W271 - Homework 4

Lei Yang, Ron Cordell, Subhashini Raghunathan

Feb 25, 2016

The Data

The file athletics.RData contains a two-year panel of data on 59 universities. Some variables relate to admissions, while others related to atheletic performance. You will use this dataset to investigate whether athletic success causes more students to apply to a university.

This data was made available by Wooldridge, and collected by Patrick Tulloch, then an economics student at MSU. It may have been further modified to test your proficiency. Sources are as follows:

1. Peterson's Guide to Four Year Colleges*, 1994 and 1995 (24th and 25th editions). Princeton University Press. Princeton, NJ.
2. The Official 1995 College Basketball Records Book*, 1994, NCAA.
3. 1995 Information Please Sports Almanac (6th edition)*. Houghton Mifflin. New York, NY.

```
# load packages
library(car)
library(lmtest)
library(sandwich)
# set work dir, clear workspace, load data, show description
setwd("~/Desktop/W271Data")
rm(list=ls())
```

Question 1:

Examine and summarize the dataset. Note that the actual data is found in the data object, while descriptions can be found in the desc object. How many observations and variables are there?

```
load('athletics.Rdata')
desc
```

```
##    variable                          label
## 1      year                    1992 or 1993
## 2      apps         # applics for admission
## 3     top25  perc frsh class in 25 hs perc
## 4    ver500 perc frsh >= 500 on verbal SAT
## 5    mth500   perc frsh >= 500 on math SAT
## 6    stufac           student-faculty ratio
## 7      bowl    = 1 if bowl game in prev yr
## 8    btitle  = 1 if men's cnf chmps prv yr
## 9   finfour    = 1 if men's final 4 prv yr
## 10    lapps                      log(apps)
## 11   avg500              (ver500+mth500)/2
## 12   school             name of university
## 13    bball        =1 if btitle or finfour
```

```
summary(data)
```

```
##       year           apps           top25           ver500
##  Min.   :1992   Min.   : 3303   Min.   :36.00   Min.   :20.00
##  1st Qu.:1992   1st Qu.: 6897   1st Qu.:54.50   1st Qu.:36.00
##  Median :1992   Median : 8646   Median :65.00   Median :49.00
##  Mean   :1992   Mean   :10489   Mean   :68.56   Mean   :54.16
##  3rd Qu.:1993   3rd Qu.:13424   3rd Qu.:85.00   3rd Qu.:71.50
##  Max.   :1993   Max.   :23342   Max.   :97.00   Max.   :94.00
##                                 NA's   :25      NA's   :30
##      mth500         stufac           bowl            btitle
##  Min.   :39.0   Min.   : 7.00   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:62.0   1st Qu.:12.00   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :81.0   Median :16.00   Median :0.0000   Median :0.0000
##  Mean   :77.6   Mean   :15.07   Mean   :0.4655   Mean   :0.1207
##  3rd Qu.:93.0   3rd Qu.:18.00   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :99.0   Max.   :24.00   Max.   :1.0000   Max.   :1.0000
##  NA's   :30
##     finfour            lapps            avg500         school
##  Min.   :0.00000   Min.   : 8.103   Min.   :32.00   Length:116
##  1st Qu.:0.00000   1st Qu.: 8.839   1st Qu.:49.50   Class :character
##  Median :0.00000   Median : 9.065   Median :66.00   Mode  :character
##  Mean   :0.06034   Mean   : 9.147   Mean   :65.88
##  3rd Qu.:0.00000   3rd Qu.: 9.505   3rd Qu.:82.12
##  Max.   :1.00000   Max.   :10.058   Max.   :96.50
##                                     NA's   :30
##      bball             perf
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :1.0000
##  Mean   :0.1552   Mean   :0.6466
##  3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :3.0000
##
```

There are **116** observations and **14** variables in the data.

Examine the variables of key interest: apps represents the number of applications for admission. bowl, btitle, and finfour are indicators of athletic success. The three athletic performance variables are all lagged by one year. Intuitively, this is because we expect a school's athletic success in the previous year to affect how many applications it recieves in the current year.

