

HW4

Subhashini R., Lei Yang, Ron Cordell

February 23, 2016

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 receives in the current year.

```
setwd('~/Documents/MIDS/W271/hw4')
load('athletics.RData')
desc
```

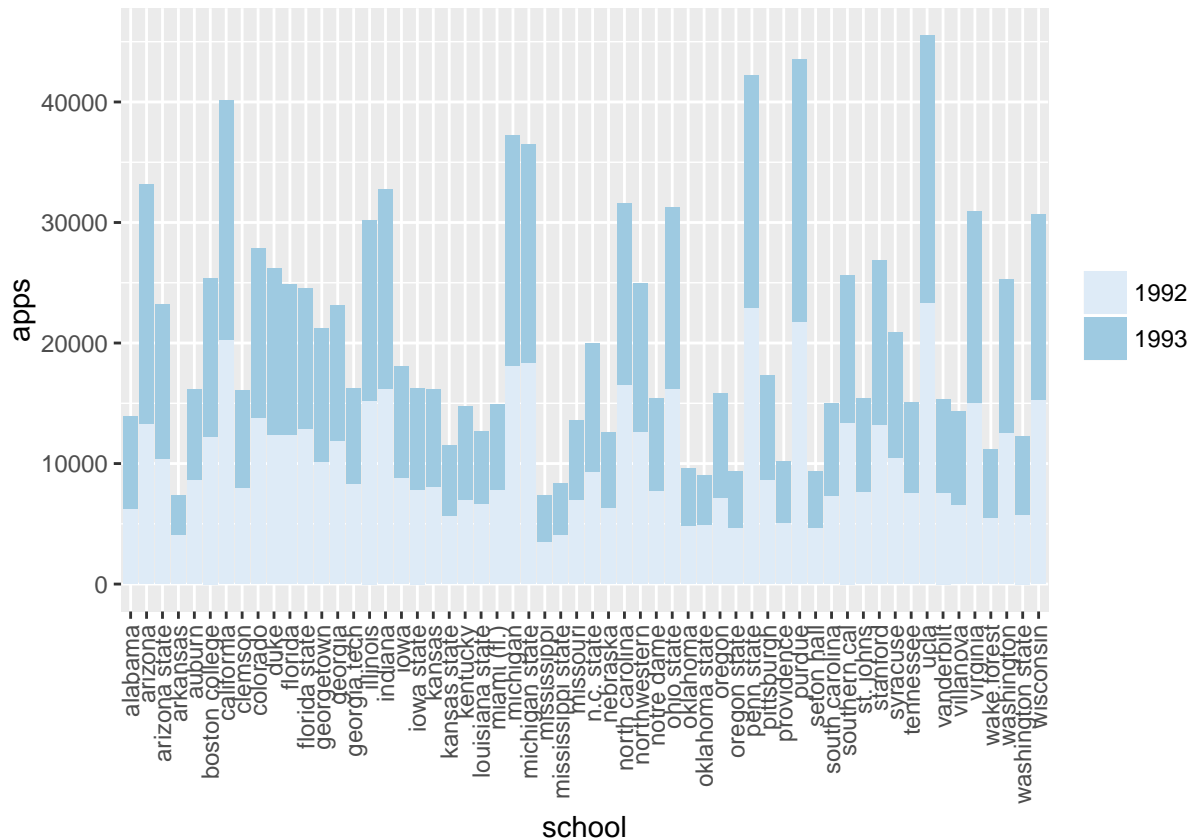
```
##      variable                                label
## 1      year                                1992 or 1993
## 2      apps                                # applcs for admission
## 3      top25    perc frsh class in 25 hs perc
## 4      ver500    perc frsh >= 500 on verbal SAT
## 5      mth500    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 ...
```

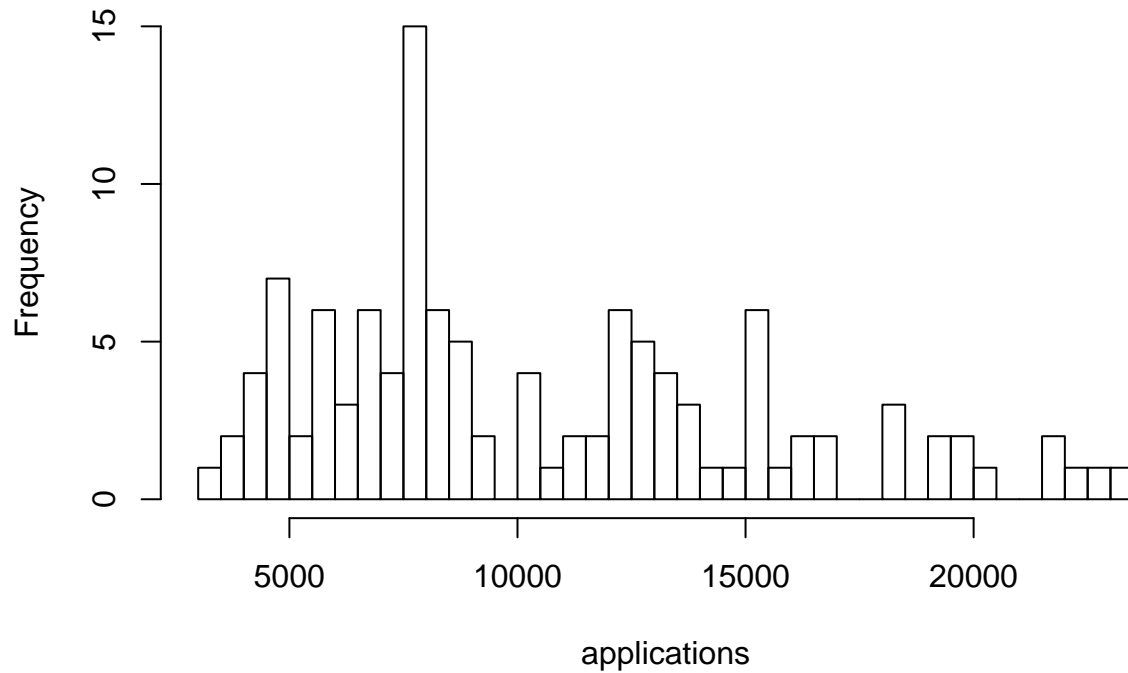
The athletics data file consists of 116 observations of data evenly divided across two years: 1992 and 1993.

