W271-2 – Spring 2016 – **HW 4**

Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song

February 24, 2016

Contents

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:

- **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
```

```
str(data)
```

```
## 'data.frame':    116 obs. of  14 variables:
##  $ year   : int  1992 1993 1992 1993 1992 1993 1992 1993 1992 1993 ...
##  $ apps   : int  6245 7677 13327 19860 10422 12809 4103 3303 8661 7548 ...
##  $ top25  : int  49 58 57 57 37 49 60 67 54 54 ...
##  $ ver500 : int  NA NA 36 36 28 31 NA NA 46 51 ...
##  $ mth500 : int  NA NA 58 58 58 62 NA NA 86 83 ...
##  $ stufac : int  20 15 16 16 20 14 16 18 16 16 ...
##  $ bowl   : int  1 1 0 1 0 0 1 0 0 0 ...
##  $ btitle : int  0 0 0 1 0 0 1 0 0 0 ...
##  $ finfour: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ lapps  : num  8.74 8.95 9.5 9.9 9.25 ...
##  $ avg500 : num  NA NA 47 47 43 46.5 NA NA 66 67 ...
##  $ school : chr  "alabama" "alabama" "arizona" "arizona" ...
##  $ bball  : int  0 0 0 1 0 0 1 0 0 0 ...
##  $ perf   : int  1 1 0 2 0 0 2 0 0 0 ...
```

```
head(data)
```

```
##   year  apps top25 ver500 mth500 stufac bowl btitle finfour    lapps
## 1 1992  6245    49     NA     NA     20    1      0       0 8.739536
## 2 1993  7677    58     NA     NA     15    1      0       0 8.945984
## 3 1992 13327    57     36     58     16    0      0       0 9.497547
## 4 1993 19860    57     36     58     16    1      1       0 9.896463
## 5 1992 10422    37     28     58     20    0      0       0 9.251675
## 6 1993 12809    49     31     62     14    0      0       0 9.457903
##   avg500        school bball perf
## 1     NA       alabama     0    1
## 2     NA       alabama     0    1
## 3   47.0       arizona     0    0
## 4   47.0       arizona     1    2
## 5   43.0 arizona state     0    0
## 6   46.5 arizona state     0    0
```

```
# summary(data)
# Omit character vectors (only column 'school') in Descriptive Statistics
round(stat.desc(data[, lapply(data, class) != "character"], desc = TRUE,
                basic = TRUE), 2)
```

```
##                    year        apps   top25  ver500  mth500  stufac   bowl
## nbr.val          116.00      116.00   91.00   86.00   86.00  116.00 116.00
## nbr.null           0.00        0.00    0.00    0.00    0.00    0.00  62.00
## nbr.na             0.00        0.00   25.00   30.00   30.00    0.00   0.00
## min             1992.00     3303.00   36.00   20.00   39.00    7.00   0.00
## max             1993.00    23342.00   97.00   94.00   99.00   24.00   1.00
## range              1.00    20039.00   61.00   74.00   60.00   17.00   1.00
## sum           231130.00  1216779.00 6239.00 4658.00 6674.00 1748.00  54.00
## median          1992.50     8646.00   65.00   49.00   81.00   16.00   0.00
## mean            1992.50    10489.47   68.56   54.16   77.60   15.07   0.47
## SE.mean            0.05      461.08    1.83    2.33    1.77    0.37   0.05
## CI.mean.0.95       0.09      913.32    3.64    4.64    3.52    0.73   0.09
## var                0.25 24661234.74  305.49  468.44  268.81   15.58   0.25
## std.dev            0.50     4966.01   17.48   21.64   16.40    3.95   0.50
## coef.var           0.00        0.47    0.25    0.40    0.21    0.26   1.08
##               btitle finfour   lapps  avg500  bball   perf
## nbr.val       116.00  116.00  116.00   86.00 116.00 116.00
## nbr.null      102.00  109.00    0.00    0.00  98.00  53.00
## nbr.na          0.00    0.00    0.00   30.00   0.00   0.00
## min             0.00    0.00    8.10   32.00   0.00   0.00
## max             1.00    1.00   10.06   96.50   1.00   3.00
## range           1.00    1.00    1.96   64.50   1.00   3.00
## sum            14.00    7.00 1061.08 5666.00  18.00  75.00
## median          0.00    0.00    9.06   66.00   0.00   1.00
## mean            0.12    0.06    9.15   65.88   0.16   0.65
## SE.mean         0.03    0.02    0.04    2.01   0.03   0.06
## CI.mean.0.95    0.06    0.04    0.09    4.00   0.07   0.13
## var             0.11    0.06    0.23  347.89   0.13   0.47
## std.dev         0.33    0.24    0.48   18.65   0.36   0.69
## coef.var        2.71    3.96    0.05    0.28   2.34   1.06
```

There are 116 observations of 14 variables (which correspond to 58 schools over two years, 1992 and 1993.

There are 115 `NA`s in the whole dataset, distributed as shown below:

