

## Lab 7

Jaime Lin

11:59PM April 22, 2021

#Rcpp

We will get some experience with speeding up R code using C++ via the Rcpp package.

First, clear the workspace and load the Rcpp package.

```
pacman::p_load(Rcpp)
```

Create a variable `n` to be 10 and a variable `Nvec` to be 100 initially. Create a random vector via `rnorm` `Nvec` times and load it into a `Nvec` x `n` dimensional matrix.

```
n = 10
Nvec = 100
X = matrix(data = rnorm(Nvec), nrow = 100, ncol = 10)
head(X)

##           [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## [1,]  0.2635986  0.2635986  0.2635986  0.2635986  0.2635986  0.2635986
## [2,] -0.5530956 -0.5530956 -0.5530956 -0.5530956 -0.5530956 -0.5530956
## [3,]  1.0264043  1.0264043  1.0264043  1.0264043  1.0264043  1.0264043
## [4,] -0.7303913 -0.7303913 -0.7303913 -0.7303913 -0.7303913 -0.7303913
## [5,] -1.2155867 -1.2155867 -1.2155867 -1.2155867 -1.2155867 -1.2155867
## [6,] -0.9892604 -0.9892604 -0.9892604 -0.9892604 -0.9892604 -0.9892604
##           [,7]      [,8]      [,9]     [,10]
## [1,]  0.2635986  0.2635986  0.2635986  0.2635986
## [2,] -0.5530956 -0.5530956 -0.5530956 -0.5530956
## [3,]  1.0264043  1.0264043  1.0264043  1.0264043
## [4,] -0.7303913 -0.7303913 -0.7303913 -0.7303913
## [5,] -1.2155867 -1.2155867 -1.2155867 -1.2155867
## [6,] -0.9892604 -0.9892604 -0.9892604 -0.9892604
```

Write a function `all_angles` that measures the angle between each of the pairs of vectors. You should measure the vector on a scale of 0 to 180 degrees with negative angles coerced to be positive.

```
angle = function(u,v){
  acos(sum(u*v)/sqrt(sum(u^2)*sum(v^2))) * (180/pi)
}
all_angles = function(X){
  A = matrix(NA, nrow = nrow(X), ncol = nrow(X))
  for (i in 1:(nrow(X)-1)) {
    for (j in (i+1):nrow(X)) {
      A[i,j] = angle(X[i,], X[j,])
    }
  }
}
```

```

    }
  }
  A
}
## all_angles(X)

```

Plot the density of these angles.

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

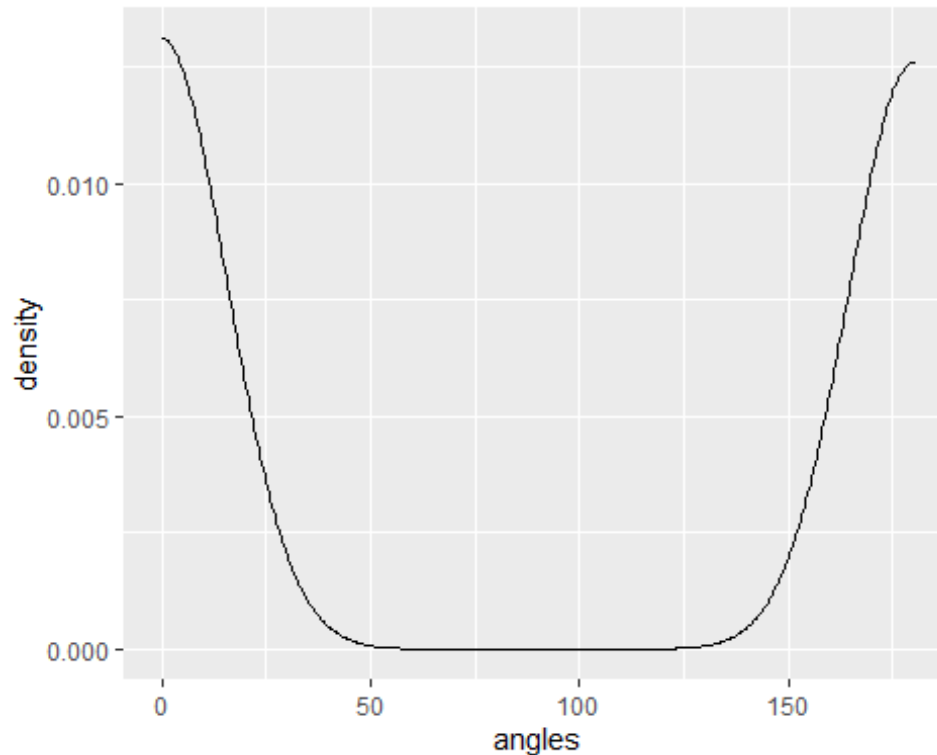
[illegible]

[illegible]

[illegible]

[illegible]

```
## Warning in acos(sum(u * v)/sqrt(sum(u^2) * sum(v^2))): NaNs produced
## Warning in acos(sum(u * v)/sqrt(sum(u^2) * sum(v^2))): NaNs produced
## Warning in acos(sum(u * v)/sqrt(sum(u^2) * sum(v^2))): NaNs produced
## Warning in acos(sum(u * v)/sqrt(sum(u^2) * sum(v^2))): NaNs produced
## Warning: Removed 6128 rows containing non-finite values (stat_density).
```