The year column is an integer as opposed to a label or level, so we may need to change that. An examination of the data doesn't reveal anything out of place. A scatterplot of applications by school for both years shows that schools received relatively similar numbers of applications for each year.



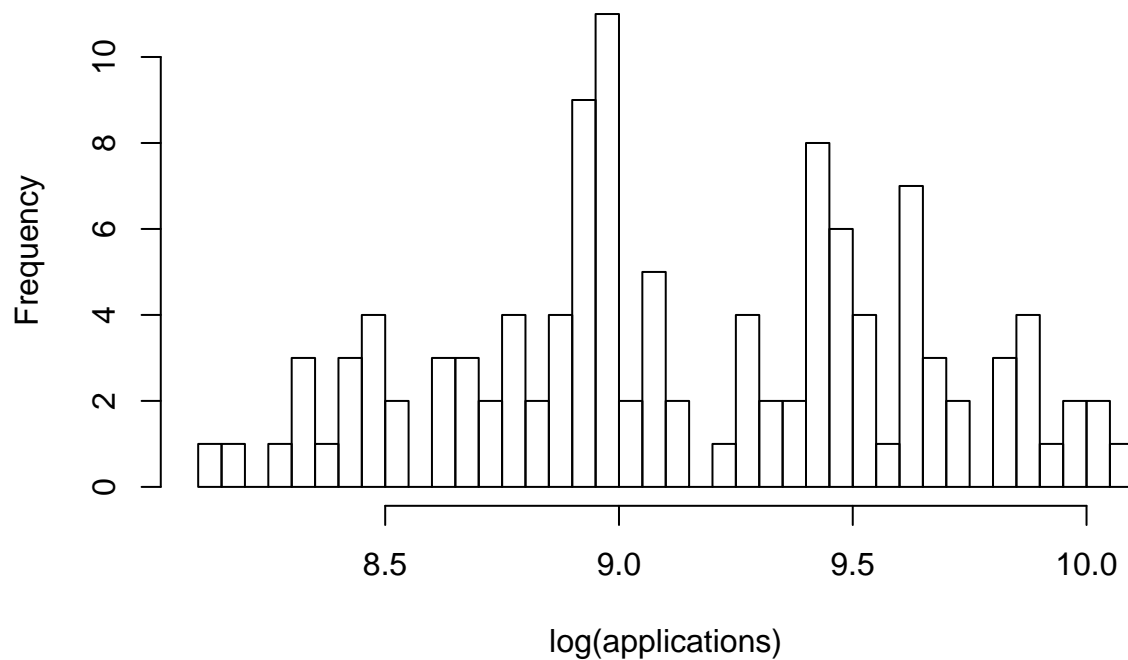
When we examine the histogram of the apps variable we can see that it has a skewed distribution.

Histogram of Applications



Taking a look at the `lapps` (`log(apps)`) variable correct the skewness of the `applications` variable distribution but the distribution almost appears bimodal.

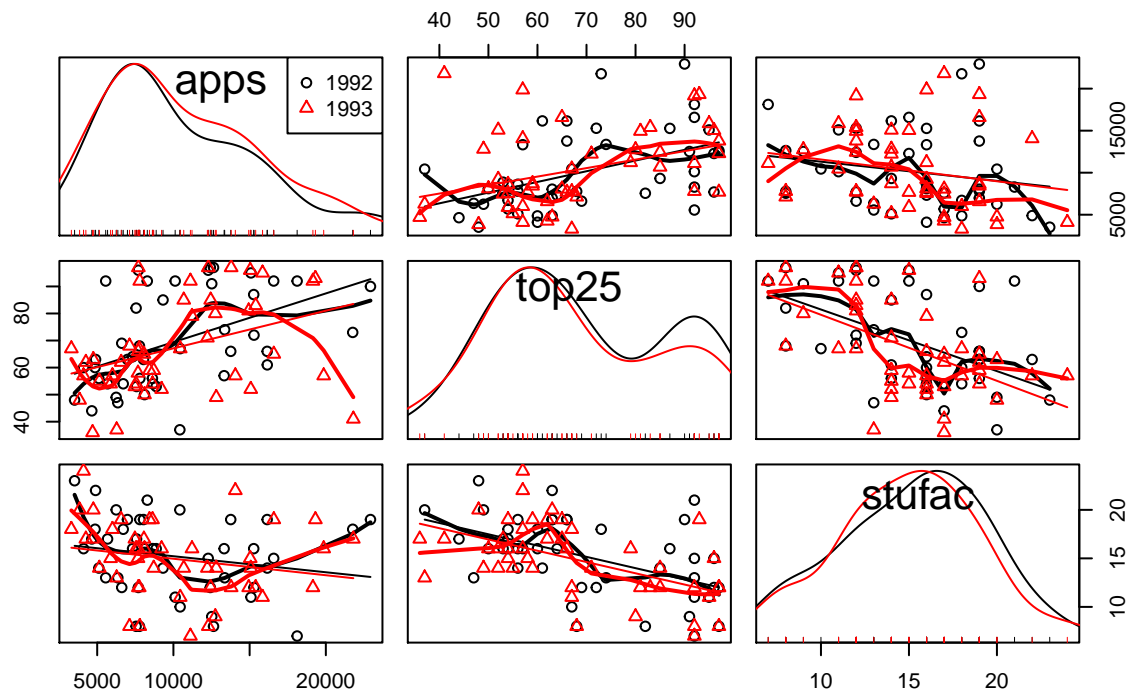
Histogram of $\log(\text{applications})$



