Applied Regression and Time Series Analysis (2016 Fall): HW4 - Week 6

Jeffrey Yau and Devesh Tiwari

October 5, 2016

Instructions:

The weekly assignment serves two purposes: (1) Review concepts, techniques, theories, statistical models covered during the week. (2) Extend the materials taught in the asynchronized lectures, assigned readings, and live sessions; some new concepts and/or techniques are introduced in the weekly assignment.

Below are specific instructions:

- **Due: 10/16/2016 (11:59pm PST)**
- You may complete this assignment on your own or in a group of no more than 3 students.
- When working in a group, you are strongly encouraged to complete the assignment on your own before discussing your group mates. Do not use the "division-of-labor" approach to complete the assignment.
- The homework is designed as a quantitative analysis. The instructions and questions are designed to guide you through the analysis of data using regression techniques. As such, you should think of it as a quantitative case study and the result of the study is a report with a set of well-written codes that can be used to reproduce the results in the report.
- Submission:
  - Submit your own assignment via ISVC
  - Submit 2 files:
    1. R-script or R markdown file
    2. A pdf file including the summary, the details of your analysis, and all the R codes used to produce the analysis
  - Each group only needs to submit one set of files
  - Use the following file naming convensation; fail to do so will receive 10% reduction in the grade:
    * **SectionNumber__hw02__LastNameFirstInitial.fileExtension**
    * Examples:
      - Section1__hw02__YauJ.Rmd
      - Section1__hw02__YauJ.pdf
      - Section1__hw02__TiwariD__YauJ.Rmd
      - Section1__hw02__TiwariD__YauJ.pdf

Objective:

The key objective of this homework is to practice the use of the difference-in-difference technique to handle potential bias arising from omitted variables.

Description of 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:

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.

Question 1:

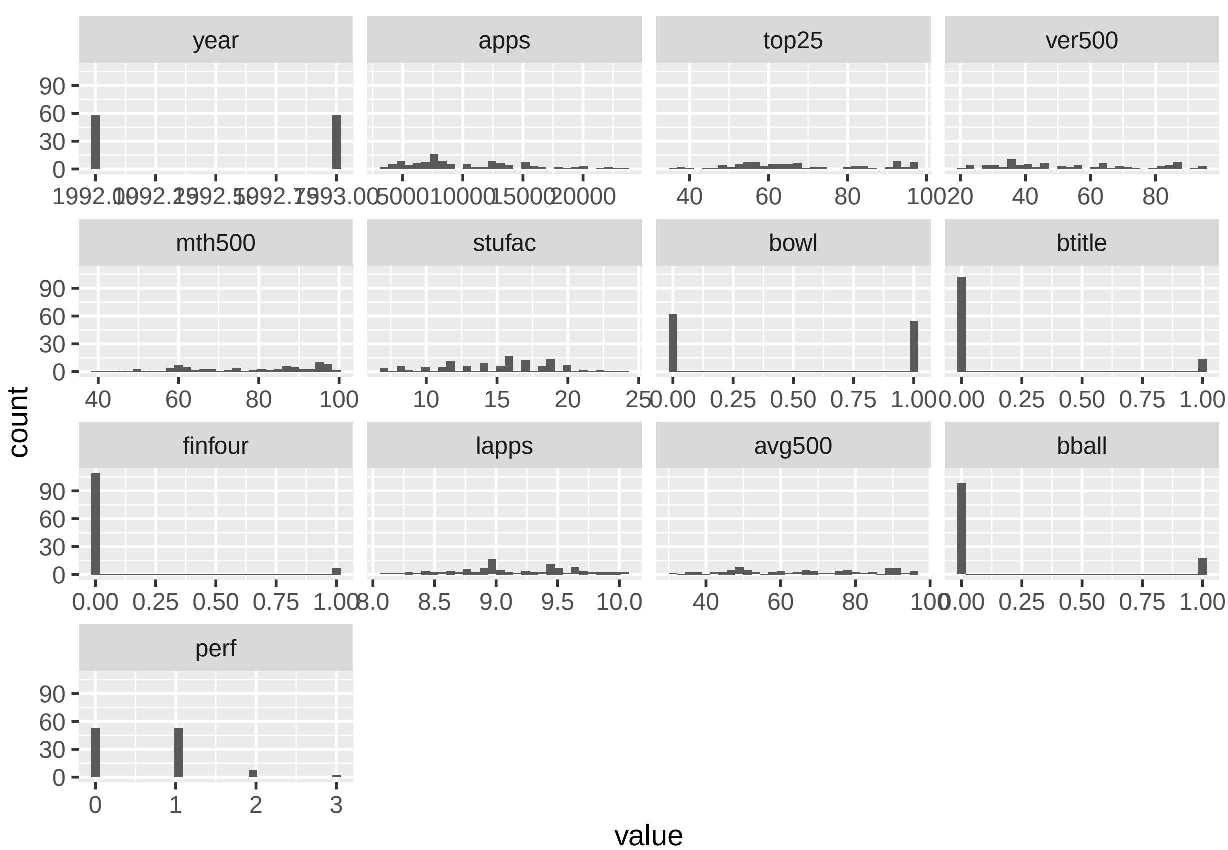
Conduct a quick examination and EDA of the dataset.