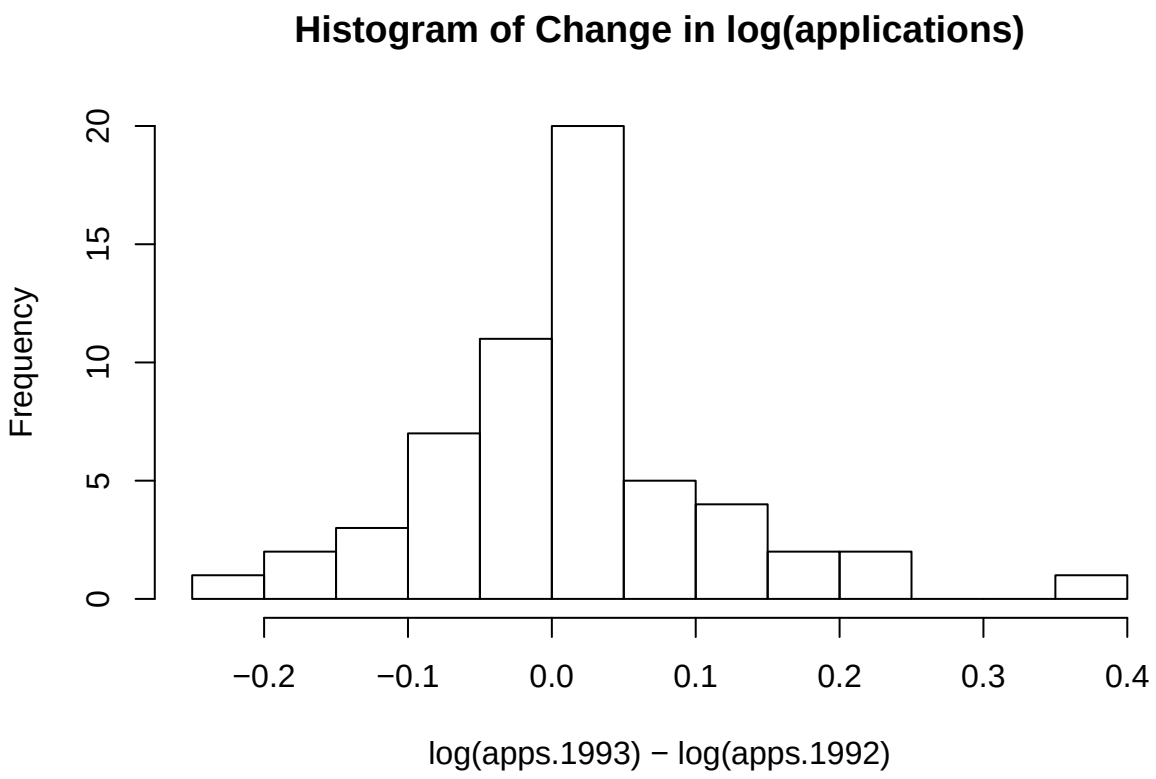
Question 2:

Note that the dataset is in long format, with a separate row for each year for each school. To prepare for a difference-in-difference analysis, transfer the dataset to wide-format. Each school should have a single row of data, with separate variables for 1992 and 1993. For example, you should have an apps.1992 variable and an apps.1993 variable to record the number of applications in either year.

```
# create wideData variable for wide format on year
wideData <- reshape(data, timevar="year", idvar=c("school"), direction="wide")
```

Create a new variable, clapps to represent the change in the log of the number of applications from 1992 to 1993. Examine this variable and its distribution. Which schools had the greatest increase and the greatest decrease in number of log applications?

```
# create clapps variable
wideData$clapps<-log(wideData$apps.1993) - log(wideData$apps.1992)
# plot histogram
hist(wideData$clapps, breaks=20, main="Histogram of Change in log(applications)",
     xlab="log(apps.1993) - log(apps.1992)")
```




```
# greatest increase
cmax <- max(wideData$clapps)
inc <- wideData[cmax==wideData$clapps, c("school")]
# greatest decrease
cmin <- min(wideData$clapps)
dec <- wideData[cmin==wideData$clapps, c("school")]
```

For the number of applications, **arizona** has the greatest increase of `0.40`, and **arkansas** has the greatest decrease of `-0.22`.