Write an Rcpp function `all_angles_cpp` that does the same thing. Use an IDE if you want, but write it below in-line.

```
"
cppFunction(
  ,
  NumericMatrix all_angles_cpp(NumericMatrix X) {
    int n = X.nrow();
    int p = X.ncol();
    NumericMatrix A(n, n);
    std::fill(A.begin(), A.end(), NA_REAL);
    for (int i_1 = 0; i_1 < (n - 1); i_1++){

      for (int i_2 = i_1 + 1; i_2 < n; i_2++){
        double sum_sqd_u = 0;
        double sum_sqd_v = 0;
        double sum_u_v = 0;
```

```

        for (int j = 0; j < p; j++){
            sum_sqd_u += pow(X(i_1, j), 2);
            sum_sqd_v += pow(X(i_2, j), 2);
            sum_u_v += X(i_1, j) * X(i_2, j);
        }
        A(i_1, i_2) = acos(sum_u_v)/sqrt((sum_sqd_u)*(sum_sqd_v)) *
(180/M_PI);
    }
}
return A;
}
,
)
all_angles_cpp(X)
"

## [1] "\ncppFunction(\n  ' \n  NumericMatrix all_angles_cpp(NumericMatrix X)
{\n    int n = X.nrow();\n    int p = X.ncol();\n    NumericMatrix A(n, n);\n
std::fill(A.begin(), A.end(), NA_REAL);\n    for (int i_1 = 0; i_1 < (n - 1);
i_1++){ \n\n      for (int i_2 = i_1 + 1; i_2 < n; i_2++){ \n          double
sum_sqd_u = 0;\n          double sum_sqd_v = 0;\n          double sum_u_v = 0;\n
for (int j = 0; j < p; j++){ \n              sum_sqd_u += pow(X(i_1, j), 2);\n
sum_sqd_v += pow(X(i_2, j), 2);\n              sum_u_v += X(i_1, j) * X(i_2, j);\n
\n          } \n          A(i_1, i_2) =
acos(sum_u_v)/sqrt((sum_sqd_u)*(sum_sqd_v)) * (180/M_PI); \n          } \n      } \n
return A;\n  } \n  '\n)\nall_angles_cpp(X)\n"

```

Test the time difference between these functions for  $n = 1000$  and  $N_{\text{vec}} = 100, 500, 1000, 5000$  using the package `microbenchmark`. Store the results in a matrix with rows representing  $N_{\text{vec}}$  and two columns for base R and Rcpp.

```

pacman::p_load(microbenchmark)
n = 1000
Nvec = c(100, 500, 1000, 5000)
"
  for (i in 1:4) {
    X = matrix(data = rnorm(Nvec[i]), nrow = Nvec[i])
    microbenchmark(all_angles(X), all_angles_cpp(X), times = 10)
  }
"

## [1] "\n  for (i in 1:4) {\n    X = matrix(data = rnorm(Nvec[i]), nrow =
Nvec[i])\n    microbenchmark(all_angles(X), all_angles_cpp(X), times = 10)\n
}\n"

```

Plot the divergence of performance (in log seconds) over  $n$  using a line geometry. Use two different colors for the R and CPP functions. Make sure there's a color legend on your plot. We will see later how to create "long" matrices that make such plots easier.



Let `Nvec = 10000` and vary `n` to be 10, 100, 1000. Plot the density of angles for all three values of `n` on one plot using color to signify `n`. Make sure you have a color legend. This is not easy.

```
n = c(10, 100, 1000)
Nvec = 10000
" for (i in 1:4) {
  X = matrix(data = rnorm(Nvec), nrow = Nvec)
  microbenchmark(all_angles(X), all_angles_cpp(X), times = n[i])
}
"
```

```
## [1] " for (i in 1:4) {\n  X = matrix(data = rnorm(Nvec), nrow = Nvec)\n\n  microbenchmark(all_angles(X), all_angles_cpp(X), times = n[i])\n}\n "
```

Write an R function `nth_fibonnaci` that finds the `nth` Fibonnaci number via recursion but allows you to specify the starting number. For instance, if the sequeency started at 1, you get the familiar 1, 1, 2, 3, 5, etc. But if it started at 0.01, you would get 0.01, 0.01, 0.02, 0.03, 0.05, etc.

```
f = c(NA)
f[1]=0.01
f[2]=0.01
nth_fibonnaci = function(n){
  for (i in 3:n) {
    f[i] = f[i-2]+ f[i-1]
  }
  f[n]
}
nth_fibonnaci(6)

## [1] 0.08
```

Write an Rcpp function `nth_fibonnaci_cpp` that does the same thing. Use an IDE if ou want, but write it below in-line.

```
"
cppFunction(
  'vector<int> f;
  f[0] = 0.01;
  f[1] = 0.01;
  NumericMatrix nth_fibonnaci_cpp(NumericMatrix n) {
    for(i = 2; i<=n; i++){
      f[i] = f[i-2]+ f[i-1];
    }
    f[n];
  }')
"
```

```
## [1] "\ncppFunction(\n  'vector<int> f;\n  f[0] = 0.01;\n  f[1] = 0.01;\n  NumericMatrix nth_fibonacci_cpp(NumericMatrix n) {\n    for(i = 2; i<=n;\n    i++){ \n      f[i] = f[i-2]+ f[i-1];\n    } \n    f[n];\n  }')\n"
```

Time the difference in these functions for  $n = 100, 200, \dots, 1500$  while starting the sequence at the smallest possible floating point value in R. Store the results in a matrix.

```
f = c(NA)
A = c(NA)
f[1]=0.01
f[2]=0.01

nth_fibonacci = function(n){

  for (i in 3:n) {
    f[i] = f[i-2]+ f[i-1]
  }

  A[i-2] = f[100]

}
```

Plot the divergence of performance (in log seconds) over  $n$  using a line geometry. Use two different colors for the R and CPP functions. Make sure there's a color legend on your plot.

