W271-2 – Spring 2016 – **HW 4**

Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song

February 24, 2016

Contents

Data **1**

Exercises **2**

Question 1 2

Question 2 5

Question 3 7

Question 4 8

Question 5 9

Question 6 9

Data

The file `athletics.RData` contains a two-year panel of data on 59 universities. Some variables relate to admissions, while others are 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:

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

Exercises

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?

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.

```
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
```

```
##                     apps   bowl btitle finfour   lapps
## nbr.val           116.00 116.00 116.00  116.00  116.00
## nbr.null            0.00  62.00 102.00  109.00    0.00
## nbr.na              0.00   0.00   0.00    0.00    0.00
## min              3303.00   0.00   0.00    0.00    8.10
## max             23342.00   1.00   1.00    1.00   10.06
## range           20039.00   1.00   1.00    1.00    1.96
## sum           1216779.00  54.00  14.00    7.00 1061.08
## median           8646.00   0.00   0.00    0.00    9.06
## mean            10489.47   0.47   0.12    0.06    9.15
## SE.mean           461.08   0.05   0.03    0.02    0.04
## CI.mean.0.95      913.32   0.09   0.06    0.04    0.09
## var          24661234.74   0.25   0.11    0.06    0.23
## std.dev          4966.01   0.50   0.33    0.24    0.48
## coef.var            0.47   1.08   2.71    3.96    0.05
```

The dataset consists of 116 observations of variables measuring athletic success and the number of applications to 58 universities. For each university, there are 2 observations, one from 1992 and one from 1993.

