# Topic 2: Exercise 1

Daniel Alonso

November 28th, 2020

### Importing libraries

```r
library(dplyr)
```

### Importing data as described by exercise

```r
d <- read.csv("../../datasets/Colleges.csv")
```

### Replacing binary variable Private with 1 and 0

```r
d$Private <- ifelse(d$Private == "Yes", 1, 0)
```

### Selecting columns

```r
data <- d %>% dplyr::select('Private','Apps','Accept','Enroll','F.Undergrad')
```

### Calculating covariances

```r
cov_matrix <- cov(data)
cov_matrix
```

```
##                   Private          Apps        Accept       Enroll  F.Undergrad
## Private         0.1986559     -745.3552     -519.2042    -235.1942    -1330.764
## Apps         -745.3552439 14978459.5301  8949859.8119 3045255.9876 15289702.474
## Accept       -519.2042169  8949859.8119  6007959.6988 2076267.7627 10393582.435
## Enroll       -235.1942393  3045255.9876  2076267.7627  863368.3923  4347529.884
## F.Undergrad -1330.7637175 15289702.4742 10393582.4355 4347529.8841 23526579.326
```

### Calculating correlations

```
corr_matrix <- cov2cor(cov_matrix)
corr_matrix
```

```
##                 Private       Apps     Accept     Enroll F.Undergrad
## Private       1.0000000 -0.4320947 -0.4752520 -0.5679078  -0.6155605
## Apps         -0.4320947  1.0000000  0.9434506  0.8468221   0.8144906
## Accept       -0.4752520  0.9434506  1.0000000  0.9116367   0.8742233
## Enroll       -0.5679078  0.8468221  0.9116367  1.0000000   0.9646397
## F.Undergrad -0.6155605  0.8144906  0.8742233  0.9646397   1.0000000
```

### Experimenting a little bit with the private variable

Let's try changing the Yes to 0 and the No to 1 and checking the covariances and correlations

```
d <- read.csv("../../datasets/Colleges.csv")
d$Private <- ifelse(d$Private == "Yes", 0, 1)
data <- d %>% dplyr::select('Private','Apps','Accept','Enroll','F.Undergrad')
```

```
cov_matrix <- cov(data)
cov_matrix
```

```
##                   Private         Apps       Accept       Enroll  F.Undergrad
## Private         0.1986559 7.453552e+02 5.192042e+02     235.1942     1330.764
## Apps          745.3552439 1.497846e+07 8.949860e+06 3045255.9876 15289702.474
## Accept        519.2042169 8.949860e+06 6.007960e+06 2076267.7627 10393582.435
## Enroll        235.1942393 3.045256e+06 2.076268e+06  863368.3923  4347529.884
## F.Undergrad 1330.7637175 1.528970e+07 1.039358e+07 4347529.8841 23526579.326
```

```
corr_matrix <- cov2cor(cov_matrix)
corr_matrix
```

```
##                Private      Apps    Accept    Enroll F.Undergrad
## Private      1.0000000 0.4320947 0.4752520 0.5679078   0.6155605
## Apps         0.4320947 1.0000000 0.9434506 0.8468221   0.8144906
## Accept       0.4752520 0.9434506 1.0000000 0.9116367   0.8742233
## Enroll       0.5679078 0.8468221 0.9116367 1.0000000   0.9646397
## F.Undergrad 0.6155605 0.8144906 0.8742233 0.9646397   1.0000000
```

We get the same numbers with reversed signs.

Let's try having the same amount of 1s and 0s and see how correlation and covariance change:

```
data$Private <- c(rep(0,length(data$Private)/2 + 1), rep(1,length(data$Private)/2))
```

```
cov_matrix <- cov(data)
cov_matrix
```

```
##                  Private         Apps       Accept       Enroll  F.Undergrad
## Private       0.2503218 3.387465e+02 2.189999e+02 8.109728e+01 4.358092e+02
## Apps        338.7465453 1.497846e+07 8.949860e+06 3.045256e+06 1.528970e+07
## Accept      218.9998740 8.949860e+06 6.007960e+06 2.076268e+06 1.039358e+07
## Enroll       81.0972764 3.045256e+06 2.076268e+06 8.633684e+05 4.347530e+06
## F.Undergrad 435.8092186 1.528970e+07 1.039358e+07 4.347530e+06 2.352658e+07
```

```
corr_matrix <- cov2cor(cov_matrix)
corr_matrix
```

```
##               Private      Apps    Accept    Enroll F.Undergrad
## Private     1.0000000 0.1749412 0.1785793 0.1744452   0.1795840
## Apps        0.1749412 1.0000000 0.9434506 0.8468221   0.8144906
## Accept      0.1785793 0.9434506 1.0000000 0.9116367   0.8742233
## Enroll      0.1744452 0.8468221 0.9116367 1.0000000   0.9646397
## F.Undergrad 0.1795840 0.8144906 0.8742233 0.9646397   1.0000000
```

## What information does the sample covariance provide?

We know that because the Private variable (binary variable) has only 2 possible values, its covariance with other variables is always going to be relatively small and will not provide much information.

## What information does the sample correlation provide?

### Scatter plot of our quantitative variables and the Private binary variable

```
model = lm(Private ~ Apps, data=d)
summary(model)
```

```
##
## Call:
## lm(formula = Private ~ Apps, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5167 -0.2087 -0.1613  0.1581  0.8649
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.235e-01  1.826e-02    6.76 2.71e-11 ***
## Apps        4.976e-05  3.731e-06   13.34  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4022 on 775 degrees of freedom
## Multiple R-squared:  0.1867, Adjusted R-squared:  0.1857
## F-statistic: 177.9 on 1 and 775 DF,  p-value: < 2.2e-16
```