*#TO-DO*

## Data Wrangling / Munging / Carpentry

Throughout this assignment you can use either the tidyverse package suite or `data.table` to answer but not base R. You can mix `data.table` with `magrittr` piping if you wish but don't go back and forth between `tbl_df`'s and `data.table` objects.

```
pacman::p_load(dplyr, magrittr, data.table)
```

Load the storms dataset from the `dplyr` package and investigate it using `str` and `summary` and `head`. Which two columns should be converted to type factor? Do so below.

```
data(storms)
str(storms)

## tibble [10,010 x 13] (S3: tbl_df/tbl/data.frame)
##  $ name      : chr [1:10010] "Amy" "Amy" "Amy" "Amy" ...
##  $ year      : num [1:10010] 1975 1975 1975 1975 1975 ...
##  $ month     : num [1:10010] 6 6 6 6 6 6 6 6 6 6 ...
##  $ day       : int [1:10010] 27 27 27 27 28 28 28 28 29 29 ...
##  $ hour      : num [1:10010] 0 6 12 18 0 6 12 18 0 6 ...
##  $ lat       : num [1:10010] 27.5 28.5 29.5 30.5 31.5 32.4 33.3 34 34.4
##  $ long      : num [1:10010] -79 -79 -79 -79 -78.8 -78.7 -78 -77 -75.8 -
```

```

74.8 ...
## $ status      : chr [1:10010] "tropical depression" "tropical depression"
"tropical depression" "tropical depression" ...
## $ category    : Ord.factor w/ 7 levels "-1"<"0"<"1"<"2"<...: 1 1 1 1 1 1 1
1 2 2 ...
## $ wind        : int [1:10010] 25 25 25 25 25 25 25 30 35 40 ...
## $ pressure    : int [1:10010] 1013 1013 1013 1013 1012 1012 1011 1006 1004
1002 ...
## $ ts_diameter: num [1:10010] NA NA NA NA NA NA NA NA NA NA ...
## $ hu_diameter: num [1:10010] NA NA NA NA NA NA NA NA NA NA ...

```

```
summary(storms)
```

```

##      name          year      month      day
## Length:10010      Min.   :1975      Min.   : 1.000      Min.   : 1.00
## Class :character  1st Qu.:1990      1st Qu.: 8.000      1st Qu.: 8.00
## Mode  :character  Median :1999      Median : 9.000      Median :16.00
##                      Mean  :1998      Mean  : 8.779      Mean  :15.86
##                      3rd Qu.:2006      3rd Qu.: 9.000      3rd Qu.:24.00
##                      Max.   :2015      Max.   :12.000      Max.   :31.00
##
##      hour          lat      long      status
## Min.   : 0.000      Min.   : 7.20      Min.   : -109.30      Length:10010
## 1st Qu.: 6.000      1st Qu.:17.50      1st Qu.: -80.70      Class :character
## Median :12.000      Median :24.40      Median : -64.50      Mode  :character
## Mean   : 9.114      Mean   :24.76      Mean   : -64.23
## 3rd Qu.:18.000      3rd Qu.:31.30      3rd Qu.: -48.60
## Max.   :23.000      Max.   :51.90      Max.   : -6.00
##
## category      wind      pressure      ts_diameter
hu_diameter
## -1:2545      Min.   : 10.00      Min.   : 882.0      Min.   :  0.00      Min.   :
0.00
##  0 :4373      1st Qu.: 30.00      1st Qu.: 985.0      1st Qu.:  69.05      1st Qu.:
0.00
##  1 :1685      Median : 45.00      Median : 999.0      Median : 138.09      Median :
0.00
##  2 : 628      Mean   : 53.49      Mean   : 992.1      Mean   : 166.76      Mean   :
21.41
##  3 : 363      3rd Qu.: 65.00      3rd Qu.:1006.0      3rd Qu.: 241.66      3rd Qu.:
28.77
##  4 : 348      Max.   :160.00      Max.   :1022.0      Max.   :1001.18      Max.
:345.23
##  5 :  68                      NA's   :6528      NA's
:6528

```

```
head(storms)
```

```

## # A tibble: 6 x 13
##   name year month day hour lat long status category wind
pressure

```

```
##   <chr> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <chr>          <ord>    <int>
<int>
## 1 Amy    1975     6    27     0  27.5 -79   tropical de~ -1      25
1013
## 2 Amy    1975     6    27     6  28.5 -79   tropical de~ -1      25
1013
## 3 Amy    1975     6    27    12  29.5 -79   tropical de~ -1      25
1013
## 4 Amy    1975     6    27    18  30.5 -79   tropical de~ -1      25
1013
## 5 Amy    1975     6    28     0  31.5 -78.8 tropical de~ -1      25
1012
## 6 Amy    1975     6    28     6  32.4 -78.7 tropical de~ -1      25
1012
## # ... with 2 more variables: ts_diameter <dbl>, hu_diameter <dbl>
```