Question 3:

Similarly to above, create three variables, cperf, cbball, and cbowl to represent the changes in the three athletic success variables. Since these variables are lagged by one year, you are actually computing the change in athletic success from 1991 to 1992.

Which of these variables has the highest variance?

```r
#create cperf, cbball, and cbowl
wideData$cperf <- wideData$perf.1993 - wideData$perf.1992
wideData$cbball <- wideData$bball.1993 - wideData$bball.1992
wideData$cbowl <- wideData$bowl.1993 - wideData$bowl.1992
wideData$cbtitle <- wideData$btitle.1993 - wideData$btitle.1992
wideData$cfinfour <- wideData$finfour.1993 - wideData$finfour.1992
# check variance
var(wideData[c("cperf", "cbball", "cbowl")])
```

```
##            cperf     cbball      cbowl
## cperf  0.8242589 0.28009679 0.42105263
## cbball 0.2800968 0.17422868 0.07017544
## cbowl  0.4210526 0.07017544 0.31578947
```

cperf has the highest variance of 0.8243.

Question 4:

We are interested in a population model,

$$lapps_i = \gamma_0 + \beta_0 I_{1993} + \beta_1 bowl_i + \beta_2 btitle_i + \beta_3 finfour_i + a_i + u_{it}$$

Here, I_{1993} is an indicator variable for the year 1993. a_i is the time-constant effect of school i. u_{it} is the idiosyncratic effect of school i at time t. The athletic success indicators are all lagged by one year as discussed above.

At this point, we assume that (1) all data points are independent random draws from this population model (2) there is no perfect multicollinearity (3) $E(a_i) = E(u_{it}) = 0$

You will estimate the first-difference equation,

$$clapps_i = \beta_0 + \beta_1 cbowl_i + \beta_2 cbtitle_i + \beta_3 finfour_i + a_i + cu_i$$

where $cu_i = u_{i1993} - u_{i1992}$ is the change in the idiosyncratic term from 1992 to 1993.

a - What additional assumption is needed for this population model to be causal? Write this in mathematical notation and also explain it intuitively in English.

b - What additional assumption is needed for OLS to consistently estimate the first-difference model? Write this in mathematical notation and also explain it intuitively in English. Comment on whether this assumption is plausible in this setting.

$$E(u_{it} \mid cbowl, cbtitle, cfinfour) = 0$$

Question 5:

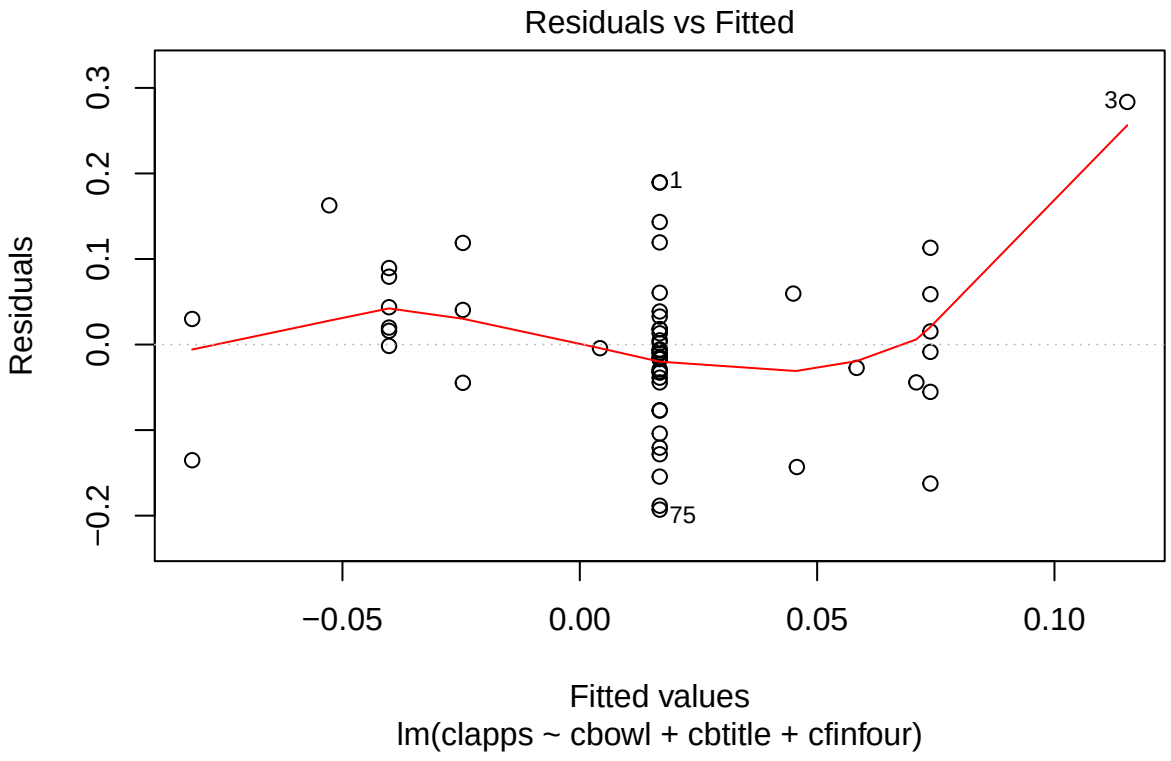
Estimate the first-difference model given above. Using the best practices descibed in class, interpret the slope coefficients and comment on their statistical significance and practical significance.

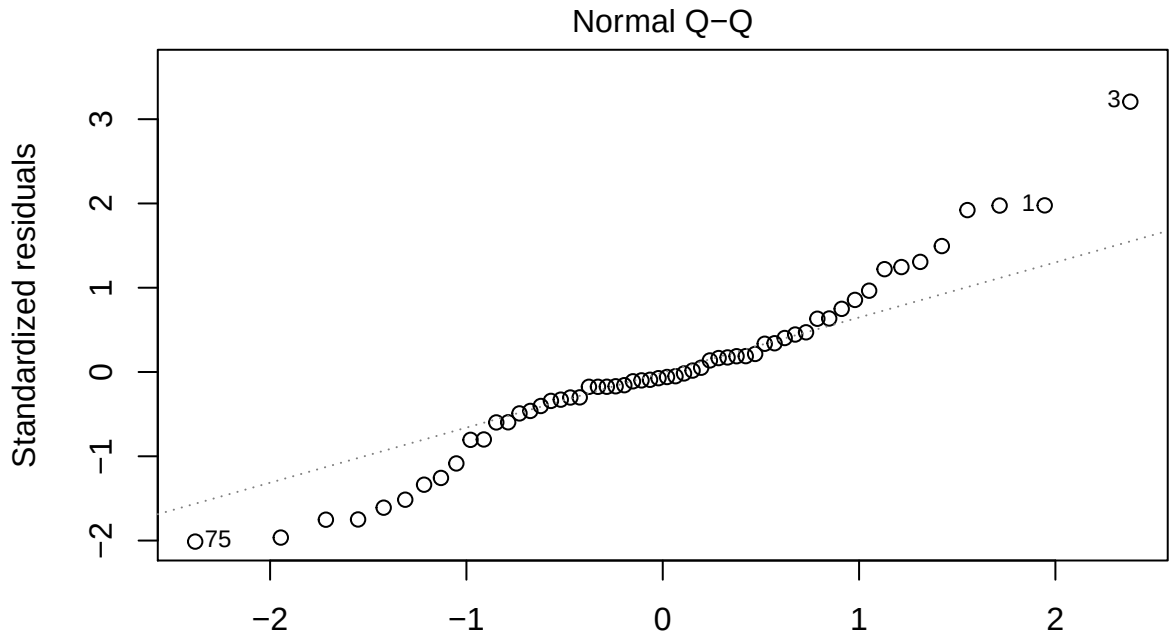
```r
# fit first-difference equation
m5 <- lm(clapps ~ cbowl+cbtitle+cfinfour, data=wideData)
summary(m5)
```

