AR_assignment_1

Alexander Staub

July 3, 2020

Question 1

What is the observed mean difference in the outcome variable Y between the presence and absence of the treatment? Show this mean difference using both a t-test as well as a regression specification. Corroborate, for example in Excel, that this observed difference equals the ratio of the covariance between Y and D over the variance of D.

Answer 1

First Load the data

```
library(readxl)

df_raw <- read_excel("C:/R work/applied econometrics/applied_econometrics/Data/Session1data.xlsx",
                     sheet = 1)
```

show the mean difference between presence and absence of treatment in the outcome variable

```
library(stats)
#regression specification
summary(lm(data = df_raw, y ~ D))[[4]]
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 0.76523179 0.02394299 31.960582 3.493854e-86
## D           0.08006954 0.04085345  1.959921 5.122324e-02
```

```
#T-test
t.test(df_raw$y~df_raw$D)
```

```
##
##  Welch Two Sample t-test
##
## data:  df_raw$y by df_raw$D
## t = -2.0047, df = 168.71, p-value = 0.04659
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.158916785 -0.001222296
## sample estimates:
## mean in group 0 mean in group 1
##       0.7652318       0.8453013
```

```
#ratio covariance to variance
cov(df_raw$y, y = df_raw$D)/var(df_raw$D)
```

```
## [1] 0.08006954
```

from the output above it is clear that the coefficient on the regression does in fact resemble the covariance of treatment and outcome divided by the variance of the treatment

Question 2

What is the difference in the outcome variable Y between the presence and absence of the treatment after controlling for observed firm characteristics (X1-X4)? i. Using a new regression, corroborate that the regression coefficient on D in the previous regression equals the ratio of the covariance between Y and that part of D that is orthogonal to X1-X4 over the variance of that part of D that is orthogonal to X1-X4. ii. What happens to the coefficients on X1-X4? iii. What happens to the regression coefficients when regressing Y on D and X1-X4 after making all independent variables orthogonal to each other? iv. At the end of the day, which regression specification should we rely on?

Answer 2

First I want to calculate the new regression specification

```r
#with treatment variable D
summary(lm(data = df_raw, y ~ x1 + x2 + x3 + x4 + D))
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + D, data = df_raw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73439 -0.12810 -0.01877  0.12033  0.68669
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.013326   0.121447   0.110 0.912727
## x1           0.085085   0.019559   4.350 2.07e-05 ***
## x2          -0.005275   0.020745  -0.254 0.799524
## x3           0.066706   0.017005   3.923 0.000116 ***
## x4           0.105799   0.018264   5.793 2.33e-08 ***
## D            0.069671   0.036538   1.907 0.057825 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2594 on 224 degrees of freedom
## Multiple R-squared:  0.2491, Adjusted R-squared:  0.2323
## F-statistic: 14.86 on 5 and 224 DF,  p-value: 1.371e-12
```

```r
#without treatment variable D
summary(lm(data = df_raw, y ~ x1 + x2 + x3 + x4 ))
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4, data = df_raw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.74989 -0.14034 -0.01333  0.11084  0.67145
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.012496   0.122155   0.102 0.918613    
## x1           0.082399   0.019622   4.199 3.86e-05 ***
## x2          -0.009183   0.020764  -0.442 0.658712    
## x3           0.065871   0.017099   3.852 0.000153 ***
## x4           0.109926   0.018241   6.026 6.79e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2609 on 225 degrees of freedom
## Multiple R-squared:  0.2369, Adjusted R-squared:  0.2233 
## F-statistic: 17.46 on 4 and 225 DF,  p-value: 1.715e-12
```

(i) In order to receive the part of d that is orthogonal to the other regressors, I need to regress the treatment D on the controls x1 - x4 and save the residuals as a new variable

```r
#regress D on the control variables
df_raw$D_orth <- residuals(lm(D~x1 + x2 + x3 + x4, data = df_raw))
#calculate the ratio of covariance between y and orthogonal d over the variance of orthogonal D
cov(df_raw$D_orth, y = df_raw$y)/var(df_raw$D_orth)
```

```
## [1] 0.06967092
```

as seen above, the regression coefficient of D using the new specification equals the ratio of covariance(y, D orthogonal to X's) and variance(D orthogonal to X's)