Reorder the columns so name is first, status is second, category is third and the rest are the same.

```
storms %>%
  select(name, status, category, everything())

## # A tibble: 10,010 x 13
##   name status      category year month  day  hour  lat  long  wind
pressure
##   <chr> <chr>          <ord>    <dbl> <dbl> <int> <dbl> <dbl> <dbl> <int>
<int>
## 1 Amy    tropical d~ -1      1975     6    27     0  27.5 -79      25
1013
## 2 Amy    tropical d~ -1      1975     6    27     6  28.5 -79      25
1013
## 3 Amy    tropical d~ -1      1975     6    27    12  29.5 -79      25
1013
## 4 Amy    tropical d~ -1      1975     6    27    18  30.5 -79      25
1013
## 5 Amy    tropical d~ -1      1975     6    28     0  31.5 -78.8    25
1012
## 6 Amy    tropical d~ -1      1975     6    28     6  32.4 -78.7    25
1012
## 7 Amy    tropical d~ -1      1975     6    28    12  33.3 -78      25
1011
## 8 Amy    tropical d~ -1      1975     6    28    18  34   -77      30
1006
## 9 Amy    tropical s~ 0       1975     6    29     0  34.4 -75.8    35
1004
## 10 Amy   tropical s~ 0       1975     6    29     6  34   -74.8    40
1002
## # ... with 10,000 more rows, and 2 more variables: ts_diameter <dbl>,
## #   hu_diameter <dbl>
```

Find a subset of the data of storms only in the 1970's.

```

storms %>%
  filter(year >= 1970 & year <= 1979)

## # A tibble: 546 x 13
##   name    year month   day  hour   lat   long status    category  wind
pressure
##   <chr> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <chr>      <ord>    <int>
<int>
##  1 Amy    1975     6    27     0  27.5 -79   tropical d~ -1      25
1013
##  2 Amy    1975     6    27     6  28.5 -79   tropical d~ -1      25
1013
##  3 Amy    1975     6    27    12  29.5 -79   tropical d~ -1      25
1013
##  4 Amy    1975     6    27    18  30.5 -79   tropical d~ -1      25
1013
##  5 Amy    1975     6    28     0  31.5 -78.8 tropical d~ -1      25
1012
##  6 Amy    1975     6    28     6  32.4 -78.7 tropical d~ -1      25
1012
##  7 Amy    1975     6    28    12  33.3 -78   tropical d~ -1      25
1011
##  8 Amy    1975     6    28    18  34    -77   tropical d~ -1      30
1006
##  9 Amy    1975     6    29     0  34.4 -75.8 tropical s~ 0       35
1004
## 10 Amy    1975     6    29     6  34    -74.8 tropical s~ 0       40
1002
## # ... with 536 more rows, and 2 more variables: ts_diameter <dbl>,
## #   hu_diameter <dbl>

```

Find a subset of the data of storm observations only with category 4 and above and wind speed 100MPH and above.

```

storms %>%
  filter(category >= 4 & wind >= 100)

## # A tibble: 416 x 13
##   name    year month   day  hour   lat   long status    category  wind
pressure
##   <chr> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <chr>      <ord>    <int>
<int>
##  1 Anita  1977     9     2     0  24.6 -96.2 hurricane 5      140
931
##  2 Anita  1977     9     2     6  24.2 -97.1 hurricane 5      150
926
##  3 Anita  1977     9     2    12  23.7 -98   hurricane 4      120
940
##  4 David  1979     8    28     0  12.2 -52.9 hurricane 4      115
947
##  5 David  1979     8    28     6  12.5 -54.4 hurricane 4      125

```

```

941
## 6 David 1979 8 28 12 12.8 -55.7 hurricane 4 130
938
## 7 David 1979 8 28 18 13.2 -56.9 hurricane 4 125
941
## 8 David 1979 8 29 0 13.7 -58 hurricane 4 120
944
## 9 David 1979 8 29 6 14.2 -59.2 hurricane 4 120
942
## 10 David 1979 8 29 12 14.8 -60.3 hurricane 4 125
938
## # ... with 406 more rows, and 2 more variables: ts_diameter <dbl>,
## # hu_diameter <dbl>

```

Create a new feature `wind_speed_per_unit_pressure`.

```

storms %>%
  mutate(wind_speed_per_unit_pressure = wind / pressure)

## # A tibble: 10,010 x 14
##   name year month day hour lat long status category wind
##   <chr> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <chr> <ord> <int>
##   <int>
## 1 Amy 1975 6 27 0 27.5 -79 tropical d~ -1 25
1013
## 2 Amy 1975 6 27 6 28.5 -79 tropical d~ -1 25
1013
## 3 Amy 1975 6 27 12 29.5 -79 tropical d~ -1 25
1013
## 4 Amy 1975 6 27 18 30.5 -79 tropical d~ -1 25
1013
## 5 Amy 1975 6 28 0 31.5 -78.8 tropical d~ -1 25
1012
## 6 Amy 1975 6 28 6 32.4 -78.7 tropical d~ -1 25
1012
## 7 Amy 1975 6 28 12 33.3 -78 tropical d~ -1 25
1011
## 8 Amy 1975 6 28 18 34 -77 tropical d~ -1 30
1006
## 9 Amy 1975 6 29 0 34.4 -75.8 tropical s~ 0 35
1004
## 10 Amy 1975 6 29 6 34 -74.8 tropical s~ 0 40
1002
## # ... with 10,000 more rows, and 3 more variables: ts_diameter <dbl>,
## # hu_diameter <dbl>, wind_speed_per_unit_pressure <dbl>