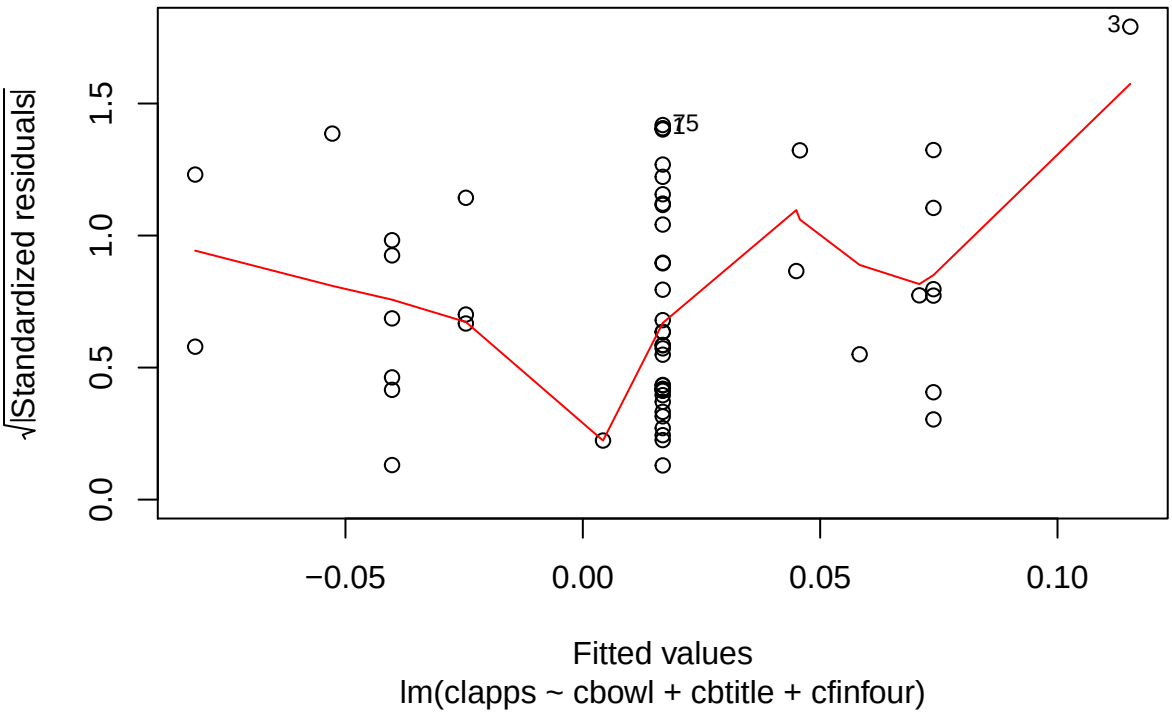
```
##
## Call:
## lm(formula = clapps ~ cbowl + cbtitle + cfinfour, data = wideData)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.192965 -0.042868 -0.006367  0.040005  0.283577
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.01684    0.01278   1.318   0.1932
## cbowl        0.05702    0.02448   2.329   0.0236 *
## cbtitle      0.04148    0.03161   1.312   0.1950
## cfinfour    -0.06961    0.04585  -1.518   0.1348
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09674 on 54 degrees of freedom
## Multiple R-squared:  0.1428, Adjusted R-squared:  0.09513
## F-statistic: 2.998 on 3 and 54 DF,  p-value: 0.03855
```