From the histograms, you can see that there are a mix of true metric variables, such as apps, and numeric variables, such as year, that should be treated as binary or categorical. From the line charts, we also see that several variables are very co-linear Ver500 and avg500 are almost perfectly co-linear.

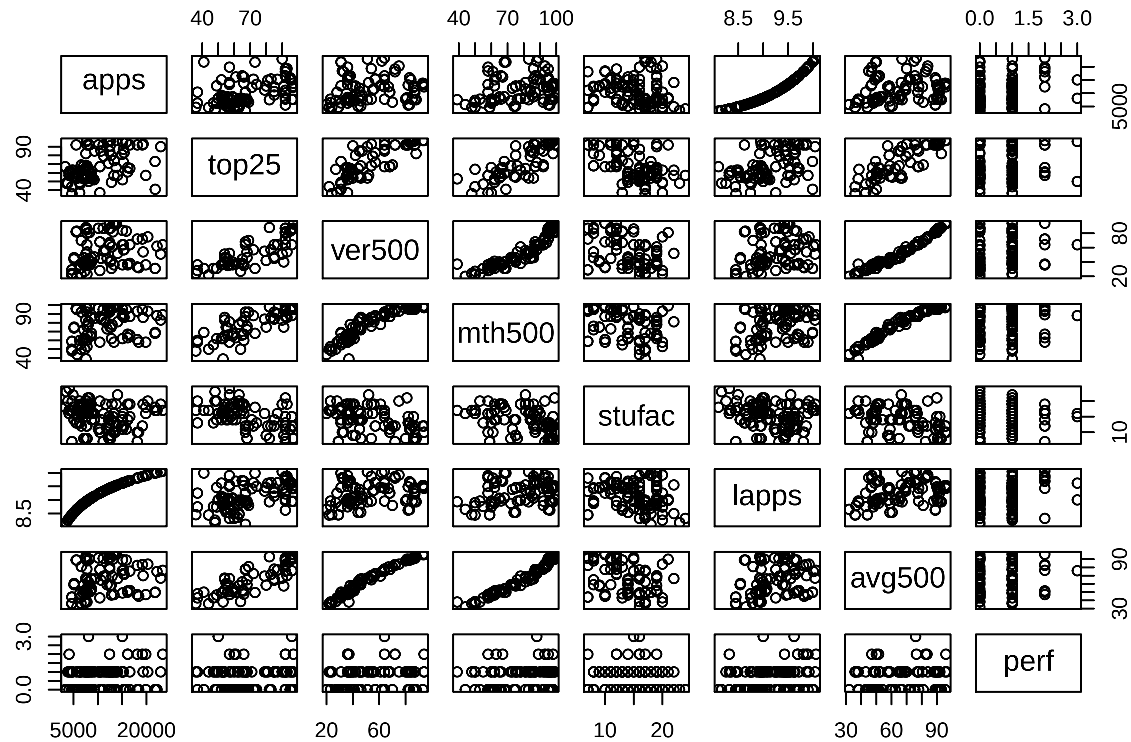
```
EDA = function(data){
  print("Summary")
print (summary(data))


d <- melt(data)
 print("Histograms of each variable")
 print(ggplot(d,aes(x = value)) +
    facet_wrap(~variable,scales = "free_x") +
    geom_histogram(),newpage=TRUE)

print("")
print("Plots of data (only metric columns with more than two values)")
plot(data[, sapply(data, function(x){is.numeric(x) && length(unique(x)) > 2})])
}

EDA(data)
```

```
## [1] "Summary"
##       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
##

## Using school as id variables

## [1] "Histograms of each variable"

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 115 rows containing non-finite values (stat_bin).
```