```
colSums(is.na(data[colSums(is.na(data)) > 0]))
```

```
##  top25 ver500 mth500 avg500
##     25     30     30     30
```

None of the variables of interest contains missing values (so we don't have to omit any observation).

For the rest of this assignment we'll focus on the 4 variables of interest mentioned above (and some others built from those ones), plus the names of the schools and the year (19922 or 1993).

> There's no need to keep `lapps` since it's just the log of `apps` and we will only use its change from 1992 to 1993 (so we will have to apply a transformation anyway: $\log(apps.1993) - \log(apps.1992)$ instead of $lapps.1993 - lapps.1992$; the results may be slightly different—and more precise—because of the rounding decimal error at storing `lapps`).

```
# Keep only variables of interest
# Also convert binary variables to logical
# (not necessary but better for plotting)
# No need to keep 'lapps', it's just log('apps')
categories <- c('bowl', 'btitle', 'finfour')
vars_of_interest <- c('year', 'school', 'apps', categories)
data2 <- data %>%
  select(match(vars_of_interest, names(data))) %>%
  mutate_each_(funs(as.logical), categories)
subset(desc, variable %in% vars_of_interest)
```

```
##    variable                          label
## 1      year                  1992 or 1993
## 2      apps       # applics for admission
## 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
## 12   school            name of university
```