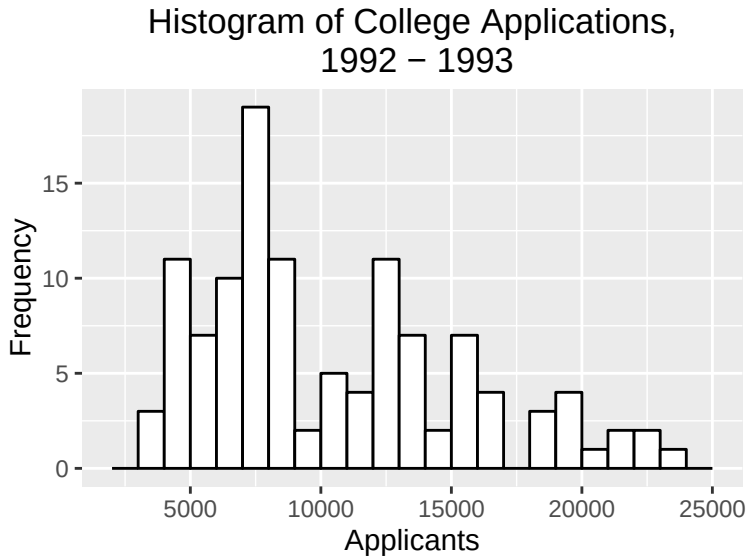

Figure 1: Histogram of College Applications, 1992 - 1993

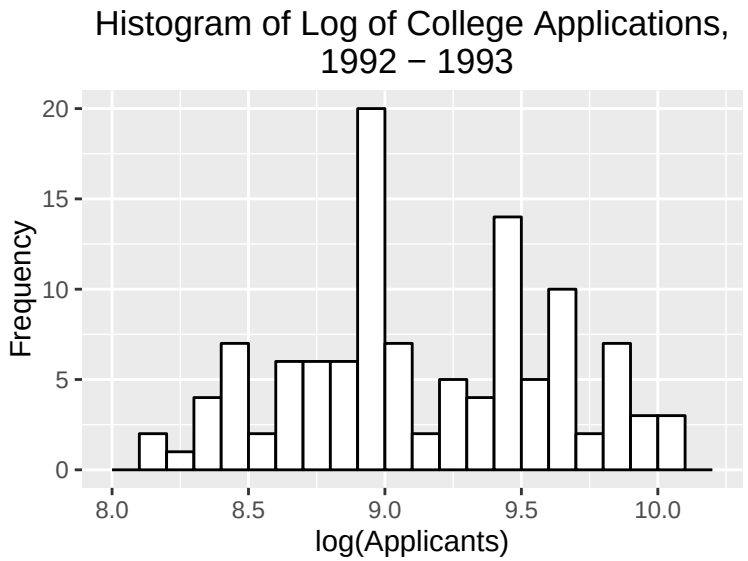

Figure 2: Histogram of Logged College Applications, 1992 - 1993

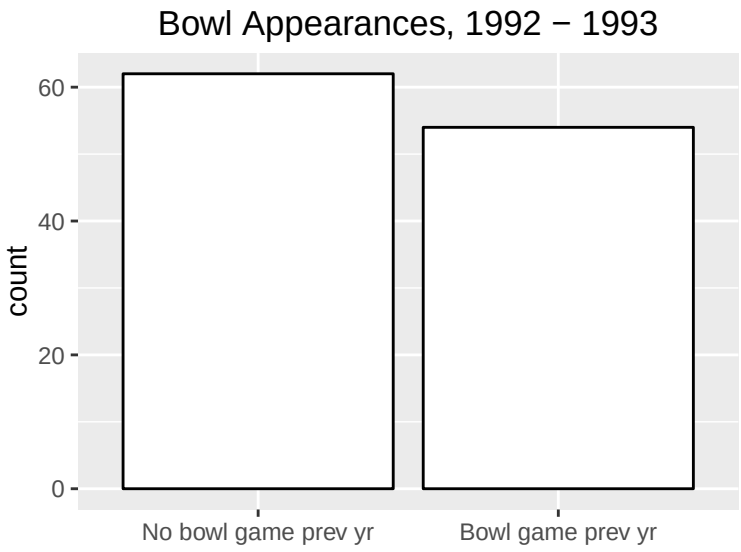

Figure 3: Histogram of Bowl Appearances, 1992 - 1993

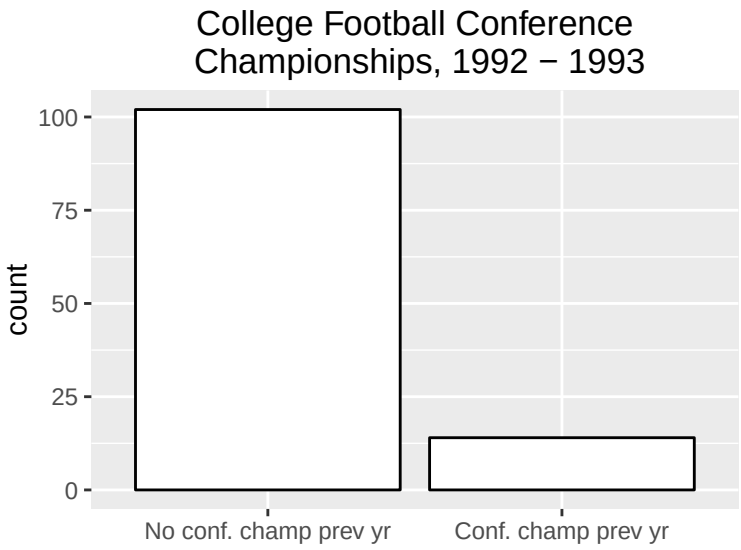

Figure 4: Histogram of Football Conference Championships, 1992 - 1993

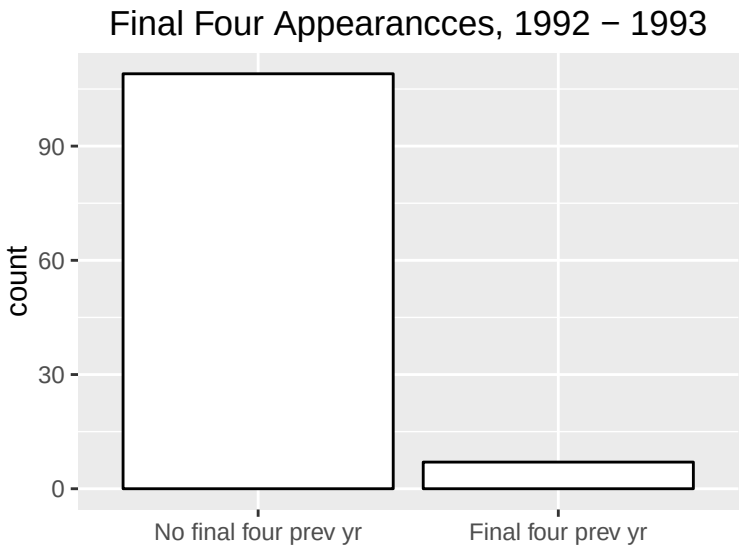

Figure 5: Histogram of Final Four Appearances, 1992 - 1993

The applicants variable is right skewed and possibly multi-model. The bowl appearance factor variable has a roughly equal split, while the majority of conference championship and final four variable observations are negative.

