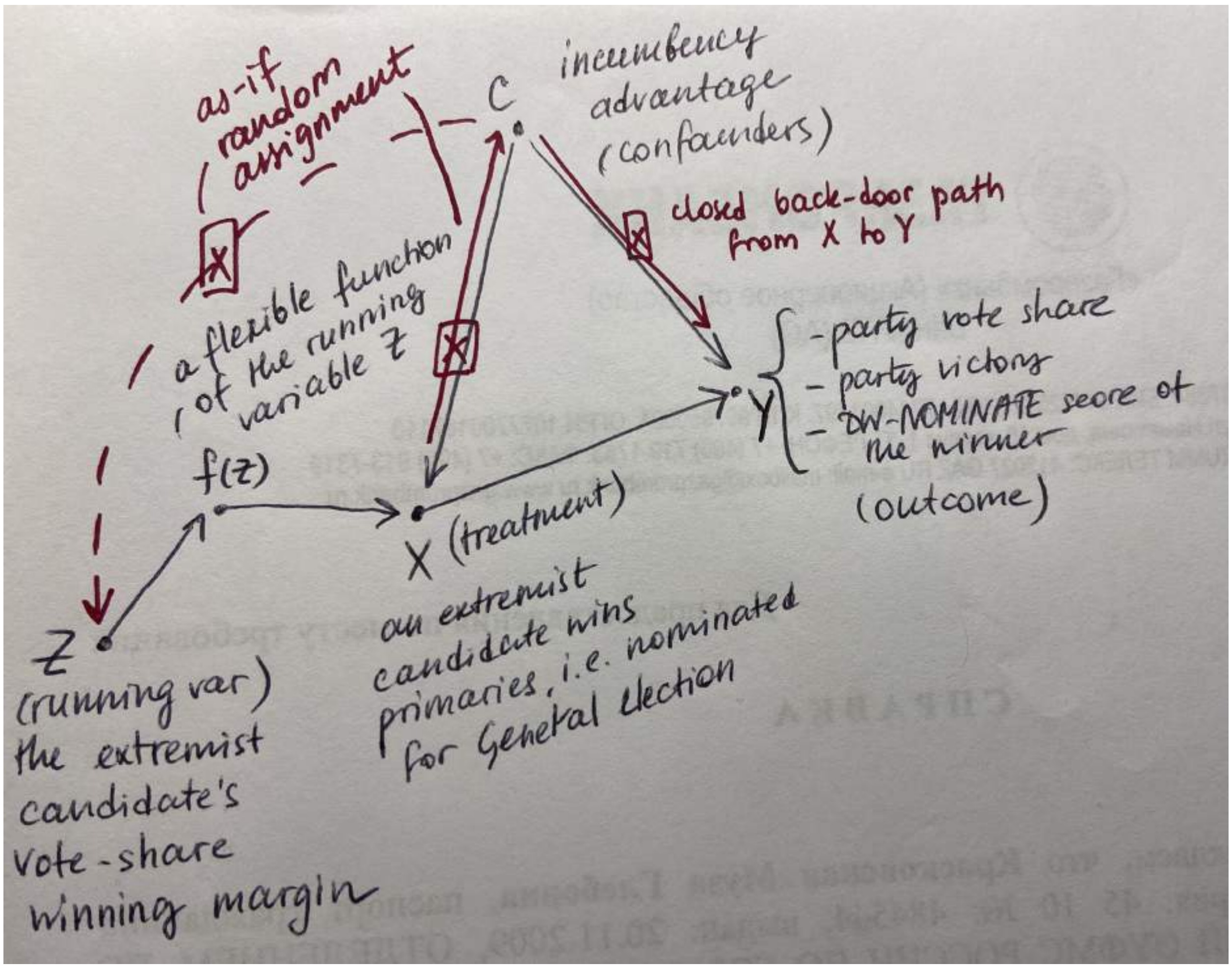


Figure 3: Causal Graph, Hall (2015)

(d) It is the difference between predicted outcomes at the cutoff for those who won or lost the coin-flip election, i.e. it is the segment on the Y-line between the intersections of two regression lines with the cutoff line (at $x = 0$).