```

Create a new feature: `average_diameter` which averages the two diameter metrics. If one is missing, then use the value of the one that is present. If both are missing, leave missing.

```

storms %>%
  rowwise() %>%
  arrange(desc(year)) %>%
  mutate(average_diameter = mean(c(ts_diameter, hu_diameter), na.rm = TRUE))

## # A tibble: 10,010 x 14
## # Rowwise:
##   name year month day hour lat long status category wind
pressure
##   <chr> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <chr> <ord> <int>
<int>
## 1 Ana 2015 5 9 6 32.2 -77.5 tropical s~ 0 50
998
## 2 Ana 2015 5 9 12 32.5 -77.8 tropical s~ 0 50
1001
## 3 Ana 2015 5 9 18 32.7 -78 tropical s~ 0 45
1001
## 4 Ana 2015 5 10 0 33.1 -78.3 tropical s~ 0 45
1001
## 5 Ana 2015 5 10 6 33.5 -78.6 tropical s~ 0 40
1002
## 6 Ana 2015 5 10 10 33.8 -78.8 tropical s~ 0 40
1002
## 7 Ana 2015 5 10 12 33.9 -78.8 tropical s~ 0 35
1002
## 8 Ana 2015 5 10 18 34.3 -78.7 tropical d~ -1 30
1006
## 9 Ana 2015 5 11 0 34.7 -78.5 tropical d~ -1 30
1009
## 10 Ana 2015 5 11 6 35.5 -78 tropical d~ -1 30
1010
## # ... with 10,000 more rows, and 3 more variables: ts_diameter <dbl>,
## # hu_diameter <dbl>, average_diameter <dbl>

```

For each storm, summarize the maximum wind speed. “Summarize” means create a new dataframe with only the summary metrics you care about.

```

storms %>%
  group_by(name) %>%
  summarise(max_wind_speed = max(wind, na.rm = TRUE))

## # A tibble: 198 x 2
##   name max_wind_speed
##   * <chr> <int>
## 1 AL011993 30
## 2 AL012000 25
## 3 AL021992 30
## 4 AL021994 30
## 5 AL021999 30
## 6 AL022000 30
## 7 AL022001 25

```

```
## 8 AL022003          30
## 9 AL022006          45
## 10 AL031987         40
## # ... with 188 more rows
```

Order your dataset by maximum wind speed storm but within the rows of storm show the observations in time order from early to late.

```
storms %>%
  group_by(name) %>%
  mutate(max_wind_speed = max(wind, na.rm = TRUE)) %>%
  select(name, max_wind_speed, everything()) %>%
  arrange(desc(max_wind_speed), year, month, day, hour)

## # A tibble: 10,010 x 14
## # Groups:   name [198]
##   name      max_wind_speed year month   day hour   lat  long status
##   <chr>          <int> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <chr>
##   <ord>
## 1 Gilbe~          160  1988     9     8    18  12   -54 tropical ~ -1
## 2 Gilbe~          160  1988     9     9     0 12.7 -55.6 tropical ~ -1
## 3 Gilbe~          160  1988     9     9     6 13.3 -57.1 tropical ~ -1
## 4 Gilbe~          160  1988     9     9    12  14   -58.6 tropical ~ -1
## 5 Gilbe~          160  1988     9     9    18 14.5 -60.1 tropical ~ 0
## 6 Gilbe~          160  1988     9    10     0 14.8 -61.5 tropical ~ 0
## 7 Gilbe~          160  1988     9    10     6  15   -62.8 tropical ~ 0
## 8 Gilbe~          160  1988     9    10    12 15.3 -64.1 tropical ~ 0
## 9 Gilbe~          160  1988     9    10    18 15.7 -65.4 tropical ~ 0
## 10 Gilbe~         160  1988     9    11     0 15.9 -66.8 hurricane 1
## # ... with 10,000 more rows, and 4 more variables: wind <int>, pressure
## #   ts_diameter <dbl>, hu_diameter <dbl>
```

Find the strongest storm by wind speed per year.

```
storms %>%
  group_by(year) %>%
  arrange(year, desc(wind)) %>%
  slice(1) %>%
  select(name, year, wind)

## # A tibble: 41 x 3
## # Groups:   year [41]
##   name      year  wind
##   <chr>   <dbl> <int>
## 1 Caroline 1975   100
## 2 Belle    1976   105
## 3 Anita    1977   150
## 4 Cora     1978    80
## 5 David    1979   150
```



```
## 6 Ivan      1980    90
## 7 Harvey    1981   115
## 8 Debby     1982   115
## 9 Alicia    1983   100
## 10 Diana    1984   115
## # ... with 31 more rows
```

For each named storm, find its maximum category, wind speed, pressure and diameters. Do not allow the max to be NA (unless all the measurements for that storm were NA).

```
maximum_category = max(storms$category)
maximum_wind_speed = max(storms$wind)
maximum_pressure = max(storms$pressure)
maximum_ts_diameter = max(storms$ts_diameters)

## Warning: Unknown or uninitialised column: `ts_diameters`.

## Warning in max(storms$ts_diameters): no non-missing arguments to max;
## returning
## -Inf

maximum_hu_diameter = max(storms$hu_diameter)
maximum_hu_diameter

## [1] NA

maximum_pressure

## [1] 1022

maximum_wind_speed

## [1] 160

maximum_category

## [1] 5
## Levels: -1 < 0 < 1 < 2 < 3 < 4 < 5

maximum_ts_diameter

## [1] -Inf
```