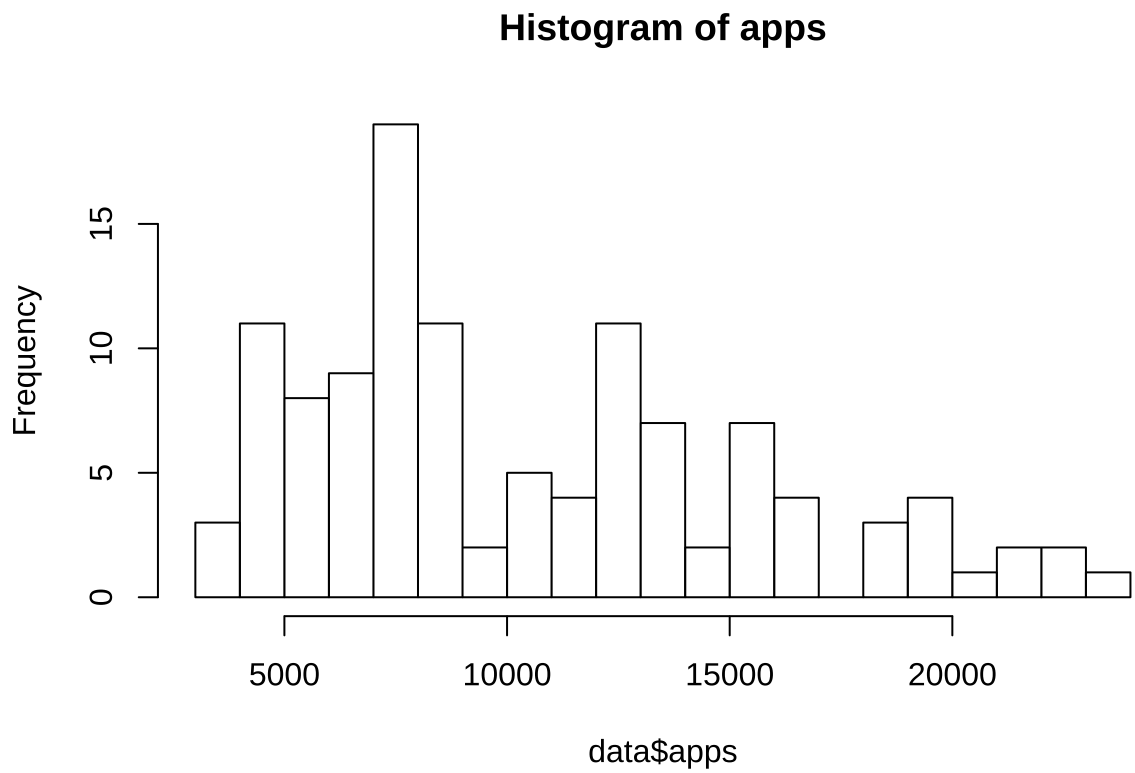
```
## [1] ""
## [1] "Plots of data (only metric columns with more than two values)"
```



