Statistics 133 Final Exam

May 11, 2010

When I ask for an "R program", I mean one or more R commands. Try your best to make your answers general, i.e. they shouldn't depend on the specific values presented in the examples.

Total: 60 points

1. Consider an SQL database table called `kids`, containing the following variables:

   - `school` representing several values of schools.
   - `gender` represented as `MALE` or `FEMALE`.
   - `height` measured in meters.
   - `weight` measured in kilograms.

   The body mass index (BMI) can be calculated as

   $$\text{BMI} = \frac{\text{weight}}{\text{height}^2}$$

   (a) (2 points) Write an SQL statement to show all the original columns for each observation, along with a new column called `BMI` that contains the body mass index.

   (b) (2 points) Write an SQL statement to display the mean height, weight, and BMI broken down by school.

   (c) (2 points) Write an SQL statement that will display the total number of observations in the table.

2. Regular Expressions

   (a) (2 points) Write an R program to extract lines from a vector names `thestring` which have more than one double quoted strings. For example, these lines

   ```
   this "line" has "quoted" "strings"
   "one" more to follow "two"
   "a""b"
   ```

   should be extracted, but these:

   ```
   there's "one" quoted string
   Here's a quote "
   No quotes here
   ```

   should not.

   (b) (2 points) Write an R program that will remove multiple blanks from before and after a vector of strings called `fullstrings`. For example, "` hello, world `" should be converted to "`hello, world`", and "`Stat 133 `" should be converted to "`Stat 133`".

(c) (2 points) Social security numbers in the United States are represented by a leading 0 followed by two digits, followed by a dash, followed by two digits, followed by a dash, finally followed by four digits. For example `023-45-7890` would be a valid value, but `05-09-1995` and `059-2-27` would not be. Write an R program that would extract Social Security numbers from a vector of strings called `text`. (You can assume there are exactly 0 or 1 Social Security numbers in each string.)

Hint: getexpr = function(s,g)substring(s,g,g + attr(g,'match.length') - 1)

3. (2 points) Consider a vector `mm`, displayed by R as follows:

```
> mm
 [1] 13 14 12 12 14 14 11 11 11 12 10 13 12 11 14 10 12 15 11 14
Levels: 10 11 12 13 14 15
```

Write an R program to calculate the sum of the values in `mm`.

4. Smoothers

   (a) (2 points) Name two or more smoothers that are available in R.

   (b) (2 points) Which smoother does *not* require an input parameter describing the bandwidth or fraction of the data that should be used for smoothing.

5. (2 points) The `rpart` package, and the `rpart` function provide recursive partitioning solutions for both classification and regression. What determines whether the `rpart` function will perform a classification analysis or a regression analysis?

6. Clustering

   (a) (2 points) Name a clustering method that requires you to specify the number of clusters in its solution. Name a clustering method that does *not* require you to specify the number of clusters.

   (b) (2 points) Write a R program to standardize the columns of `X` by subtracting the median and dividing by the mean average deviation.

   (c) (2 points) Name 2 distance measures that can be used in `dist()`

7. Consider a data frame called `mydat`, whose summary output is shown below:

```
     group             y            
 Min.   :1.000   Min.   :-2.923601  
 1st Qu.:2.000   1st Qu.:-0.704193  
 Median :3.000   Median : 0.010546  
 Mean   :2.978   Mean   : 0.007019  
 3rd Qu.:4.000   3rd Qu.: 0.735157  
 Max.   :5.000   Max.   : 3.160055  
```

   (a) (2 points) Write an R program that will perform an analysis of variance (ANOVA) to test the hypothesis that the mean of `y` is the same for the five groups defined by `group`.

   (b) (2 points) Write an R program to produce a lattice plot with five panels, each containing a histogram of `y` broken down by the value of `group`.

   (c) (2 points) Write an R program to show how many observations there are for each `group`.

   (d) (2 points) How many missing values are there in `mydat$y`?

8. (2 points) When extracting information from XML files, the double square bracket subscripting operator was used instead of the usual single bracket. For example, we would write

```
doc[['value']]
```

   instead of

```
doc['value']
```

   Why do we need to use double brackets in these cases?

9. (2 points) What is the principal difference between a regression model fit by the `lm` function, and a regression model fit by the `gam` function from the `mgcv` package?

10. Consider a data frame called `fitness`, with seven variables. Here's the output of the summary command:

```
      Age            Weight          Oxygen         RunTime     
 Min.   :38.00   Min.   :59.08   Min.   :37.39   Min.   : 8.17  
 1st Qu.:44.00   1st Qu.:73.20   1st Qu.:44.96   1st Qu.: 9.78  
 Median :48.00   Median :77.45   Median :46.77   Median :10.47  
 Mean   :47.68   Mean   :77.44   Mean   :47.38   Mean   :10.59  
 3rd Qu.:51.00   3rd Qu.:82.33   3rd Qu.:50.13   3rd Qu.:11.27  
 Max.   :57.00   Max.   :91.63   Max.   :60.05   Max.   :14.03  
   RestPulse        RunPulse        MaxPulse    
 Min.   :40.00   Min.   :146.0   Min.   :155.0  
 1st Qu.:48.00   1st Qu.:163.0   1st Qu.:168.0  
 Median :52.00   Median :170.0   Median :172.0  
 Mean   :53.45   Mean   :169.6   Mean   :173.8  
 3rd Qu.:58.50   3rd Qu.:176.0   3rd Qu.:180.0  
 Max.   :70.00   Max.   :186.0   Max.   :192.0  
```

   (a) (2 points) Write an R program that will perform a linear regression with `Oxygen` as the dependent variable, and `Age`, `Weight`, and `RunTime` as independent variables, and which will display t-tests and probabilities for the test that the slope of each of the independent variables is 0.

   (b) (2 points) What facilities are available in R to check if the assumptions of the linear regression are met by this model for the `fitness` data.

11. (a) (2 points) Write an R program that will determine the number of days between today (May 11, 2010) and next Christmas (December 25, 2010).

   (b) (2 points) Write an R program that will convert character strings stored in a vector `str` into proper R `Date` values. The contents of `str` are shown below:

```
> str
[1] "4-12-2010" "5-15-2010" "7-4-2010"
```

12. (2 points) In the R formula language, what is the difference between a term like `A:B`, and a term like `A*B`?

13. (2 points) We compute the error rate of LDA on the wine dataset as follows:

```
> wine.lda = lda(Cultivar ~ .,data=wine)
> pred = predict(wine.lda,wine)
> tt=table(wine$Cultivar,pred$class)
> error = sum(tt[row(tt) != col(tt)]) / sum(tt)
```

Is this an accurate measure of how good the classifier is? Why or why not?

14. (4 points) Hypothesis Testing. Answer True or False.

    (a) A test at 5% significance level is expected to falsely reject the null hypothesis 5% of the time.

    (b) If the test statistic is in the rejection region, then the null hypothesis is not reasonable (according to the data).

    (c) The power of the hypothesis test depends on the true parameter value.

    (d) Type I error = 1 - Type II error.

15. (2 points) Write a R callback function `buttonclick` (part of a GUI) that increments the global variable `nclicks` by 1.

16. (2 points) With the `CGIwithR` library, the CGI variables (defined in the forms) are stored in a R object. What is its name and structure?

17. (2 points) Write a R function `srctail` that takes in an url and returns the last 10 lines of its page source.