```
head(data2)
```

```
##   year        school  apps  bowl btitle finfour
## 1 1992       alabama  6245  TRUE  FALSE   FALSE
## 2 1993       alabama  7677  TRUE  FALSE   FALSE
## 3 1992       arizona 13327 FALSE  FALSE   FALSE
## 4 1993       arizona 19860  TRUE   TRUE   FALSE
## 5 1992 arizona state 10422 FALSE  FALSE   FALSE
## 6 1993 arizona state 12809 FALSE  FALSE   FALSE
```

Table 1: Number of observations per group

bowl	btitle	finfour	Freq
FALSE	FALSE	FALSE	53
TRUE	FALSE	FALSE	45
FALSE	TRUE	FALSE	7
TRUE	TRUE	FALSE	4
FALSE	FALSE	TRUE	1
TRUE	FALSE	TRUE	3
FALSE	TRUE	TRUE	1
TRUE	TRUE	TRUE	2

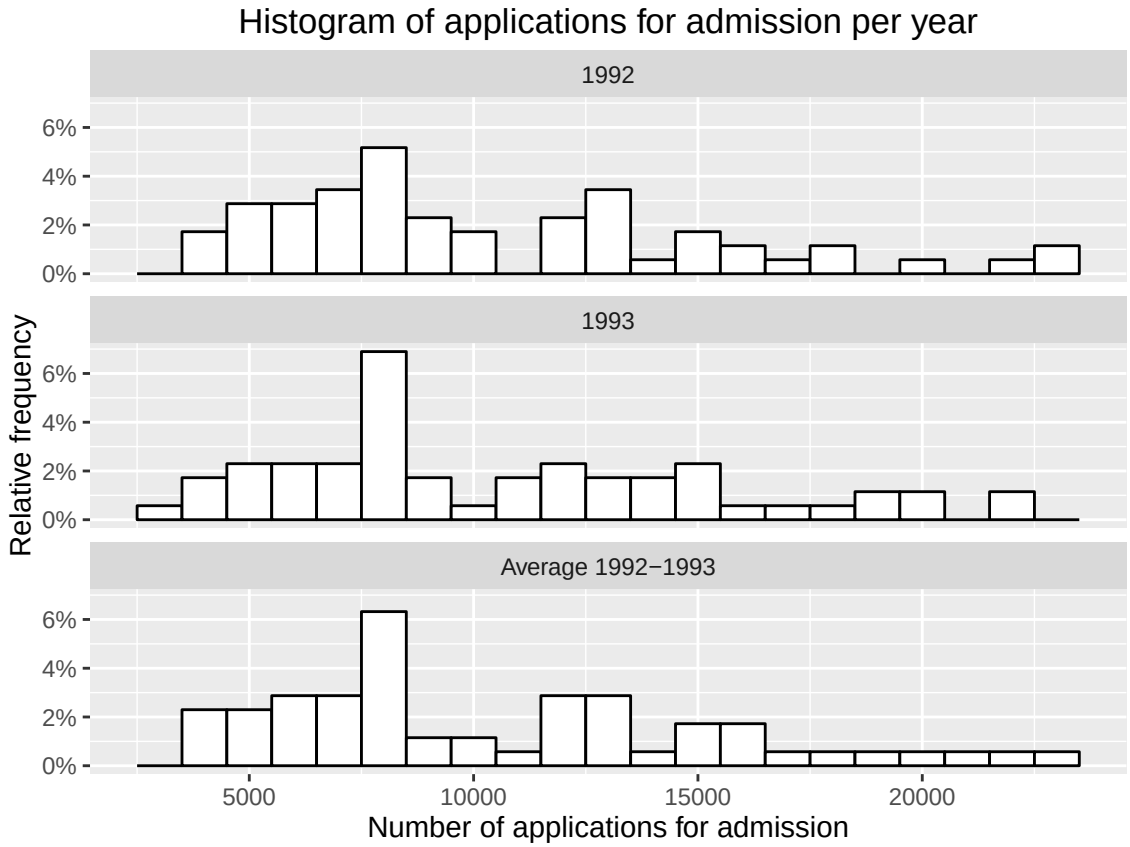

Figure 1: Histogram of applications for admission per year

As shown above, in Figure 1, the sample distribution of `apps` is right-skewed (i.e., its skewness is positive skewed) and platykurtic (its excess kurtosis is negative, so it has thinner tails than the normal distribution). Both things happen each year, as well as to the average number of applications