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 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?

```
#Transer dataset to wide-format
wide <-  reshape(data,
                 v.names=c("apps", "top25", "ver500", "mth500", "stufac",
                           "bowl", "btitle", "finfour", "lapps", "avg500", "bball", "perf"),
                 idvar = "school", timevar = "year", direction = "wide")
```

```
wide$clapps <- wide$lapps.1993 - wide$lapps.1992
```

```
##                apps.1992   apps.1993 lapps.1992 lapps.1993 clapps
## nbr.val            58.00       58.00      58.00      58.00  58.00
## nbr.null            0.00        0.00       0.00       0.00   3.00
## nbr.na              0.00        0.00       0.00       0.00   0.00
```

```
## min              3516.00     3303.00       8.17       8.10  -0.22
## max             23342.00    22165.00      10.06      10.01   0.40
## range           19826.00    18862.00       1.89       1.90   0.62
## sum            604302.00   612477.00     530.13     530.95   0.82
## median           8641.50     8669.00       9.06       9.07   0.01
## mean            10419.00    10559.95       9.14       9.15   0.01
## SE.mean           659.81      649.86       0.06       0.06   0.01
## CI.mean.0.95     1321.25     1301.32       0.13       0.13   0.03
## var          25250560.49 24494454.75       0.23       0.23   0.01
## std.dev          5024.99     4949.19       0.48       0.48   0.10
## coef.var            0.48        0.47       0.05       0.05   7.16
##              bowl.1992 bowl.1993 btitle.1992 btitle.1993 finfour.1992
## nbr.val          58.00     58.00       58.00       58.00        58.00
## nbr.null         31.00     31.00       50.00       52.00        55.00
## nbr.na            0.00      0.00        0.00        0.00         0.00
## min               0.00      0.00        0.00        0.00         0.00
## max               1.00      1.00        1.00        1.00         1.00
## range             1.00      1.00        1.00        1.00         1.00
## sum              27.00     27.00        8.00        6.00         3.00
## median            0.00      0.00        0.00        0.00         0.00
## mean              0.47      0.47        0.14        0.10         0.05
## SE.mean           0.07      0.07        0.05        0.04         0.03
## CI.mean.0.95      0.13      0.13        0.09        0.08         0.06
## var               0.25      0.25        0.12        0.09         0.05
## std.dev           0.50      0.50        0.35        0.31         0.22
## coef.var          1.08      1.08        2.52        2.97         4.32
##              finfour.1993
## nbr.val             58.00
## nbr.null            54.00
## nbr.na               0.00
## min                  0.00
## max                  1.00
## range                1.00
## sum                  4.00
## median               0.00
## mean                 0.07
## SE.mean              0.03
## CI.mean.0.95         0.07
## var                  0.07
## std.dev              0.26
## coef.var             3.71
```

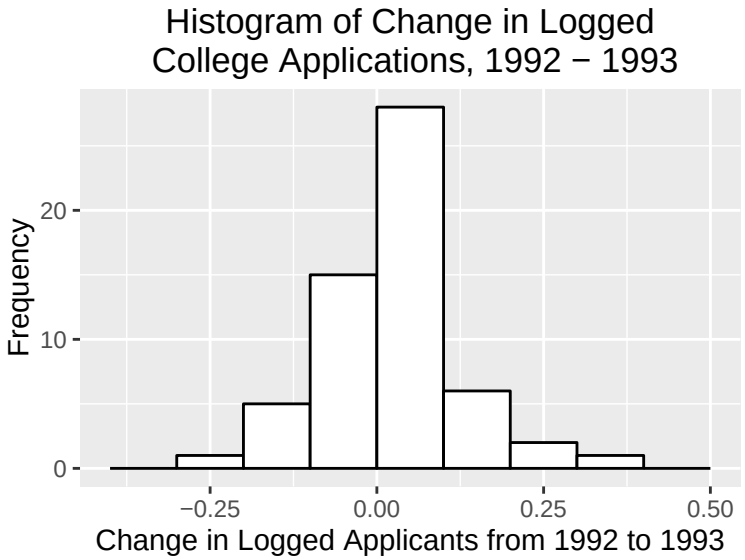

Figure 6: Histogram of the change in logged applications from 1992 to 1993

```
wide[wide$clapps == max(wide$clapps), 'school']
```

```
## [1] "arizona"
```

```
wide[wide$clapps == min(wide$clapps), 'school']
```

```
## [1] "arkansas"
```

The University of Arizona had the highest increase in logged applications and the University of Arkansas had the highest decrease in logged applications.