(ii) as seen in the output of the regression summary above, the coefficient on x1, x2 and x3 increases, while it decreases on x4. Standard errors on all 4 control variables hardly see any change and thus significance values of the coefficients don't change either

(iii) the process above will be repeated with all the x variables to make them orthogonal to eachother

```r
#create orthogonal variables
df_raw$x1_orth <- residuals(lm(x1~D + x2 + x3 + x4, data = df_raw))
df_raw$x2_orth <- residuals(lm(x2~D + x1 + x3 + x4, data = df_raw))
df_raw$x3_orth <- residuals(lm(x3~D + x2 + x1 + x4, data = df_raw))
df_raw$x4_orth <- residuals(lm(x4~D + x2 + x3 + x1, data = df_raw))
#run regression with orthogonal variables
summary(lm(data = df_raw, y~x1_orth + x2_orth + x3_orth + x4_orth + D_orth))
```

```
## 
## Call:
## lm(formula = y ~ x1_orth + x2_orth + x3_orth + x4_orth + D_orth, 
##     data = df_raw)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.73439 -0.12810 -0.01877  0.12033  0.68669 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.79273    0.01710  46.350  < 2e-16 ***
## x1_orth      0.03588    0.02138   1.678   0.0947 .  
## x2_orth      0.06324    0.02470   2.560   0.0111 *  
## x3_orth      0.08060    0.01716   4.698 4.58e-06 ***
## x4_orth      0.13934    0.02331   5.978 8.83e-09 ***
## D_orth       0.08240    0.03707   2.223   0.0272 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2594 on 224 degrees of freedom
## Multiple R-squared:  0.2491, Adjusted R-squared:  0.2323
## F-statistic: 14.86 on 5 and 224 DF,  p-value: 1.371e-12
```

after making each variable orthogonal to the remaining variables in the regression, any correlation among them is partialled out, making a multivariate regression unnecessary. All that changed is that the intercept has increased, making the effect size of the variables lower in the multivariate regression with orthogonal variables compared to non-orthogonal variable regression.

(iv) It would make more sense to use the first specification (i.e. not including orthogonal variables) rather than the second regression as in the second regression, interpretation of the coefficients is no longer straight forward. I.e. a change in the treatment variable from 0 to 1 gives the change in the outcome variable (keeping control variables constant), while the coefficient in the orthogonal treatment variable controlling for orthogonal control variables does not yield the same interpretation.

Question 3

What is the difference in the outcome variable Y between the presence and absence of the treatment after controlling for observed firm characteristics (X1-X4) as well as industry effects?

Answer 3

```r
#regression with treatment and industry effects
summary(lm(data = df_raw, y~ x1 + x2 + x3 + x4 + as.factor(industry) + D))[[4]]
```

```
##                          Estimate Std. Error     t value     Pr(>|t|)
## (Intercept)          -0.123675007 0.14200342 -0.87092977 3.847685e-01
## x1                    0.057022481 0.02768903  2.05938865 4.066744e-02
## x2                   -0.004625521 0.02085694 -0.22177374 8.247015e-01
## x3                    0.081515330 0.01780985  4.57698026 7.993262e-06
## x4                    0.116809483 0.02002426  5.83339732 1.987526e-08
## as.factor(industry)2 -0.003602800 0.10786957 -0.03339960 9.733871e-01
## as.factor(industry)3 -0.004964950 0.10511675 -0.04723271 9.623718e-01
## as.factor(industry)4  0.208037201 0.20139406  1.03298577 3.027765e-01
## as.factor(industry)5 -0.231467467 0.27055929 -0.85551477 3.932231e-01
## as.factor(industry)6  0.495574073 0.15858782  3.12491890 2.025069e-03
## as.factor(industry)7  0.025086661 0.10033378  0.25003205 8.028024e-01
## as.factor(industry)8  0.114452560 0.12201755  0.93800080 3.493014e-01
## as.factor(industry)9 -0.117132358 0.12550154 -0.93331414 3.517096e-01
## as.factor(industry)10  0.096914381 0.10009716  0.96820313 3.340359e-01
## as.factor(industry)11  0.081504602 0.11510945  0.70806179 4.796774e-01
## D                     0.094283549 0.03822337  2.46664646 1.442405e-02
```

the outcome variable changes by 0.094 after treatment with industry effects and observed firm characteristics, however remains negative. Including industry effects one can see that hardly any have a significant impact on the outcome variable (apart for industry 6)

Question 4

Given that the assumption of homoskedasticity is probably not valid and the data are not independent across firms because of industry affiliation, which standard errors should we rely on? The default OLS standard errors (“conventional”), robust standard errors, clustered standard errors, or some other standard errors? Using the regression specification of question 3, determine the significance of the coefficient on D using the standard error that you think is most appropriate and explain why you think it is most appropriate.