Exploring the `Applications` a little more we compare the `lapps` variable to the other continuous variables in the data set using a couple of scatterplot matrices. The first matrix compares `applications`, `top 25%` and `student faculty ratio` grouping by `year`. There don't appear to be any significant correlations between the

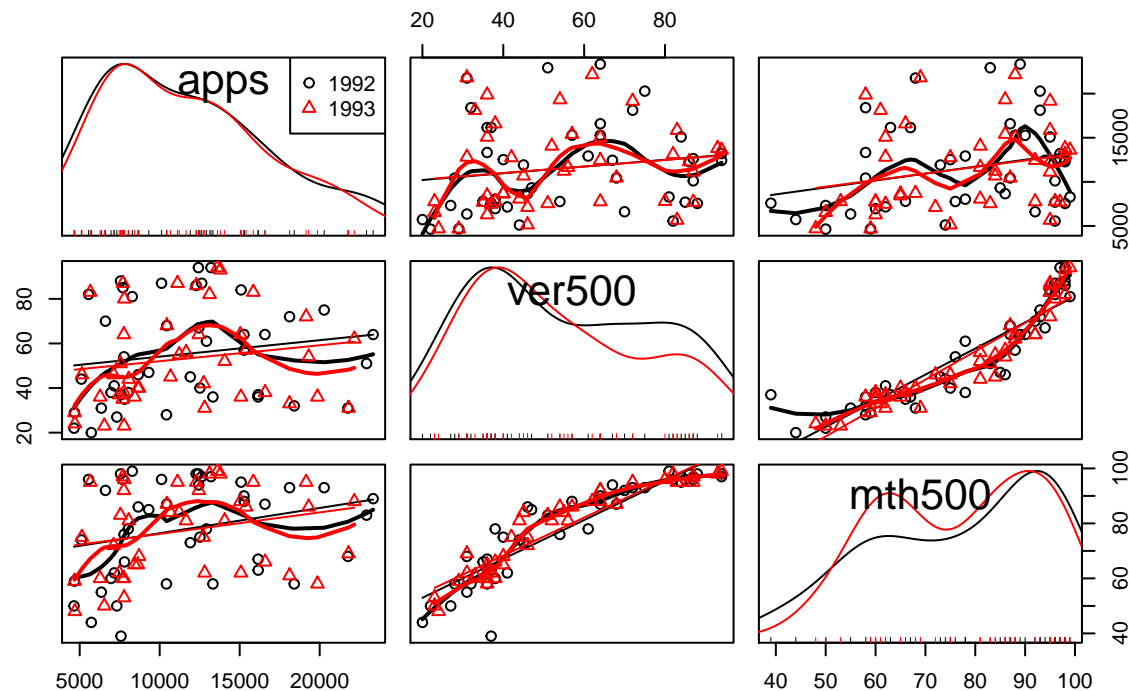
variables on visual inspection.

Applications, Top 25%, Student-Faculty Ratio for 1992 & 1993



The second matrix compares $\log(\text{apps})$ with SAT scores. There is a very strong correlation between the verbal and math SAT scores but neither appears to correlate with the apps variable