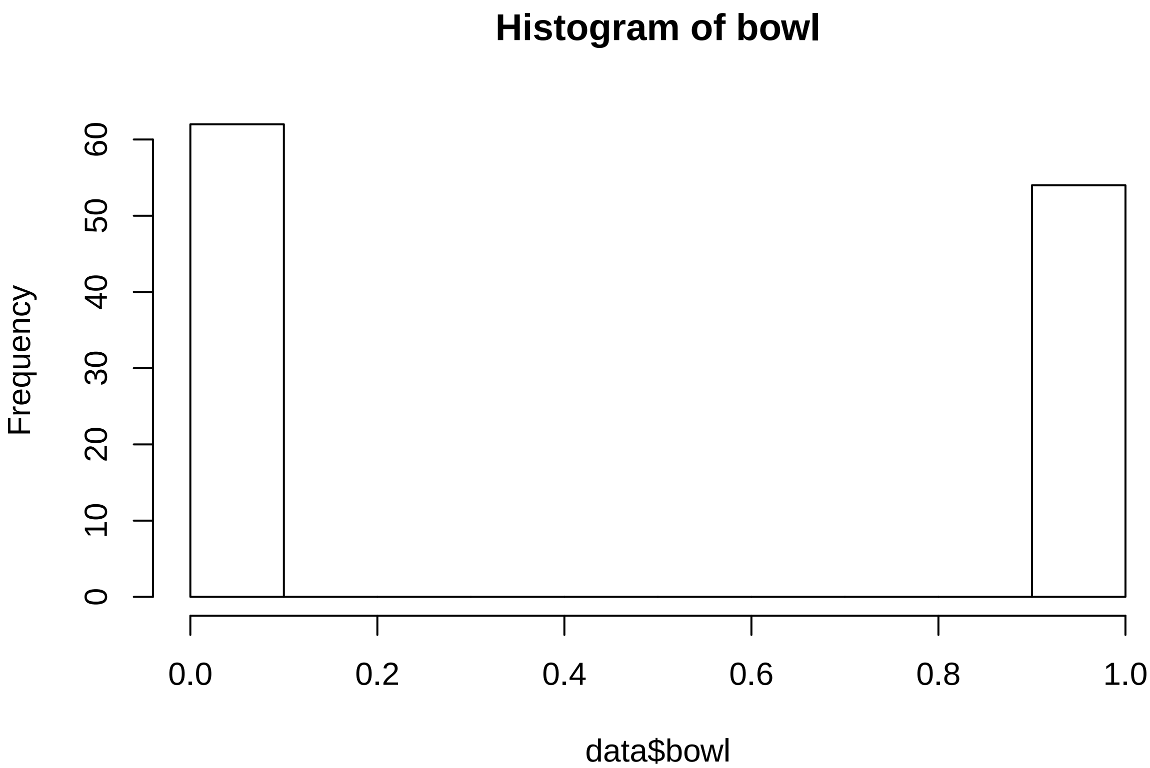
Examine the variables of interest: *apps* represents the number of applications for admission. *bowl, btitle*, and *finfour* are indicators of athletic success. The three athletic performance variables are all lagged by one year. Intuitively, this is because we expect a school's athletic success in the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

The apps variable has a somewhaat skewed right distribution and appears to be bi-modal. So it will not be a good fit for OLS regression in its current state. The bowl variable is a binary variable that seems pretty evenly distributed, while btitle and finfour are much more one sided.

