dataframes

Michael Ghattas

Part 1 Examples

Read in data

General text files may be read in using *read.table* using appropriate arguments. CSV format, an "industry standard" typically uses *read.csv*

```
dat<-read.csv("worms.csv")
```

Inspecting the data

You can view dataframes by double clicking in RStudio. There are some standard summary tools.

```
dat
```

```
##           Field.Name Area Slope Vegetation Soil.pH  Damp Worm.density
## 1        Nashs.Field  3.6    11  Grassland     4.1 FALSE            4
## 2     Silwood.Bottom  5.1     2     Arable     5.2 FALSE            7
## 3      Nursery.Field  2.8     3  Grassland     4.3 FALSE            2
## 4        Rush.Meadow  2.4     5     Meadow     4.9  TRUE            5
## 5    Gunness.Thicket  3.8     0      Scrub     4.2 FALSE            6
## 6           Oak.Mead  3.1     2  Grassland     3.9 FALSE            2
## 7       Church.Field  3.5     3  Grassland     4.2 FALSE            3
## 8            Ashurst  2.1     0     Arable     4.8 FALSE            4
## 9        The.Orchard  1.9     0    Orchard     5.7 FALSE            9
## 10     Rookery.Slope  1.5     4  Grassland     5.0  TRUE            7
## 11       Garden.Wood  2.9    10      Scrub     5.2 FALSE            8
## 12      North.Gravel  3.3     1  Grassland     4.1 FALSE            1
## 13      South.Gravel  3.7     2  Grassland     4.0 FALSE            2
## 14 Observatory.Ridge  1.8     6  Grassland     3.8 FALSE            0
## 15        Pond.Field  4.1     0     Meadow     5.0  TRUE            6
## 16      Water.Meadow  3.9     0     Meadow     4.9  TRUE            8
## 17         Cheapside  2.2     8      Scrub     4.7  TRUE            4
## 18        Pound.Hill  4.4     2     Arable     4.5 FALSE            5
## 19        Gravel.Pit  2.9     1  Grassland     3.5 FALSE            1
## 20         Farm.Wood  0.8    10      Scrub     5.1  TRUE            3
```

```
dim(dat)
```

```
## [1] 20  7
```

```
names(dat)
```

```
## [1] "Field.Name"   "Area"         "Slope"        "Vegetation"   "Soil.pH"
## [6] "Damp"         "Worm.density"
```

```
str(dat)
```

```
## 'data.frame':    20 obs. of  7 variables:
##  $ Field.Name  : chr  "Nashs.Field" "Silwood.Bottom" "Nursery.Field" "Rush.Meadow" ...
##  $ Area        : num  3.6 5.1 2.8 2.4 3.8 3.1 3.5 2.1 1.9 1.5 ...
##  $ Slope       : int  11 2 3 5 0 2 3 0 0 4 ...
##  $ Vegetation  : chr  "Grassland" "Arable" "Grassland" "Meadow" ...
##  $ Soil.pH     : num  4.1 5.2 4.3 4.9 4.2 3.9 4.2 4.8 5.7 5 ...
##  $ Damp        : logi  FALSE FALSE FALSE TRUE FALSE FALSE ...
##  $ Worm.density: int  4 7 2 5 6 2 3 4 9 7 ...
```

```
summary(dat)
```

```
##   Field.Name             Area           Slope        Vegetation
##  Length:20          Min.   :0.800   Min.   : 0.00   Length:20
##  Class :character   1st Qu.:2.175   1st Qu.: 0.75   Class :character
##  Mode  :character   Median :3.000   Median : 2.00   Mode  :character
##                     Mean   :2.990   Mean   : 3.50
##                     3rd Qu.:3.725   3rd Qu.: 5.25
##                     Max.   :5.100   Max.   :11.00
##     Soil.pH         Damp          Worm.density
##  Min.   :3.500   Mode :logical   Min.   :0.00
##  1st Qu.:4.100   FALSE:14        1st Qu.:2.00
##  Median :4.600   TRUE :6         Median :4.00
##  Mean   :4.555                   Mean   :4.35
##  3rd Qu.:5.000                   3rd Qu.:6.25
##  Max.   :5.700                   Max.   :9.00
```

```
head(dat)
```

```
##        Field.Name Area Slope Vegetation Soil.pH  Damp Worm.density
## 1     Nashs.Field  3.6    11  Grassland     4.1 FALSE            4
## 2  Silwood.Bottom  5.1     2     Arable     5.2 FALSE            7
## 3   Nursery.Field  2.8     3  Grassland     4.3 FALSE            2
## 4     Rush.Meadow  2.4     5     Meadow     4.9  TRUE            5
## 5 Gunness.Thicket  3.8     0      Scrub     4.2 FALSE            6
## 6        Oak.Mead  3.1     2  Grassland     3.9 FALSE            2
```

```
dat[2,4]
```

```
## [1] "Arable"
```