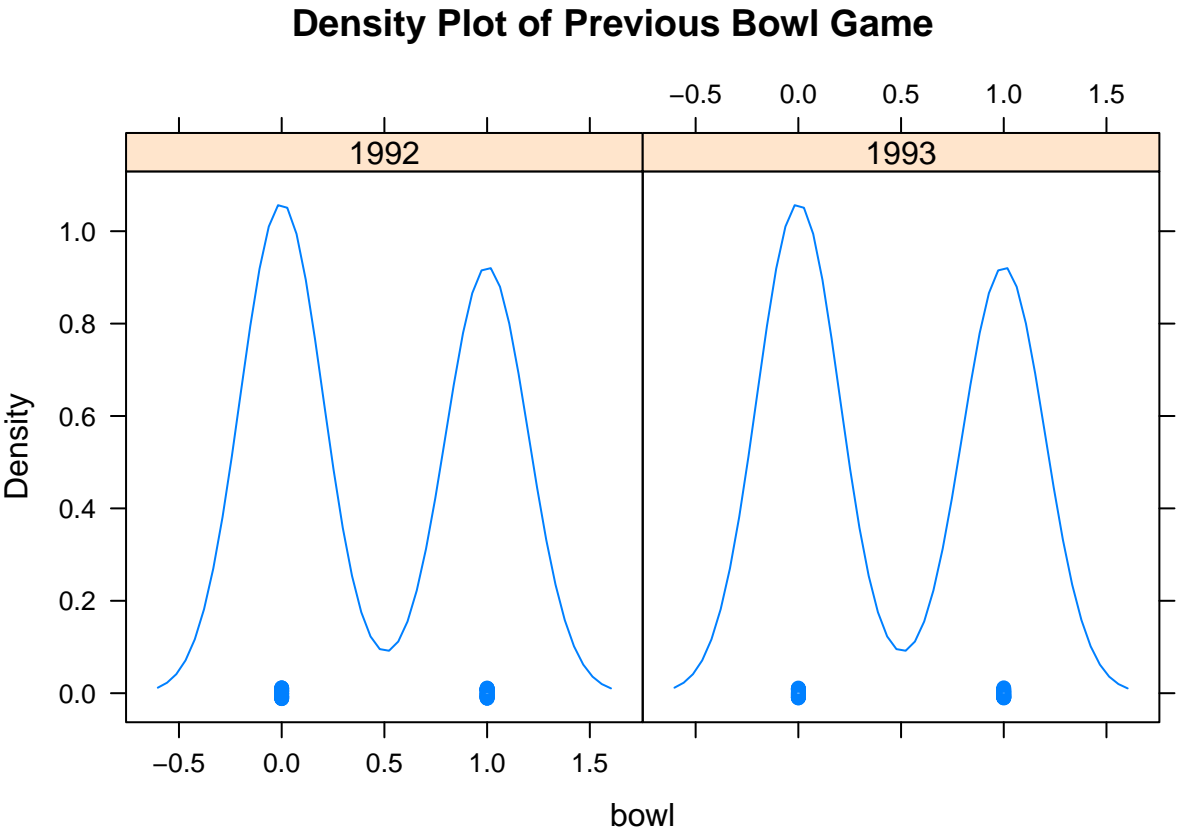
Applications, SAT Verbal > 500, SAT Math > 500 for 1992 & 1993



Exploring the athletic univariate data of bowl, btitle and finfour, showing a stats summary and density plot

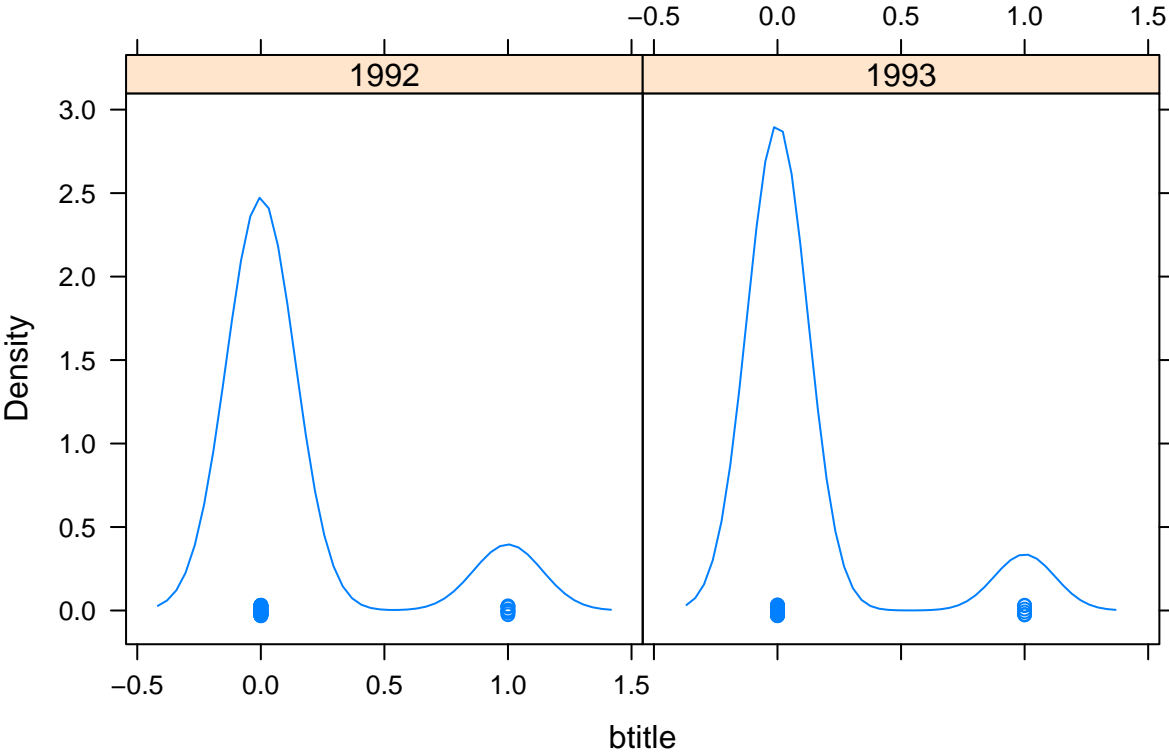
for each by year:

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	0.0000	0.0000	0.0000	0.4655	1.0000	1.0000



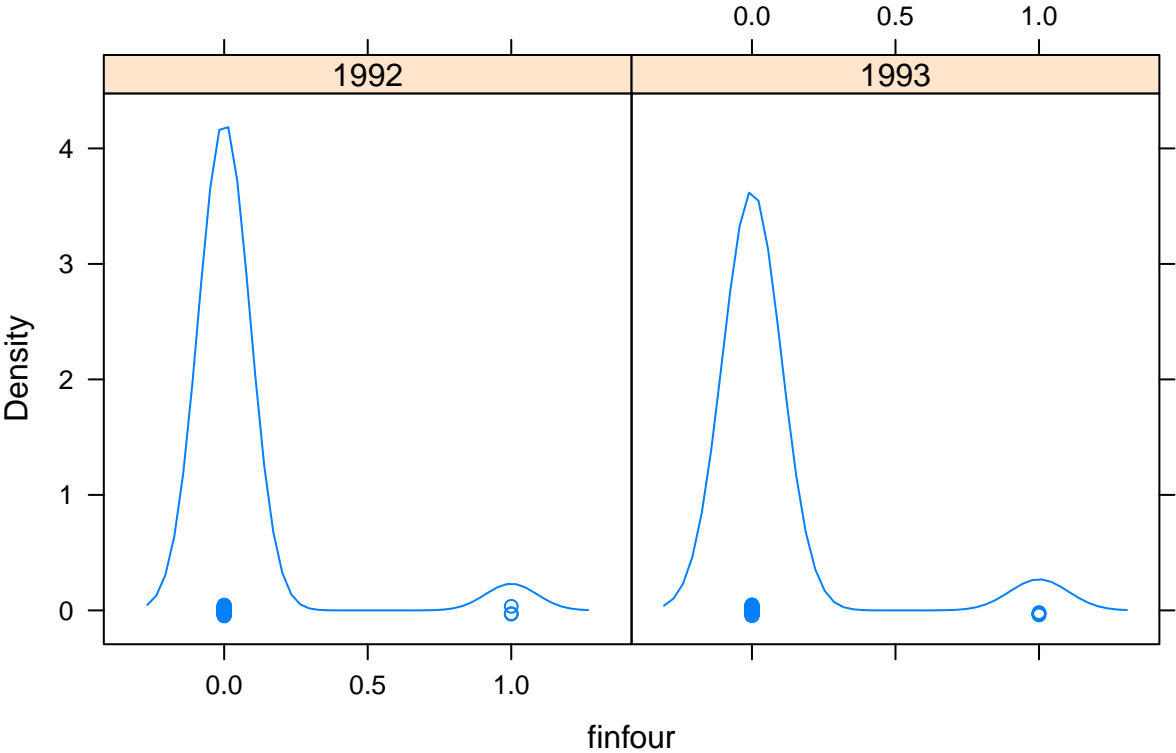
##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	0.0000	0.0000	0.0000	0.1207	0.0000	1.0000