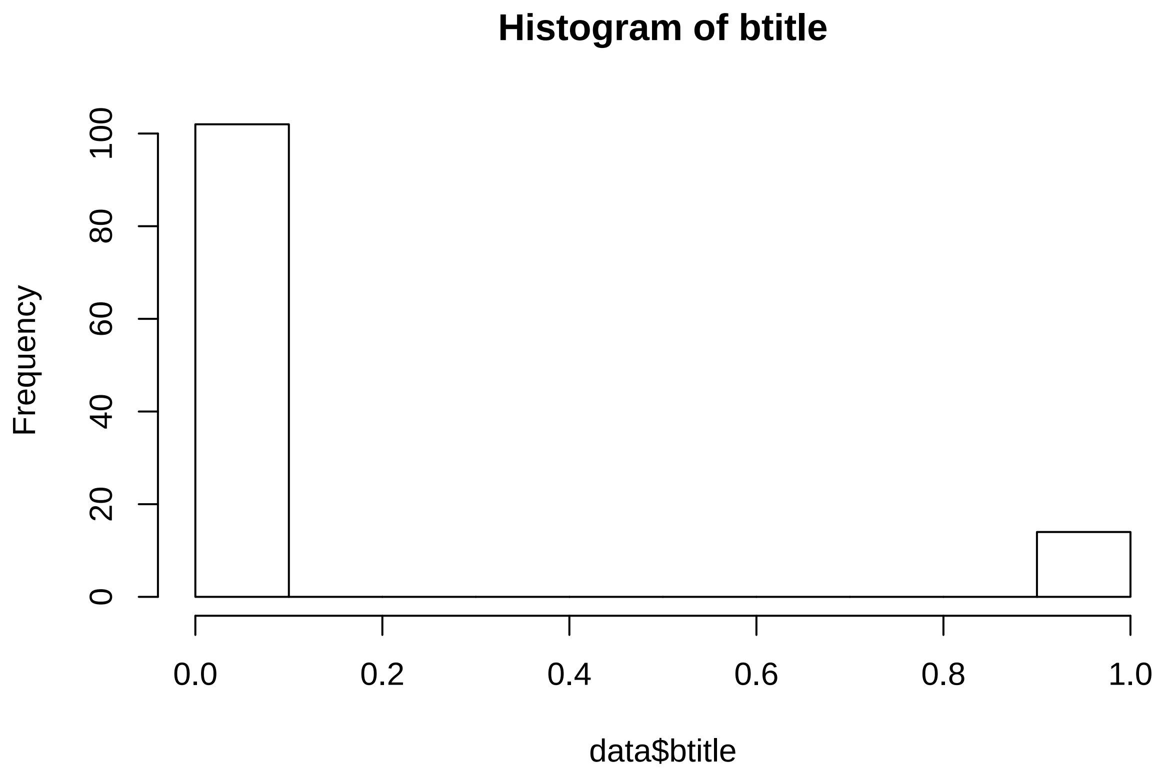
```r
hist(data$apps, main="Histogram of apps",breaks=20)
```