For each year in the dataset, tally the number of storms. “Tally” is a fancy word for “count the number of”. Plot the number of storms by year. Any pattern?

```
storms %>%
  group_by(year) %>%
  tally()

## # A tibble: 41 x 2
##   year      n
##   * <dbl> <int>
```

```
## 1 1975 86
## 2 1976 52
## 3 1977 53
## 4 1978 54
## 5 1979 301
## 6 1980 161
## 7 1981 164
## 8 1982 105
## 9 1983 79
## 10 1984 236
## # ... with 31 more rows
```

For each year in the dataset, tally the storms by category.

```
storms %>%
  group_by(year, category) %>%
  tally()

## # A tibble: 233 x 3
## # Groups:   year [41]
##   year category     n
##   <dbl> <ord>   <int>
## 1 1975 -1       30
## 2 1975 0       33
## 3 1975 1       12
## 4 1975 2        9
## 5 1975 3        2
## 6 1976 -1      10
## 7 1976 0       20
## 8 1976 1       10
## 9 1976 2        9
## 10 1976 3        3
## # ... with 223 more rows
```

For each year in the dataset, find the maximum wind speed per status level.

```
storms %>%
  group_by(year, wind) %>%
  tally()

## # A tibble: 837 x 3
## # Groups:   year [41]
##   year wind     n
##   <dbl> <int> <int>
## 1 1975 20     2
## 2 1975 25    25
## 3 1975 30     3
## 4 1975 35     2
## 5 1975 40     2
## 6 1975 45     2
## 7 1975 50     7
```

```
## 8 1975 55 9
## 9 1975 60 11
## 10 1975 65 5
## # ... with 827 more rows
```

For each storm, summarize its average location in latitude / longitude coordinates.

```
storms %>%
  group_by(name, lat) %>%
  tally()

## # A tibble: 8,170 x 3
## # Groups:   name [198]
##   name      lat     n
##   <chr>    <dbl> <int>
## 1 AL011993  21.5     1
## 2 AL011993  22.3     1
## 3 AL011993  23.2     1
## 4 AL011993  24.5     1
## 5 AL011993  25.4     1
## 6 AL011993  26.1     1
## 7 AL011993  26.7     1
## 8 AL011993  27.8     1
## 9 AL012000  20.7     1
## 10 AL012000  20.8     1
## # ... with 8,160 more rows
```

For each storm, summarize its duration in number of hours (to the nearest 6hr increment).

```
storms %>%
  group_by(name, hour<=6) %>%
  tally()

## # A tibble: 396 x 3
## # Groups:   name [198]
##   name      `hour <= 6`     n
##   <chr>    <lgl>         <int>
## 1 AL011993 FALSE           4
## 2 AL011993 TRUE            4
## 3 AL012000 FALSE           2
## 4 AL012000 TRUE            2
## 5 AL021992 FALSE           3
## 6 AL021992 TRUE            2
## 7 AL021994 FALSE           3
## 8 AL021994 TRUE            3
## 9 AL021999 FALSE           1
## 10 AL021999 TRUE            3
## # ... with 386 more rows
```

For storm in a category, create a variable `storm_number` that enumerates the storms 1, 2, ... (in date order).

```

storms %>%
  group_by(category, name) %>%
  tally()

## # A tibble: 687 x 3
## # Groups:   category [7]
##   category name      n
##   <ord>    <chr>  <int>
## 1 -1      AL011993     8
## 2 -1      AL012000     4
## 3 -1      AL021992     5
## 4 -1      AL021994     6
## 5 -1      AL021999     4
## 6 -1      AL022000    12
## 7 -1      AL022001     5
## 8 -1      AL022003     4
## 9 -1      AL022006     1
## 10 -1     AL031987    28
## # ... with 677 more rows

```

Convert year, month, day, hour into the variable timestamp using the lubridate package. Although the new package clock just came out, lubridate still seems to be standard. Next year I'll probably switch the class to be using clock.

```

pacman::p_load("lubridate")
storms_date = storms %>%
  mutate(Date = paste(year, month, day, hour, sep = "-")) %>%
  select(name, Date)

storms_date

## # A tibble: 10,010 x 2
##   name   Date
##   <chr> <chr>
## 1 Amy   1975-6-27-0
## 2 Amy   1975-6-27-6
## 3 Amy   1975-6-27-12
## 4 Amy   1975-6-27-18
## 5 Amy   1975-6-28-0
## 6 Amy   1975-6-28-6
## 7 Amy   1975-6-28-12
## 8 Amy   1975-6-28-18
## 9 Amy   1975-6-29-0
## 10 Amy  1975-6-29-6
## # ... with 10,000 more rows

```

Using the lubridate package, create new variables day\_of\_week which is a factor with levels "Sunday", "Monday", ... "Saturday" and week\_of\_year which is integer 1, 2, ..., 52.

```

storms_date %>%
  mutate(day_of_week = wday(storms_date$Date, label = TRUE))