The figures in the next three pages show the number of observations depending on the three athletic success indicators for each of the two years under study, as well as the mean number of applications for admission is different whether if the school won the bowl game, the men's conference championship, or the final four in the previous year, but that difference is never significant (see how the confidence intervals always overlap in Figures 2, 4, and 6).

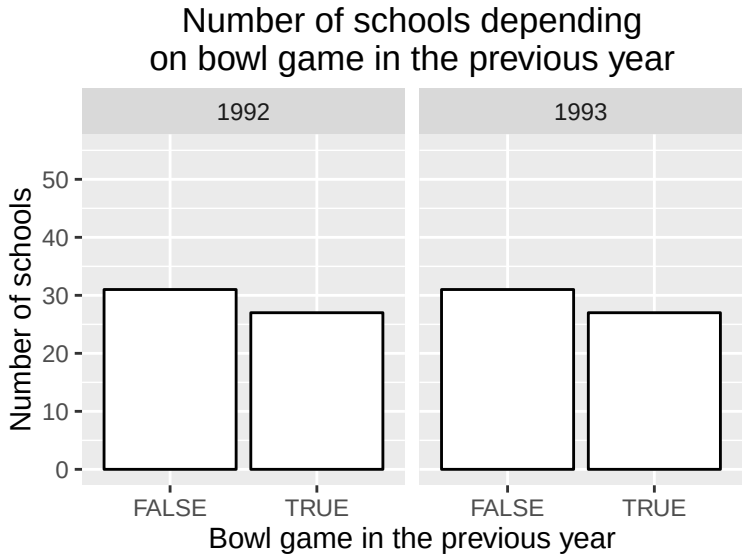

Figure 2: Number of schools depending on bowl game in the previous year

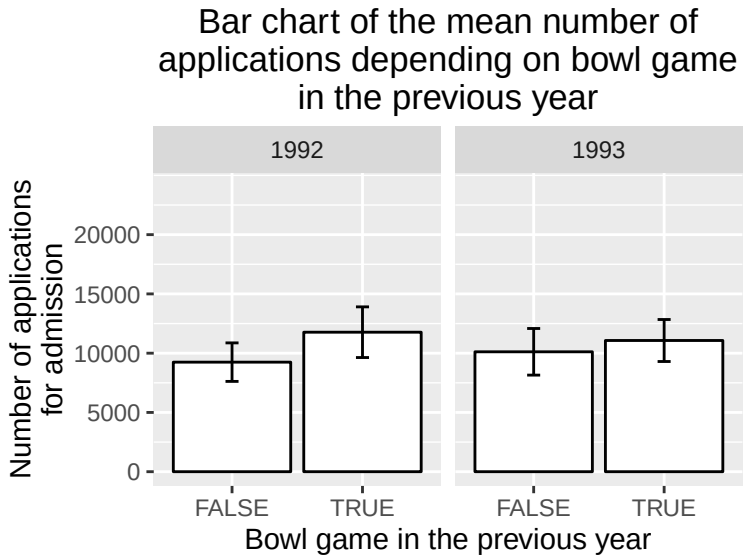

Figure 3: Bar chart of the mean number of applications depending on bowl game in the previous year

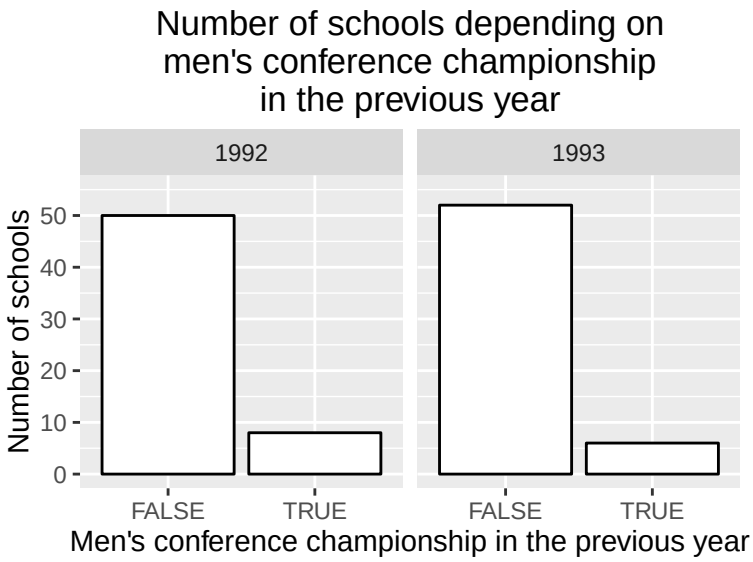

Figure 4: Number of schools depending on men's conference championship in the previous year

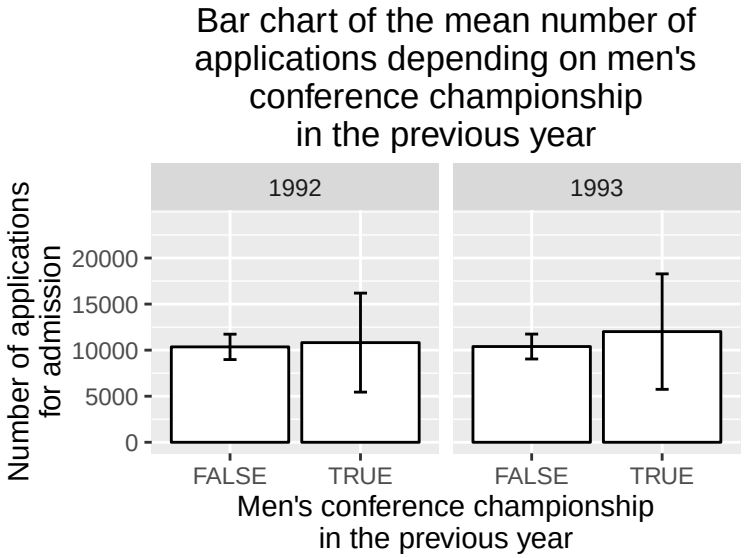

Figure 5: Bar chart of the mean number of applications depending on men's conference championship in the previous year

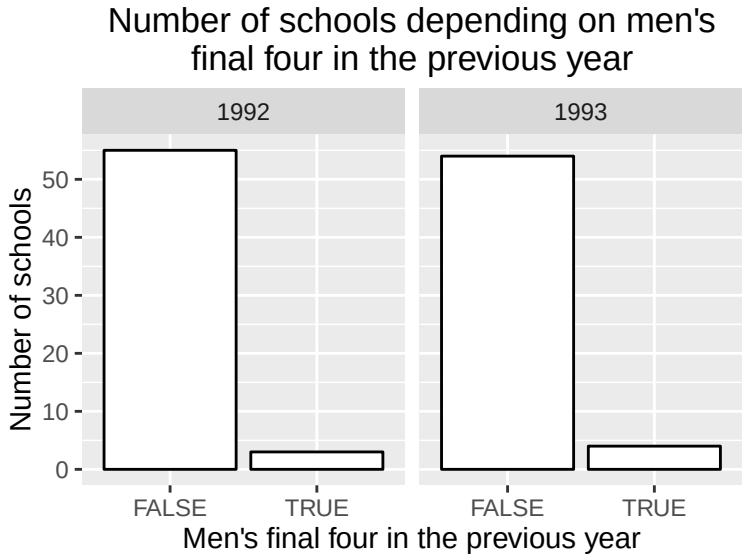

Figure 6: Number of schools depending on men's final four in the previous year

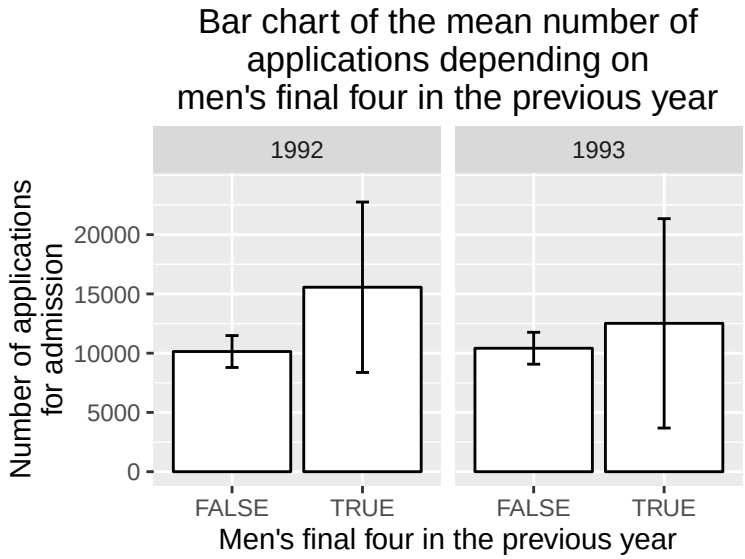

Figure 7: Bar chart of the mean number of applications depending on men's final four in the previous year

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.