Select columns

Leaving the row index blank and providing column indices allows selection of a subset of the columns.

```
dat[,3]
```

```
##  [1] 11  2  3  5  0  2  3  0  0  4 10  1  2  6  0  0  8  2  1 10
```

```
dat$Vegetation
```

```
##  [1] "Grassland" "Arable"    "Grassland" "Meadow"    "Scrub"     "Grassland"
##  [7] "Grassland" "Arable"    "Orchard"   "Grassland" "Scrub"     "Grassland"
## [13] "Grassland" "Grassland" "Meadow"    "Meadow"    "Scrub"     "Arable"
## [19] "Grassland" "Scrub"
```

```
dat[,2:4]
```

```
##    Area Slope Vegetation
## 1   3.6    11  Grassland
## 2   5.1     2     Arable
## 3   2.8     3  Grassland
## 4   2.4     5     Meadow
## 5   3.8     0      Scrub
## 6   3.1     2  Grassland
## 7   3.5     3  Grassland
## 8   2.1     0     Arable
## 9   1.9     0    Orchard
## 10  1.5     4  Grassland
## 11  2.9    10      Scrub
## 12  3.3     1  Grassland
## 13  3.7     2  Grassland
## 14  1.8     6  Grassland
## 15  4.1     0     Meadow
## 16  3.9     0     Meadow
## 17  2.2     8      Scrub
## 18  4.4     2     Arable
## 19  2.9     1  Grassland
## 20  0.8    10      Scrub
```

Select rows

Leaving the column index blank and providing a vector of indices or boolean values allows selection of a subset of the rows.

```
dat[1:4,]
```

```
##       Field.Name Area Slope Vegetation Soil.pH  Damp Worm.density
## 1    Nashs.Field  3.6    11  Grassland     4.1 FALSE            4
## 2 Silwood.Bottom  5.1     2     Arable     5.2 FALSE            7
## 3  Nursery.Field  2.8     3  Grassland     4.3 FALSE            2
## 4    Rush.Meadow  2.4     5     Meadow     4.9  TRUE            5
```

```
damp<-dat[dat$Damp,]
summary(damp)
```

```
##   Field.Name             Area           Slope        Vegetation
##  Length:6           Min.   :0.800   Min.   : 0.00   Length:6
##  Class :character   1st Qu.:1.675   1st Qu.: 1.00   Class :character
##  Mode  :character   Median :2.300   Median : 4.50   Mode  :character
##                     Mean   :2.483   Mean   : 4.50
##                     3rd Qu.:3.525   3rd Qu.: 7.25
##                     Max.   :4.100   Max.   :10.00
##     Soil.pH        Damp          Worm.density
##  Min.   :4.700   Mode:logical   Min.   :3.00
##  1st Qu.:4.900   TRUE:6         1st Qu.:4.25
##  Median :4.950                  Median :5.50
##  Mean   :4.933                  Mean   :5.50
##  3rd Qu.:5.000                  3rd Qu.:6.75
##  Max.   :5.100                  Max.   :8.00
```

```
flat<-dat[dat$Slope<4.5,]
summary(flat)
```

```
##   Field.Name             Area           Slope        Vegetation
##  Length:14          Min.   :1.500   Min.   :0.000   Length:14
##  Class :character   1st Qu.:2.825   1st Qu.:0.000   Class :character
##  Mode  :character   Median :3.400   Median :1.500   Mode  :character
##                     Mean   :3.293   Mean   :1.429
##                     3rd Qu.:3.875   3rd Qu.:2.000
##                     Max.   :5.100   Max.   :4.000
##     Soil.pH         Damp          Worm.density
##  Min.   :3.500   Mode :logical   Min.   :1.00
##  1st Qu.:4.125   FALSE:11        1st Qu.:2.00
##  Median :4.400   TRUE :3         Median :4.50
##  Mean   :4.521                   Mean   :4.50
##  3rd Qu.:4.975                   3rd Qu.:6.75
##  Max.   :5.700                   Max.   :9.00
```