```

```
## # A tibble: 10,010 x 3
##   name   Date       day_of_week
##   <chr> <chr>       <ord>
## 1 Amy    1975-6-27-0   Fri
## 2 Amy    1975-6-27-6   Fri
## 3 Amy    1975-6-27-12  Fri
## 4 Amy    1975-6-27-18  Fri
## 5 Amy    1975-6-28-0    Sat
## 6 Amy    1975-6-28-6    Sat
## 7 Amy    1975-6-28-12  Sat
## 8 Amy    1975-6-28-18  Sat
## 9 Amy    1975-6-29-0    Sun
## 10 Amy   1975-6-29-6    Sun
## # ... with 10,000 more rows
```

storms\_date

```
## # A tibble: 10,010 x 2
##   name   Date
##   <chr> <chr>
## 1 Amy    1975-6-27-0
## 2 Amy    1975-6-27-6
## 3 Amy    1975-6-27-12
## 4 Amy    1975-6-27-18
## 5 Amy    1975-6-28-0
## 6 Amy    1975-6-28-6
## 7 Amy    1975-6-28-12
## 8 Amy    1975-6-28-18
## 9 Amy    1975-6-29-0
## 10 Amy   1975-6-29-6
## # ... with 10,000 more rows
```

For each storm, summarize the day in which is started in the following format “Friday, June 27, 1975”.

```
storms_date %>%
  mutate(day_of_week = wday(storms_date$Date, label = TRUE)) %>%
  mutate(Month = month(storms_date$Date, label = TRUE))
```

```
## # A tibble: 10,010 x 4
##   name   Date       day_of_week Month
##   <chr> <chr>       <ord>    <ord>
## 1 Amy    1975-6-27-0   Fri      Jun
## 2 Amy    1975-6-27-6   Fri      Jun
## 3 Amy    1975-6-27-12  Fri      Jun
## 4 Amy    1975-6-27-18  Fri      Jun
## 5 Amy    1975-6-28-0    Sat      Jun
## 6 Amy    1975-6-28-6    Sat      Jun
## 7 Amy    1975-6-28-12  Sat      Jun
## 8 Amy    1975-6-28-18  Sat      Jun
## 9 Amy    1975-6-29-0    Sun      Jun
```

```
## 10 Amy    1975-6-29-6  Sun          Jun
## # ... with 10,000 more rows
```

```
storms_date
```

```
## # A tibble: 10,010 x 2
##   name    Date
##   <chr> <chr>
## 1 Amy    1975-6-27-0
## 2 Amy    1975-6-27-6
## 3 Amy    1975-6-27-12
## 4 Amy    1975-6-27-18
## 5 Amy    1975-6-28-0
## 6 Amy    1975-6-28-6
## 7 Amy    1975-6-28-12
## 8 Amy    1975-6-28-18
## 9 Amy    1975-6-29-0
## 10 Amy   1975-6-29-6
## # ... with 10,000 more rows
```

Create a new factor variable `decile_windspeed` by binning wind speed into 10 bins.

*#TO-DO*

Create a new data frame `serious_storms` which are category 3 and above hurricanes.

```
storms %>%
  mutate(serious_storms = category >= 3) %>%
  select(name, serious_storms)
```

```
## # A tibble: 10,010 x 2
##   name    serious_storms
##   <chr> <lgl>
## 1 Amy    FALSE
## 2 Amy    FALSE
## 3 Amy    FALSE
## 4 Amy    FALSE
## 5 Amy    FALSE
## 6 Amy    FALSE
## 7 Amy    FALSE
## 8 Amy    FALSE
## 9 Amy    FALSE
## 10 Amy   FALSE
## # ... with 10,000 more rows
```

In `serious_storms`, merge the variables `lat` and `long` together into `lat_long` with values `lat / long` as a string.

```
storms %>%
  mutate(lat_long = paste(lat, long, sep = ", ")) %>%
  select(name, lat_long)
```

```
## # A tibble: 10,010 x 2
##   name lat_long
##   <chr> <chr>
## 1 Amy 27.5, -79
## 2 Amy 28.5, -79
## 3 Amy 29.5, -79
## 4 Amy 30.5, -79
## 5 Amy 31.5, -78.8
## 6 Amy 32.4, -78.7
## 7 Amy 33.3, -78
## 8 Amy 34, -77
## 9 Amy 34.4, -75.8
## 10 Amy 34, -74.8
## # ... with 10,000 more rows
```

Let's return now to the original storms data frame. For each category, find the average wind speed, pressure and diameters (do not count the NA's in your averaging).

```
Average_wind = mean(storms$wind)
Average_pressure = mean(storms$pressure)
Average_ts_diameter = mean(storms$ts_diameter)
Average_hu_diameter = mean(storms$hu_diameter)
```

For each named storm, find its maximum category, wind speed, pressure and diameters (do not allow the max to be NA) and the number of readings (i.e. observations).

*## this a repeat question from above in line 259*

Calculate the distance from each storm observation to Miami in a new variable `distance_to_miami`. This is very challenging. You will need a function that computes distances from two sets of latitude / longitude coordinates.

```
MIAMI_LAT_LONG_COORDS = c(25.7617, -80.1918)
distance_to_miami = c(NA)
distance_to_miami_left = c(NA)
distance_to_miami_right = c(NA)
distance = function(d) {
  for (i in 1:d) {
    distance_to_miami_left[i] = MIAMI_LAT_LONG_COORDS[1]-storms$lat[i]
    distance_to_miami_right[i] = MIAMI_LAT_LONG_COORDS[2] - storms$long[i]
  }
  distance_to_miami[i] = c(distance_to_miami_left[i],
distance_to_miami_right[i])
}
```

For each storm observation, use the function from the previous question to calculate the distance it moved since the previous observation.