```
data3 <- reshape(data2, idvar = "school", timevar = "year",
                 v.names = c("apps","bowl", "btitle", "finfour"),
                 varying = list(c("apps.1992", "apps.1993"),
                                c("bowl.1992", "bowl.1993"),
                                c("btitle.1992", "btitle.1993"),
                                c("finfour.1992", "finfour.1993")),
                 direction = "wide")
head(data3, 4)
```

```
##          school apps.1992 bowl.1992 btitle.1992 finfour.1992 apps.1993
## 1       alabama      6245      TRUE       FALSE        FALSE      7677
## 3       arizona     13327     FALSE       FALSE        FALSE     19860
## 5 arizona state     10422     FALSE       FALSE        FALSE     12809
## 7      arkansas      4103      TRUE        TRUE        FALSE      3303
##   bowl.1993 btitle.1993 finfour.1993
## 1      TRUE       FALSE        FALSE
## 3      TRUE        TRUE        FALSE
## 5     FALSE       FALSE        FALSE
## 7     FALSE       FALSE        FALSE
```

```
# Same using tidyr: melt/gather columns + unite variable w/ year + spread/dcast
data3 <- data2 %>%
  gather(variable, value, -(year:school)) %>%
  unite(temp, variable, year, sep = ".") %>%
  spread(temp, value)
# Convert to logical values again
vars_to_convert <- unlist(lapply(categories, function(x) grep(x, names(data3))))
data3 <- data3 %>%
  mutate_each_(funs(as.logical), names(data3)[vars_to_convert])
head(data3, 4)
```

```
##          school apps.1992 apps.1993 bowl.1992 bowl.1993 btitle.1992
## 1       alabama      6245      7677      TRUE      TRUE       FALSE
## 2       arizona     13327     19860     FALSE      TRUE       FALSE
## 3 arizona state     10422     12809     FALSE     FALSE       FALSE
## 4      arkansas      4103      3303      TRUE     FALSE        TRUE
##   btitle.1993 finfour.1992 finfour.1993
## 1       FALSE        FALSE        FALSE
## 2        TRUE        FALSE        FALSE
## 3       FALSE        FALSE        FALSE
## 4       FALSE        FALSE        FALSE
```

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.