Practice Problem for Part 1

Please supply code to create a new data frame “grass” the restricts “dat” to observations in which the “Vegetation” variable equals “Grassland”. Output the values of the variable “Damp” for this data frame.

```
# Create the 'grass' data frame
grass <- subset(dat, Vegetation == "Grassland")

# Output the values of the 'Damp' variable for the 'grass' data frame
damp_values <- grass$Damp

# Display the 'grass' data frame and 'Damp' values
grass
```

```
##           Field.Name Area Slope Vegetation Soil.pH  Damp Worm.density
## 1        Nashs.Field  3.6    11  Grassland     4.1 FALSE            4
## 3      Nursery.Field  2.8     3  Grassland     4.3 FALSE            2
## 6           Oak.Mead  3.1     2  Grassland     3.9 FALSE            2
## 7       Church.Field  3.5     3  Grassland     4.2 FALSE            3
## 10     Rookery.Slope  1.5     4  Grassland     5.0  TRUE            7
## 12      North.Gravel  3.3     1  Grassland     4.1 FALSE            1
## 13      South.Gravel  3.7     2  Grassland     4.0 FALSE            2
## 14 Observatory.Ridge  1.8     6  Grassland     3.8 FALSE            0
## 19        Gravel.Pit  2.9     1  Grassland     3.5 FALSE            1
```

```
damp_values
```

```
## [1] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
```

Part 2 Example

Find Maximum Likelihood, binomial

Maximize $p^{30}\left(1-p\right)^{20}$. The "optimize" function works for single variable functions.

```
# function to calculate and return x^30*(1-x)^20, the function to be optimized.
f<-function(x){
  return (x^30*(1-x)^20)
}
(ml_ext<-optimize(f,c(0,1),maximum=TRUE)) # Look for the value that maximizes the function f
```

```
## $maximum
## [1] 0.6000077
##
## $objective
## [1] 2.430733e-15
```

```
                                          # in the range [0,1], the range of valid probabilities.
30/50 # theoretical value =30/(30+20)
```

```
## [1] 0.6
```

```
f(.6) # check
```

```
## [1] 2.430733e-15
```

Practice Problem for Part 2

Please revise the code above to calculate the value of p that maximizes $p^{15}\left(1-p\right)^{85}$. Check the result by computing the relevant ratio. Calculate the maximum value of the function.

```r
# Function to calculate and return p^15*(1-p)^85, the function to be optimized.
f <- function(p) {
  return (p^15 * (1-p)^85)
}

# Look for the value that maximizes the function f in the range [0,1].
ml_ext <- optimize(f, c(0, 1), maximum = TRUE)

# Theoretical value
theoretical_value = (15 / (15 + 85))

# Check the value at the theoretical maximum
f_theoretical <- f(theoretical_value)

# Calculate the maximum value of the function
max_value <- ml_ext$objective

# Output the results
ml_ext
```

```
## $maximum
## [1] 0.1499918
##
## $objective
## [1] 4.38508e-19
```

```r
theoretical_value
```

```
## [1] 0.15
```

```r
f_theoretical
```

```
## [1] 4.385081e-19
```

```r
max_value
```

```
## [1] 4.38508e-19
```