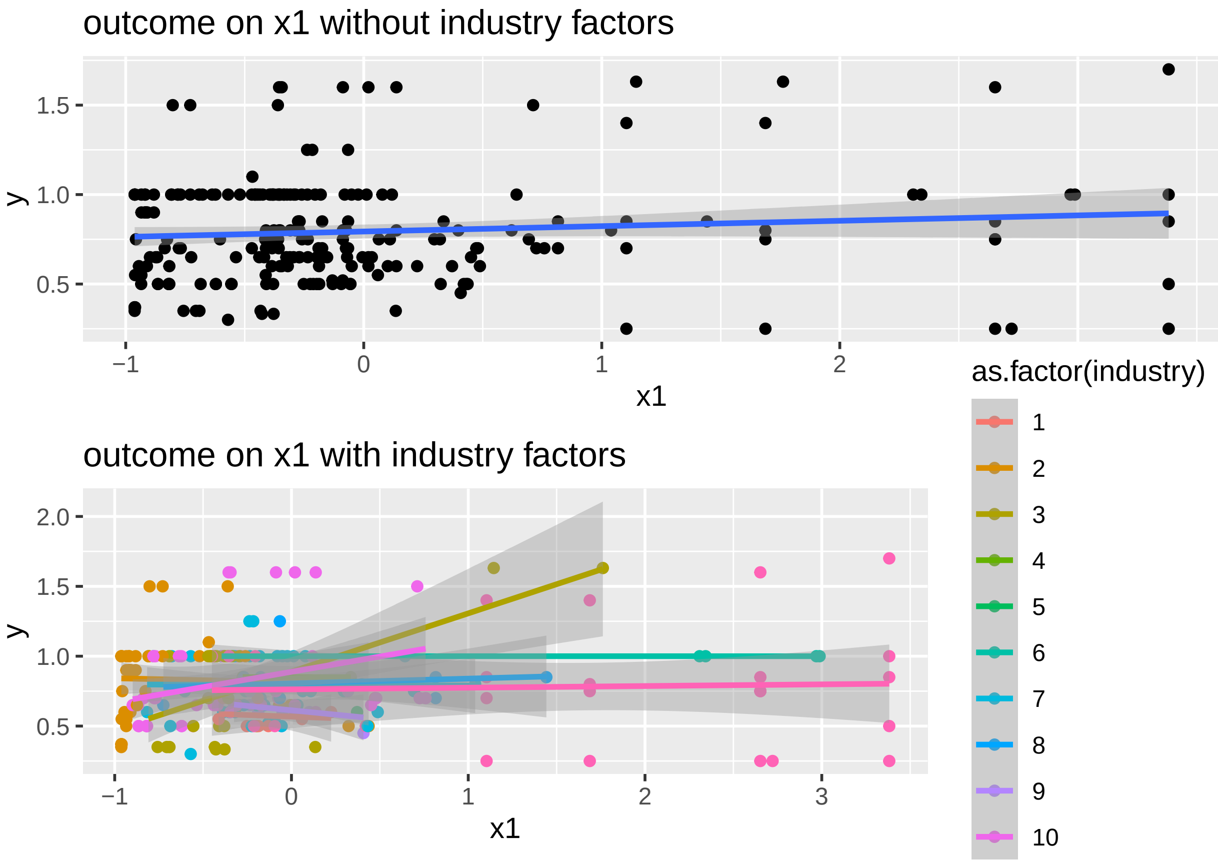
Answer 4

a brief visual investigation when plotting the outcome variable and the x1 variable with and without industry factors shows 2 things. - 1st, that variance increases as x1 increases (heteroskedasticity) - 2nd, that variance is differs for each industry cluster