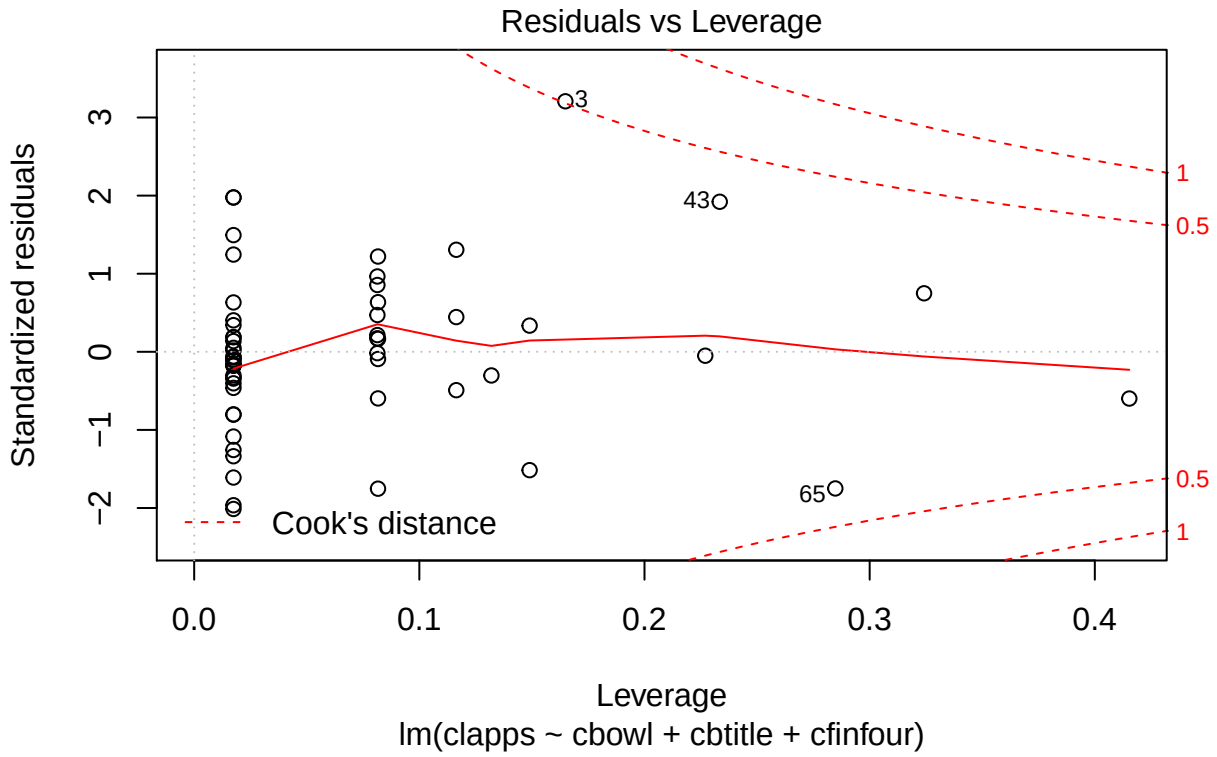
From the coefficients, the improvement in bowl game has statistically significant impact on application number, which will increase **5.70%** if a scholl has played in a bowl game the previous year while hasn't in the year before. Earning a men's basketball conference title can also increase application by **4.15%**, but the effect is not statistically significant. Basketball is not as popular/influential as football? Finally the model suggests that being in Final Four would actually decrease the application by **-6.96%**, this is clearly unreasonable and further model diagnostic is needed to evaluate the validity of the model.

```
# model diagnostic
plot(m5)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(clapps ~ cbowl + cbtitle + cfinfour)


Scale−Location
√|Standardized residuals|
Fitted values
lm(clapps ~ cbowl + cbtitle + cfinfour)



it seems the model has violation in both zero-conditional mean and homoskedasticity. In addition, school #3 (`arizona state`) has the number of an outlier.

Question 6:

Test the joint signifance of the three indicator variables. This is the test of the overall model. What impact does the result have on your conclusions?