```
# data3$clapps <- log(data3$apps.1993) - log(data3$apps.1992)
# data3$clapps <- log(data3$apps.1993 / data3$apps.1992)
data3 <- data3 %>%
  mutate(clapps = log(apps.1993) - log(apps.1992))
# Results may differ from those we'd obtain using lapps due to decimals!!!
head(data3, 4)
```

```
##          school apps.1992 apps.1993 bowl.1992 bowl.1993 btitle.1992
## 1       alabama      6245      7677      TRUE      TRUE       FALSE
## 2       arizona     13327     19860     FALSE      TRUE       FALSE
## 3 arizona state     10422     12809     FALSE     FALSE       FALSE
## 4      arkansas      4103      3303      TRUE     FALSE        TRUE
##   btitle.1993 finfour.1992 finfour.1993     clapps
## 1       FALSE        FALSE        FALSE  0.2064477
## 2        TRUE        FALSE        FALSE  0.3989156
## 3       FALSE        FALSE        FALSE  0.2062291
## 4       FALSE        FALSE        FALSE -0.2168873
```

Which schools had the greatest increase and the greatest decrease in number of log applications?

```
# (schools_greatest_increase <- head(data3[order(data3$clapps,
#                                                decreasing = TRUE), ], 5))
schools_greatest_increase <- data3 %>%
  arrange(desc(clapps)) %>%
  select(school, apps.1992, apps.1993,clapps) %>%
  head(5)
```

Table 2: Schools with the greatest increase in number of log applications

school	apps.1992	apps.1993	clapps
arizona	13327	19860	0.3989156
alabama	6245	7677	0.2064477
arizona state	10422	12809	0.2062291
oregon	7159	8631	0.1869901
villanova	6611	7759	0.1601185

```
tableCount <- incCount(tableCount, "table-Q2-1")
```

```
# (schools_greatest_decrease <- tail(data3[order(data3$clapps,
#                                                decreasing = TRUE), ], 5))
schools_greatest_decrease <- data3 %>%
  arrange(clapps) %>%
  select(school, apps.1992, apps.1993,clapps) %>%
  head(5)
```

Table 3: Schools with the greatest decrease in number of log applications

school	apps.1992	apps.1993	clapps
arkansas	4103	3303	-0.2168873
oklahoma state	4892	4102	-0.1761266
penn state	22930	19315	-0.1715641
auburn	8661	7548	-0.1375476
louisiana state	6707	6000	-0.1113923

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.

```
# data3$cbowl <- data3$bowl.1993 - data3$bowl.1992
data3 <- data3 %>%
  mutate(cbowl = bowl.1993 - bowl.1992, cperf = btitle.1993 - btitle.1992,
         cbball = finfour.1993 - finfour.1992)
head(data3, 4)
```

```
##          school apps.1992 apps.1993 bowl.1992 bowl.1993 btitle.1992
## 1       alabama      6245      7677      TRUE      TRUE       FALSE
## 2       arizona     13327     19860     FALSE      TRUE       FALSE
## 3 arizona state     10422     12809     FALSE     FALSE       FALSE
## 4      arkansas      4103      3303      TRUE     FALSE        TRUE
##   btitle.1993 finfour.1992 finfour.1993     clapps cbowl cperf cbball
## 1       FALSE        FALSE        FALSE  0.2064477     0     0      0
## 2        TRUE        FALSE        FALSE  0.3989156     1     1      0
## 3       FALSE        FALSE        FALSE  0.2062291     0     0      0
## 4       FALSE        FALSE        FALSE -0.2168873    -1    -1      0
```

Which of these variables has the highest variance?

First, let's see how many of the 58 schools in the sample won each title each year:

```
data3 %>% select(matches('bowl.1|finfour|btitle')) %>% summarise_each(funs(sum))
```

```
##   bowl.1992 bowl.1993 btitle.1992 btitle.1993 finfour.1992 finfour.1993
## 1        27        27           8           6            3            4
```

Of course, this does not tell us anything about the variance: the same 27 schools that won the bowl game in 1991 could have won it again in 1992.

```
(v <- data3 %>% select(cbowl, cperf, cbball) %>% summarise_each(funs(var)))
```

```
##       cbowl     cperf     cbball
## 1 0.3157895 0.1742287 0.08741682
```

As shown above, **cbowl** is the variable with the highest variance.

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.

First of all, we'll change the names of `cperf`, `cfinfour`, and `cbball` (as defined in Question 3) to follow the notation in the formula above:

```
data3 <- data3 %>%
  rename(cbtitle = cperf, cfinfour = cbball)
```

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.**

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.

```
model1 <- lm(clapps ~ cbowl + cbtitle + cfinfour, data3)
```

Table 4: Regression summary

	Dependent variable:
	Change in log(applications)
Change in bowl game in previous year	0.0570*
	(0.0272)
Change in men's conference championship in previous year	0.0415
	(0.0443)
Change in men's final four in previous year	−0.0696
	(0.0668)
Intercept (Constant): year 1993	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

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?