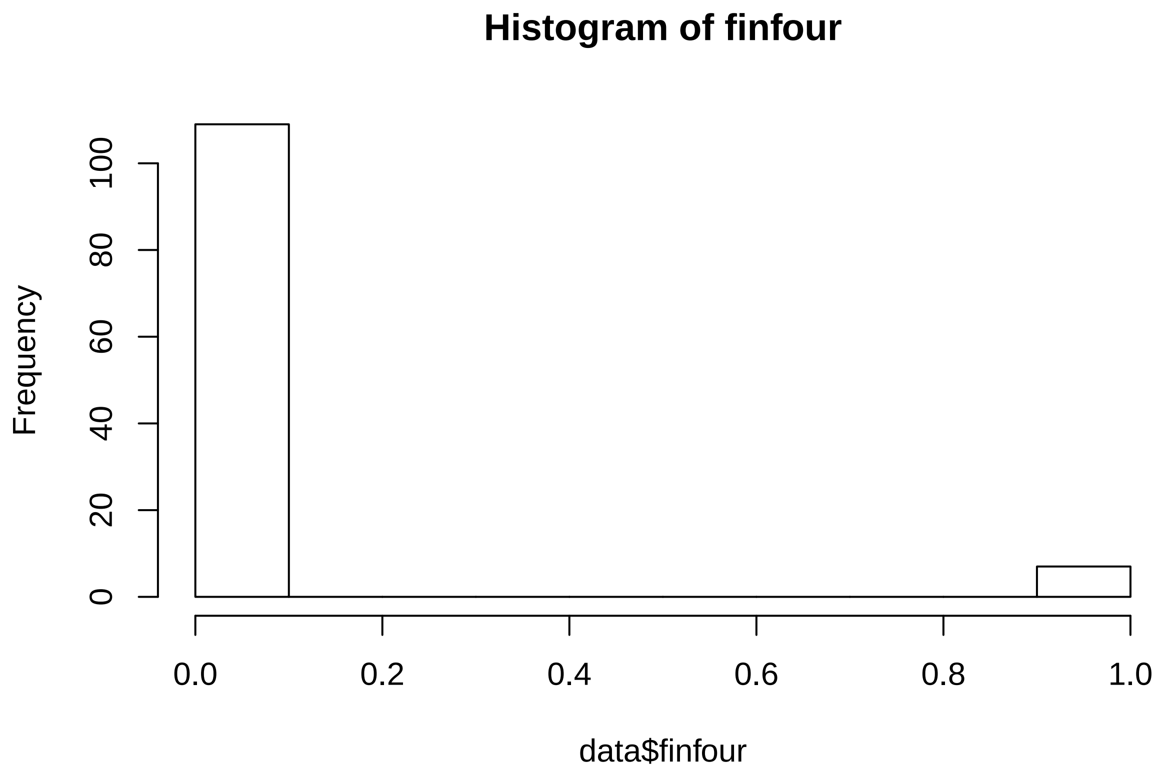

```
hist(data$bowl, main="Histogram of bowl")
```




```
hist(data$btitle, main="Histogram of btitle")
```




```
hist(data$finfour, main="Histogram of finfour")
```



Question 2:

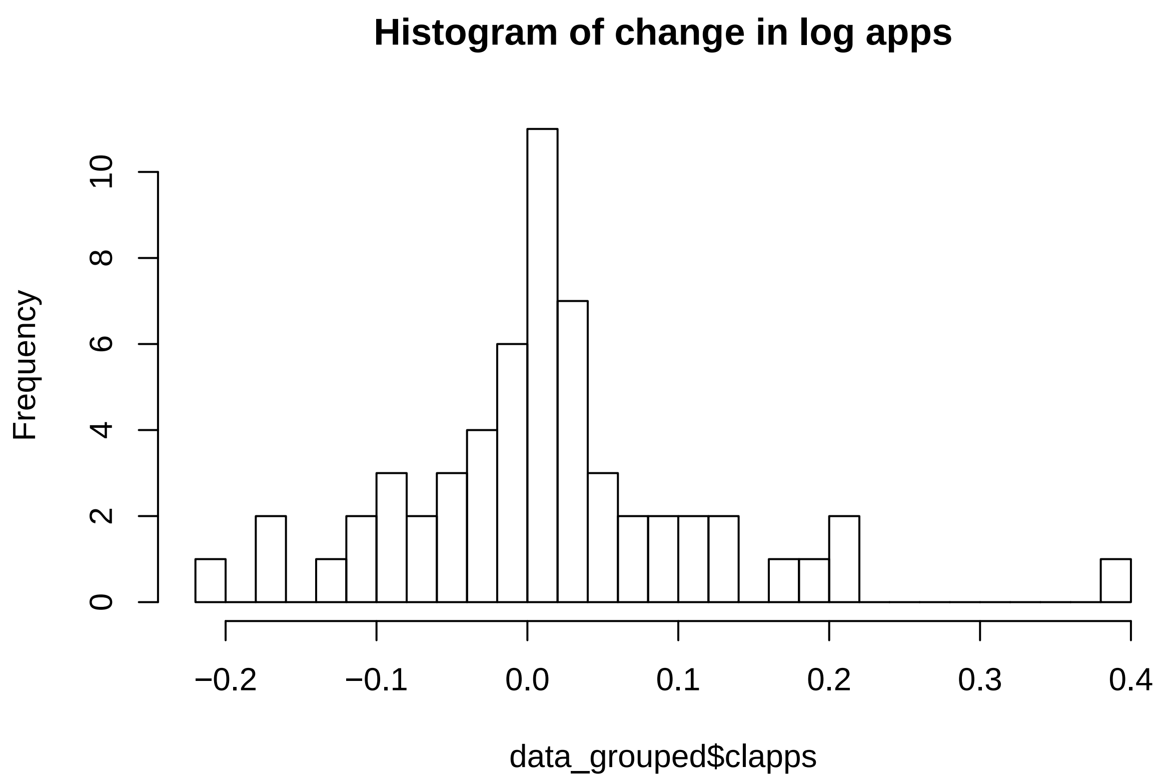
Note that the data set is in long format, with a separate row for each year for each school. To prepare for a difference-in-difference analysis, transfer the dataset to wide-format. Each school should have a single row of data, with separate variables for *1992* and *1993*. For example, you should have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

```
data_grouped <- dcast(melt(data,id.vars = c("school","year")),school~variable+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.

The clapps variabble appears to be somewhat normally distributed, although there are more values clustered at or around 0. There's one outlier at .4.