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning in qt((1 - level)/2, df): NaNs produced
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning -
## Inf
```



below, the regression coefficients including robust standard errors and clustered standard errors on the industry level

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sandwich)
if (!require(miceadds)) install.packages("miceadds"); library(miceadds)
```

```
## Loading required package: miceadds

## Warning: package 'miceadds' was built under R version 4.0.2

## Loading required package: mice

## Warning: package 'mice' was built under R version 4.0.2

##
## Attaching package: 'mice'

## The following objects are masked from 'package:base':
##
##     cbind, rbind

## * miceadds 3.9-14 (2020-05-09 11:27:27)
```

```
#compute the regression specification from question 3 with robust standard errors
m_1 <- lm(data = df_raw, y~ x1 + x2 + x3 + x4 + as.factor(industry) + D)
coeftest(m_1, vcov = vcovHC(m_1, type = "HC1"))
```

```
##
## t test of coefficients:
##
##                         Estimate Std. Error t value  Pr(>|t|)
## (Intercept)           -0.1236750  0.1059529 -1.1673  0.244403
## x1                     0.0570225  0.0355992  1.6018  0.110677
## x2                    -0.0046255  0.0172529 -0.2681  0.788880
## x3                     0.0815153  0.0191646  4.2534 3.147e-05 ***
## x4                     0.1168095  0.0184551  6.3294 1.423e-09 ***
## as.factor(industry)2  -0.0036028  0.0698093 -0.0516  0.958888
## as.factor(industry)3  -0.0049649  0.0756815 -0.0656  0.947755
## as.factor(industry)4   0.2080372  0.0421921  4.9307 1.642e-06 ***
## as.factor(industry)5  -0.2314675  0.0489958 -4.7242 4.180e-06 ***
## as.factor(industry)6   0.4955741  0.0963736  5.1422 6.119e-07 ***
## as.factor(industry)7   0.0250867  0.0422408  0.5939  0.553208
## as.factor(industry)8   0.1144526  0.0465343  2.4595  0.014705 *
## as.factor(industry)9  -0.1171324  0.0595353 -1.9674  0.050424 .
## as.factor(industry)10  0.0969144  0.0503480  1.9249  0.055569 .
## as.factor(industry)11  0.0815046  0.0719388  1.1330  0.258494
## D                      0.0942835  0.0349300  2.6992  0.007506 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#compute with clustered standard errors
m_2 <- lm.cluster(data = df_raw, y~ x1 + x2 + x3 + x4 + as.factor(industry) + D, cluster = "industry")

summary(m_2)
```

```
## R^2= 0.31978
```

```
##
##                           Estimate Std. Error     t value     Pr(>|t|)
## (Intercept)           -0.123675007 0.07556404 -1.63669137 1.016950e-01
## x1                     0.057022481 0.04221261  1.35084008 1.767467e-01
## x2                    -0.004625521 0.03780530 -0.12235111 9.026210e-01
## x3                     0.081515330 0.02677004  3.04502099 2.326641e-03
## x4                     0.116809483 0.01392975  8.38561039 5.046321e-17
## as.factor(industry)2  -0.003602800 0.06070322 -0.05935106 9.526725e-01
## as.factor(industry)3  -0.004964950 0.02572151 -0.19302712 8.469377e-01
## as.factor(industry)4   0.208037201 0.05382941  3.86474945 1.112033e-04
## as.factor(industry)5  -0.231467467 0.05021439 -4.60958403 4.034754e-06
## as.factor(industry)6   0.495574073 0.09271754  5.34498743 9.042333e-08
## as.factor(industry)7   0.025086661 0.03665233  0.68444922 4.936915e-01
## as.factor(industry)8   0.114452560 0.04279233  2.67460451 7.481745e-03
## as.factor(industry)9  -0.117132358 0.02524581 -4.63967549 3.489567e-06
## as.factor(industry)10  0.096914381 0.03793282  2.55489518 1.062198e-02
## as.factor(industry)11  0.081504602 0.08554458  0.95277343 3.407049e-01
## D                      0.094283549 0.06021783  1.56570826 1.174169e-01
```

above results show that standard errors on the treatment variable double when including clustered standard errors vs robust standard errors. The former is the most plausible specification, given the firms industry affiliation related non-independence. dadsfasdf.