Density Plot of Previous Year Conf. Champs



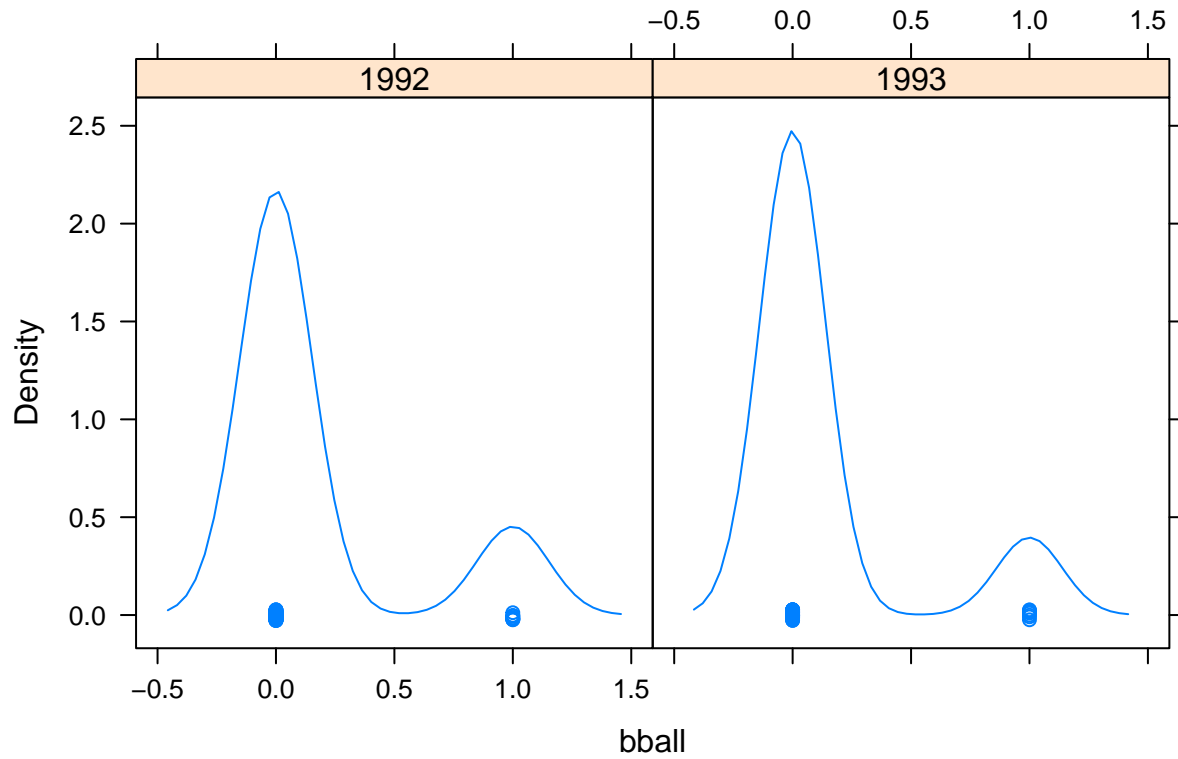
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.06034 0.00000 1.00000
```

Density Plot of Previous Year Final Four

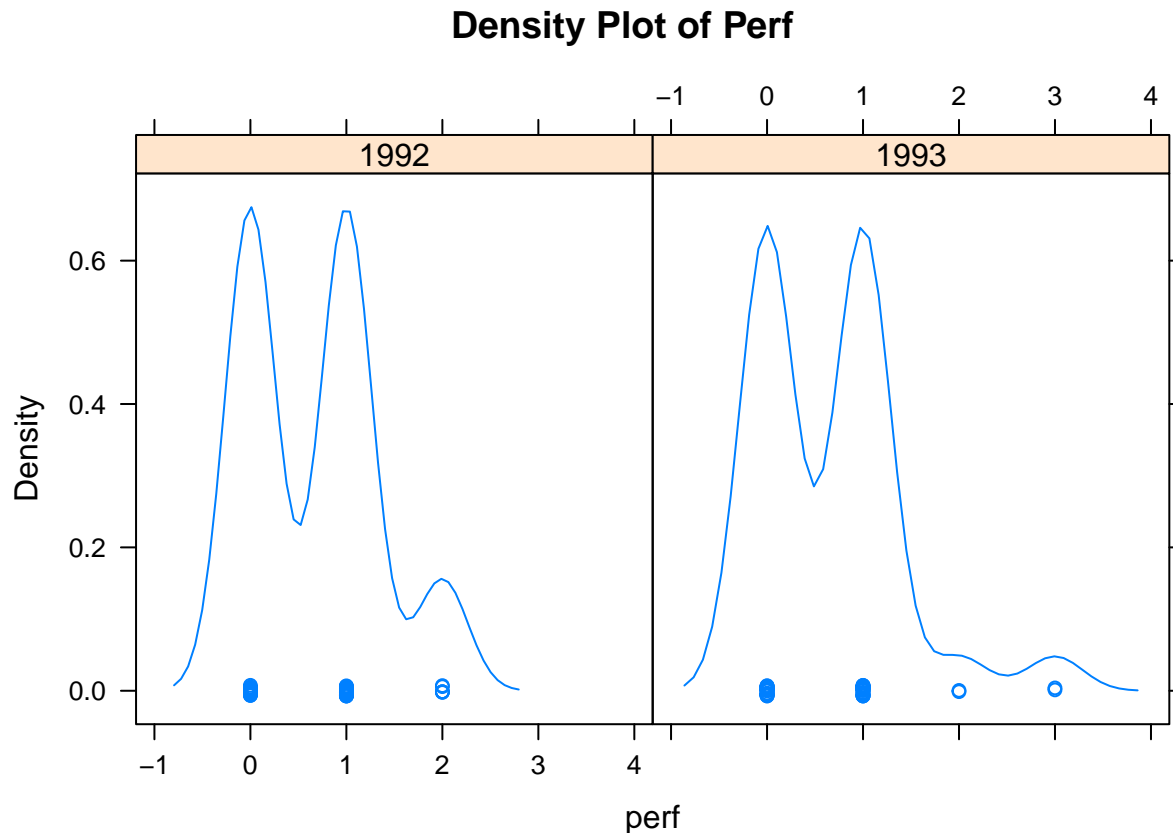


##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	0.0000	0.0000	0.0000	0.1552	0.0000	1.0000

Density Plot of Previous Year Conference & Final Four



##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	0.0000	0.0000	1.0000	0.6466	1.0000	3.0000



The bowl athletic success variable appears relatively even across the two years but the btitle and finfour variables are heavily skewed to 1992. This matches the expectation as set out in Question 1 in that previous year's athletic indicators are more apparent for the current year's applications.