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?

```
wide$cfinfour <-  wide$finfour.1993 - wide$finfour.1992
wide$cbtitle <- wide$btitle.1993 - wide$btitle.1992
wide$cbowl <- wide$bowl.1993 - wide$bowl.1992
c(var(wide$cfinfour), var(wide$cbtitle), var(wide$cbowl))
```

```
## [1] 0.08741682 0.17422868 0.31578947
```

The change in bowl appearance, cbowl, has the highest variance among the athletic performance variables.

Question 4

We are interested in a population model,

$$\mathbf{lapps_i} = \delta_0 + \beta_0 \mathbf{I_{1993}} + \beta_1 \mathbf{bowl_i} + \beta_2 \mathbf{btitle_i} + \beta_3 \mathbf{finfour_i} + \mathbf{a_i} + \mathbf{u_{it}}$$

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,

$$\mathbf{clapps_i} = \beta_0 + \beta_1 \mathbf{cbowl} + \mathbf{i} + \beta_2 \mathbf{cbtitle_i} + \beta_3 \mathbf{cfinfour_i} + \mathbf{a_i} + \mathbf{cu_i}$$

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.**

We need to assume that the error term has the same variance given any values of the explanatory variables. This can be stated mathematically as:

$$Var(u|x_1, \ldots, x_k) = \sigma^2$$

Intuitively, this assumption means that the variance of the error does not depend on the values of x. We can examine this assumption by looking at the fitted vs. residuals plot and noting if the spread of the residuals tends to be the same over the entire range of the dataset.

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.**

To estimate the first-difference model we must assume 'parallel trends'. Take an example with period A(after) and B(before), indicator variable D = 1 if treated, and potential outcome Y^0

$$E(Y_A^0 - Y_B^0|D = 1) - E(Y_A^0 - Y_B^0|D = 0) = 0$$

Intuitively, this assumption says that if we consider the counter-factual where the treated group was not treated, then the change in the explanatory variable would have been the same for both groups. By comparing the trend for the treated group with a similar untreated control group, we can argue in favor of the parallel trends assumption, although it is not formally testable since the hypothetical untreated outcome for treated units is not observed.

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.

Table 1: Regression summary

	Dependent variable:
	clapps
cfinfour	−0.0696
	(0.0668)
cbtitle	0.0415
	(0.0443)
cbowl	0.0570*
	(0.0272)
Constant	0.0168
	(0.0140)
F Statistic	1.472
df	3; 54
Observations	58
R^2	0.1428
Adjusted R^2	0.0951
Residual Std. Error	0.0967

·p<0.1; *p<0.05; **p<0.01; ***p<0.001

The change in bowl appearances variable was associated with a statistically significant increase in change logged college applications, (β = 0.06, t = 2.1, p = 0.04). The coefficient for change in final four appearances was -0.07 and was not statistically significant (β = -0.07, t = -1.04). The coefficient for change in football conference titles was 0.04 and was not statistically significant (β = 0.04, t = 0.94).

To illustrate the magnitude of the effect for the change in bowl appearances variable, a college with average logged college applicants in 1992 would have an expected increase of 6377 applicants due to the effect of appearing in a bowl.

Both the change in final four appearances and change in college football conference titles variables suffer from the fact that there are very few observations in the yes group. This makes the assumption of parallel trends less plausible because the individual factors effecting this small group of colleges probably outweigh the national and regional factors effecting all colleges. If there was relative equality among the group sizes, we would expect these factors to be overwhelmed by national trends, but that is not the case here.

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?

The overall significance test for the model was statistically significant (F(3, 54) = 3.00, p = 0.04). This suggests that there is some effect of athletic performance on college enrollment. However, the issues with the group sizes in some variables, as well as the fact that this study was conducted over a two year period, with a relatively small sample of schools, means that this model is far from conclusive. Further investigation over a longer period of time would be needed to achieve a more conclusive result.