```
data_grouped$clapps = data_grouped$lapps_1993 - data_grouped$lapps_1992
hist(data_grouped$clapps,main="Histogram of change in log apps",breaks=30)
```



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

```r
print("Top 5 schools with greatest increase")
```

```
## [1] "Top 5 schools with greatest increase"
```

```r
head(arrange(data_grouped[c("school","clapps")],desc(clapps)), n = 5)
```

```
##           school    clapps
## 1        arizona 0.3989162
## 2        alabama 0.2064476
## 3  arizona state 0.2062283
## 4         oregon 0.1869907
## 5      villanova 0.1601181
```

```r
print("Top 5 schools with greatest decrease")
```

```
## [1] "Top 5 schools with greatest decrease"
```

```r
head(arrange(data_grouped[c("school","clapps")],clapps), n = 5)
```

```
##             school     clapps
## 1         arkansas -0.2168865
```

```
## 2  oklahoma state -0.1761265
## 3      penn state -0.1715641
## 4          auburn -0.1375475
## 5 louisiana state -0.1113930
```

Question 3 Similarly to above, create three variables, *cbowl*, *cbtitle*, and *cfinfour*, where each of these variables represents *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.

```
data_grouped$cbowl = data_grouped$bowl_1993 - data_grouped$bowl_1992
data_grouped$cbtitle = data_grouped$btitle_1993 - data_grouped$btitle_1992
data_grouped$cfinfour = data_grouped$finfour_1993 - data_grouped$finfour_1992
```

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,

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

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.

In order to be causal, there needs to be no other variables that are correlated (either postive or negative) with the three explanatory variables and that is causal to the dependent variable.

$$\nexists_{z \notin X}(z \rightarrow y \wedge \forall_{xinX} cov(x, z) > 0)$$

Where X is the set of explanatory variables in the regression

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.

In order for OLS to consistently estimate the dependent variable, the error term must be uncorrelated with each of the three explanatory variables.

$$\forall_{x \in X} cov(\epsilon | x) = 0)$$

Where X is the set of explanatory variables in the regression

This asssumption is plausible. While there may be other factors that influence number of applications, there isn't any reason to suspect that these factors might be more or less correlated with the dependent variable if a college has worse or better athletic performance between years.

Question 5 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 null hypothesis is that the three explanatory variables all have coefficients of 0

```
model = lm(clapps ~ cbowl + cbtitle + cfinfour,data=data_grouped)
summary(model)
```

```
##
## Call:
## lm(formula = clapps ~ cbowl + cbtitle + cfinfour, data = data_grouped)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.192965 -0.042868 -0.006367  0.040005  0.283578
##
## 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
```

The model is jointly significant, as the p value is less than .05. This suggests that the change in athletic performance between 1991 and 1992 would effect change in applications between 1992 and 1993.