*#TO-DO*

For each storm, find the total distance it moved over its observations and its total displacement. “Distance” is a scalar quantity that refers to “how much ground an object has covered” during its motion. “Displacement” is a vector quantity that refers to “how far out of place an object is”; it is the object’s overall change in position.

#TO-DO

For each storm observation, calculate the average speed the storm moved in location.

#TO-DO

For each storm, calculate its average ground speed (how fast its eye is moving which is different from windspeed around the eye).

#TO-DO

Is there a relationship between average ground speed and maximum category attained? Use a dataframe summary (not a regression).

#TO-DO

Now we want to transition to building real design matrices for prediction. This is more in tune with what happens in the real world. Large data dump and you convert it into  $X$  and  $y$  how you see fit.

Suppose we wish to predict the following: given the first three readings of a storm, can you predict its maximum wind speed? Identify the  $y$  and identify which features you need  $x_1, \dots, x_p$  and build that matrix with `dplyr` functions. This is not easy, but it is what it’s all about. Feel free to “featurize” as creatively as you would like. You aren’t going to overfit if you only build a few features relative to the total 198 storms.

#TO-DO

Fit your model. Validate it.

#TO-DO

Assess your level of success at this endeavor.

#TO-DO

## The Forward Stepwise Procedure for Probability Estimation Models

Set a seed and load the `adult` dataset and remove missingness and randomize the order.

```
set.seed(1)
pacman::p_load_gh("coatless/ucidata")
data(adult)
adult = na.omit(adult)
adult = adult[sample(1 : nrow(adult)), ]
```



Copy from the previous lab all cleanups you did to this dataset.

*#TO-DO*

We will be doing model selection. We will split the dataset into 3 distinct subsets. Set the size of our splits here. For simplicity, all three splits will be identically sized. We are making it small so the stepwise algorithm can compute quickly. If you have a faster machine, feel free to increase this.

```
Nsplitsize = 1000
```

Now create the following variables: Xtrain, ytrain, Xselect, yselect, Xtest, ytest with Nsplitsize observations. Binarize the y values.

```
Xtrain = adult[1 : Nsplitsize, ]
Xtrain$income = NULL
ytrain = ifelse(adult[1 : Nsplitsize, "income"] == ">50K", 1, 0)
Xselect = adult[(Nsplitsize + 1) : (2 * Nsplitsize), ]
Xselect$income = NULL
yselect = ifelse(adult[(Nsplitsize + 1) : (2 * Nsplitsize), "income"] == ">50K", 1, 0)
Xtest = adult[(2 * Nsplitsize + 1) : (3 * Nsplitsize), ]
Xtest$income = NULL
ytest = ifelse(adult[(2 * Nsplitsize + 1) : (3 * Nsplitsize), "income"] == ">50K", 1, 0)
```

Fit a vanilla logistic regression on the training set.

```
logistic_mod = glm(ytrain ~ ., Xtrain, family = "binomial")
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

and report the log scoring rule, the Brier scoring rule.

*#TO-DO*

We will be doing model selection using a basis of linear features consisting of all first-order interactions of the 14 raw features (this will include square terms as squares are interactions with oneself).

Create a model matrix from the training data containing all these features. Make sure it has an intercept column too (the one vector is usually an important feature). Cast it as a data frame so we can use it more easily for modeling later on. We're going to need those model matrices (as data frames) for both the select and test sets. So make them here too (copy-paste). Make sure their dimensions are sensible.

```
#TO-DO
"
dim(Xmm_train)
dim(Xmm_select)
dim(Xmm_test)
"
```

```
## [1] "\ndim(Xmm_train)\ndim(Xmm_select)\ndim(Xmm_test)\n"
```

Write code that will fit a model stepwise. You can refer to the chunk in the practice lecture. Use the negative Brier score to do the selection. The negative of the Brier score is always positive and lower means better making this metric kind of like `s_e` so the picture will be the same as the canonical U-shape for oos performance.

Run the code and hit “stop” when you begin to see the Brier score degrade appreciably oos. Be patient as it will wobble.

```
pacman::p_load(Matrix)
"
p_plus_one = ncol(Xmm_train)
predictor_by_iteration = c() #keep a growing list of predictors by iteration
in_sample_brier_by_iteration = c() #keep a growing list of briers by iteration
oos_brier_by_iteration = c() #keep a growing list of briers by iteration
i = 1

repeat {

  #TO-DO
  #wrap glm and predict calls with use suppressWarnings() so the console is
  clean during run

  if (i > Nsplitsize || i > p_plus_one){
    break
  }
}
"

## [1] "\np_plus_one = ncol(Xmm_train)\npredictor_by_iteration = c() #keep a
growing list of predictors by iteration\nin_sample_brier_by_iteration = c()
#keep a growing list of briers by iteration\nnoos_brier_by_iteration = c()
#keep a growing list of briers by iteration\nni = 1\n\nrepeat {\n\n  #TO-DO \n
#wrap glm and predict calls with use suppressWarnings() so the console is
clean during run\n  \n  if (i > Nsplitsize || i > p_plus_one){\n    break\n
}\n}\n"
```

Plot the in-sample and oos (select set) Brier score by  $p$ . Does this look like what's expected?

```
#TO-DO
```