Question 2:

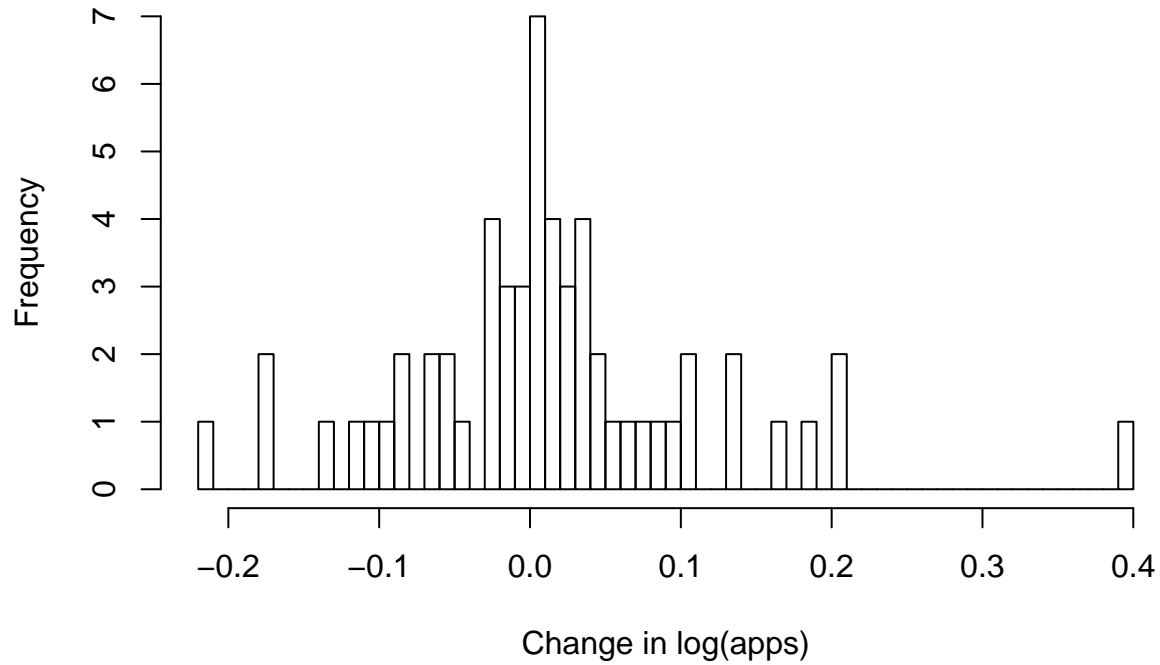
Note that the dataset is in long format, with a separate row for each year for each school. To prepare for a difference-in-difference analysis, transfer the dataset to wide-format. Each school should have a single row of data, with separate variables for 1992 and 1993. For example, you should have an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

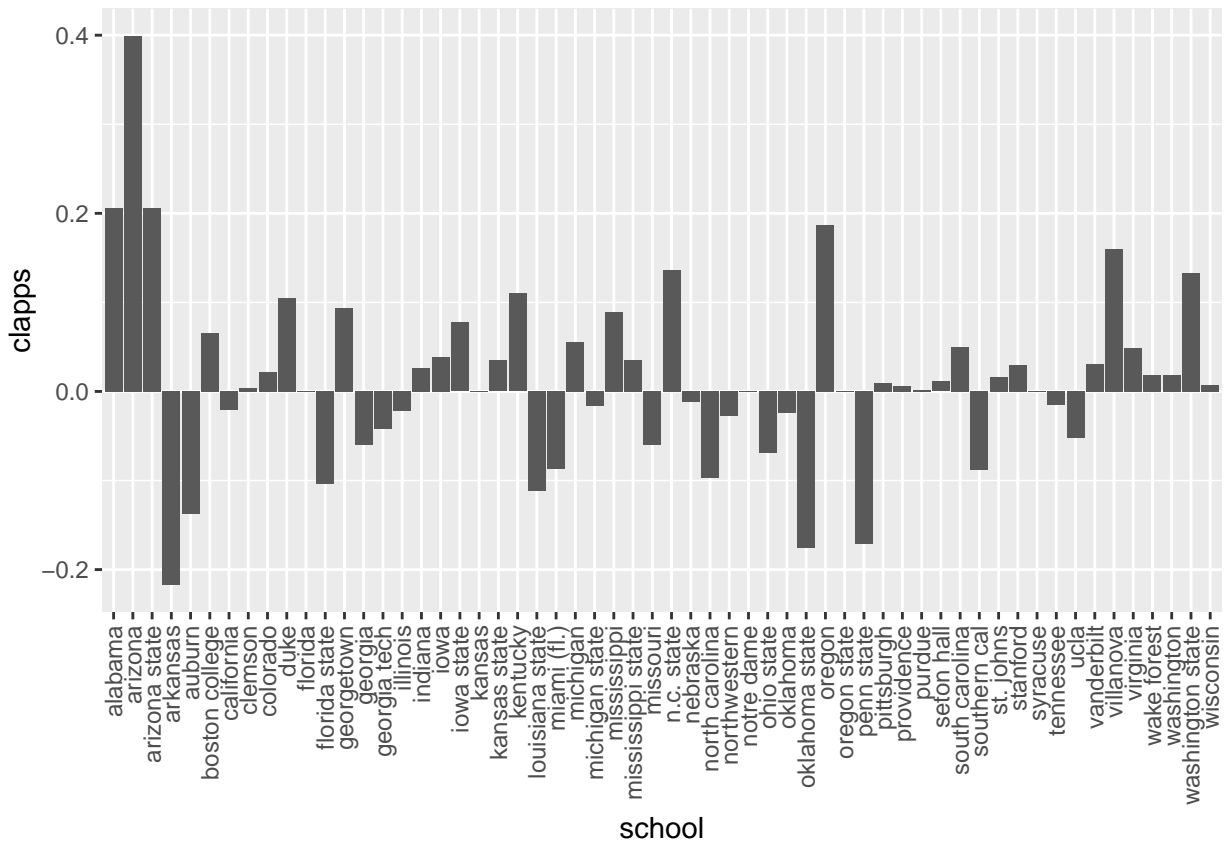
```
dw <- reshape(data, idvar='school', timevar='year', direction='wide',
              v.names=c('apps', 'top25', 'ver500', 'mth500',
                        'stufac', 'bowl', 'btitle', 'finfour',
                        'lapps', 'avg500', 'bball', 'perf'))
dw$clapps <- dw$lapps.1993 - dw$lapps.1992
summary(dw$clapps)
```

```
##      Min.   1st Qu.   Median     Mean   3rd Qu.    Max.
## -0.216900 -0.026630  0.006774  0.014210  0.049490  0.398900
```

Histogram of Change in log(apps) from 1992 to 1993



Warning: Stacking not well defined when ymin != 0



Arizona had the greatest increase while Arkansas had the greatest decrease in applications.

Question 3

Similarly to above, create three variables, *cperf*, *cball*, 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?

```
dw$cperf <- dw$perf.1993 - dw$perf.1992
dw$cball <- dw$bball.1993 - dw$bball.1992
dw$cbowl <- dw$bowl.1993 - dw$bowl.1992
dw$cbtitle <- dw$btitle.1993 - dw$btitle.1992
dw$cfinfour <- dw$finfour.1993 - dw$finfour.1992
```

Summary statistics and variances are calculated as:

```
summary(dw$cperf)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.00000  0.00000  0.00000 -0.01724  0.00000  3.00000
```

```
summary(dw$cball)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.00000  0.00000  0.00000 -0.03448  0.00000  1.00000
```

```
summary(dw$cbowl)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      -1         0         0         0         0         1
```

```
var(dw[c('cperf', 'cball', 'cbowl')])
```

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

From the variance calculation on each of the new variables it appears that the *cperf* variable has the highest variance. However, the *cperf* variance is not negatively correlated with the component years like *cball* and *cbowl*:

```
var(dw[c('cperf', 'perf.1992', 'perf.1993')])
```

```
##               cperf   perf.1992  perf.1993
## cperf         0.8242589 -0.37447066 0.44978826
## perf.1992 -0.3744707  0.44041137 0.06594071
## perf.1993  0.4497883  0.06594071 0.51572898
```

There is a negative covariance between *cball* and *bball.1992*

```
var(dw[c('cball', 'bball.1992', 'bball.1993')]))
```

```
##               cball   bball.1992  bball.1993
## cball         0.17422868 -0.09921355 0.07501512
## bball.1992 -0.09921355  0.14519056 0.04597701
## bball.1993  0.07501512  0.04597701 0.12099214
```

There is a negative covariance between *cbowl* and *bowl.1992*:

```
var(dw[c('cbowl', 'bowl.1992', 'bowl.1993')]))
```

```
##               cbowl   bowl.1992  bowl.1993
## cbowl         0.3157895 -0.15789474 0.15789474
## bowl.1992 -0.1578947  0.25317604 0.09528131
## bowl.1993  0.1578947  0.09528131 0.25317604
```

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.

The difference equation contains the a_i term, which is probably a mistake since a_i should be removed by taking the differences between the equation

The additional assumption for causality is

$$\text{cov}(\Delta u, \Delta \mathbf{X}_i) = 0$$

or stated differently:

$$E(\Delta u_i \mid \text{cbowl}, \text{cbtitle}, \text{cfinfour}) = 0$$

There must be no correlation between the residual error term and any of the random variables. If there is correlation between Δu and $\Delta \mathbf{X}_i$ then there will be bias in the estimators.

- 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.

The additional assumption of OLS to consistently estimate the first difference model is there must be some variance in the random variables across the time periods.

$$\text{var}(\Delta \mathbf{X}_i) > 0$$

Question 5

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

```
modell1 <- lm(clapps ~ cbowl + cbtitle + cfinfour, data=dw)
summary(modell1)
```

```
##
## Call:
## lm(formula = clapps ~ cbowl + cbtitle + cfinfour, data = dw)
##
## 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 β_1 coefficient is the only statistically significant coefficient, indicating that an increase in chance of a bowl game over the previous two years results in an increase of **5.7%** in the number of applications received.

The β_2 and β_3 coefficients are not statistically significant, and indeed the change in finfour has a negative correlation with application rates, which is not what is expected.

We check the joint hypothesis that : $H_0 : \beta_2, \beta_3 = 0$ using the F test between the unrestricted model RSS and the restricted model RSS.

```
modell1.r <- lm(clapps ~ cbowl, data=dw)
rss.r = sum(residuals(modell1.r)^2)
rss.ur = sum(residuals(modell1)^2)

linearHypothesis(modell1, c("cbtitle", "cfinfour"), vcov=vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## cbtitle = 0
## cfinfour = 0
##
## Model 1: restricted model
## Model 2: clapps ~ cbowl + cbtitle + cfinfour
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df       F Pr(>F)
## 1      56
## 2      54  2 0.9207 0.4044
```

We can't reject the null hypothesis that β_2 and β_3 are statistically different than 0.

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?

```
waldtest(modell1, test='F', vcov=vcovHC)

## Wald test
##
## Model 1: clapps ~ cbowl + cbtitle + cfinfour
## Model 2: clapps ~ 1
##   Res.Df Df       F Pr(>F)
## 1      54
## 2      57 -3 1.4717 0.2325
```

The result of the Wald test for joint significance of the entire model shows that the model is basically crap.