Stat 416 Homework 4

Due Date: Thursday, April 2

Gene Expression Omnibus (GEO) is a gene expression repository that allows data browsing, query and retrieval. The GEO website is

`http://www.ncbi.nlm.nih.gov/geo/`

From this page, it is possible to access data from thousands of microarray experiments. One example data set is available at

`http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6297`

Please examine the material on this Web page and download the data by clicking on the *Series Matrix File(s)* link near the bottom of the page. To answer some of these questions, you may wish to examine the corresponding publication at

`http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=18231591`

1. Name the factor or factors in this experiment. Name the levels of each factor.
2. Name the experimental units in this experiment.
3. Describe the experimental design of this experiment.
4. Can you identify an factors in the experiment that are not included in the data set?
5. Read the data into an R workspace. This can be done by setting the working directory to the path where the data file is stored on your computer and then issuing the following command.

```
d=read.table("GSE6297_series_matrix.txt",skip=67,nrows=45101,header=T)
```

6. Make side-by-side boxplots of all the slides and comment on any notable features.
7. What method was used to compute these expression measures?
8. The R function `lm` can be used to fit linear models. An example is provided below for fitting a two-factor additive model of the form

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \quad (i = 1, 2; j = 1, 2; k = 1, 2, 3)$$

where the ε_{ijk} values are independent and normally distributed with mean 0 and some unknown variance σ^2 and $\mu, \alpha_1, \alpha_2, \beta_1$, and β_2 are unknown mean parameters.

```
y=c(5.81,6.18,6.36,5.03,4.54,5.19,5.88,6.01,5.86,5.32,4.55,5.27)
a=c(1,1,1,1,1,1,2,2,2,2,2,2)
b=c(1,1,1,2,2,2,1,1,1,2,2,2)

out=lm(y˜as.factor(a)+as.factor(b))

anova(out)

#Note that you can extract specific elements of the ANOVA
#table as follows.

anova(out)[2,5]
```

(a) Run this R code and examine the output. (If you cut and paste, be sure to replace the tilde.) Provide a p-value for testing the significance of factor a. Do the same for factor b.

(b) Note that for this two-factor model, there are essentially four different means given in the table below.

	$b = 1$	$b = 2$
$a = 1$	$\mu + \alpha_1 + \beta_1$	$\mu + \alpha_1 + \beta_2$
$a = 2$	$\mu + \alpha_2 + \beta_1$	$\mu + \alpha_2 + \beta_2$

For any set of means satisfying the additive model, there are an infinite number of choices for $\mu, \alpha_1, \alpha_2, \beta_1$, and β_2 that will yield those means. For example, the table

	$b = 1$	$b = 2$
$a = 1$	5.2	7.5
$a = 2$	1.1	3.4

can be obtained by choosing $\mu = 0$, $\alpha_1 = 3.1$, $\alpha_2 = -1.0$ $\beta_1 = 2.1$, and $\beta_2 = 4.4$; OR by choosing $\mu = 5.2$, $\alpha_1 = 0$, $\alpha_2 = -4.1$ $\beta_1 = 0$, and $\beta_2 = 2.3$. Give another set of values for the mean parameters that will yield the same table.

(c) Because there are an infinite number of values for $\mu, \alpha_1, \alpha_2, \beta_1$, and β_2 that give exactly the same four estimated means, there are no unique best estimates of these parameters. However, there are unique best estimates of the four means $\mu + \alpha_1 + \beta_1$, $\mu + \alpha_1 + \beta_2$, $\mu + \alpha_2 + \beta_1$, and $\mu + \alpha_2 + \beta_2$. R will produce values that – when added together appropriately – give these best estimates of the four means. By default, R sets the value corresponding to the first level of each factor to be 0 and reports only values for μ and levels other than the first. For example, the values R reports will be like the second set of values provided at near the end of part (b). To see R's values for the mean parameters use the command `coef(out)`. Using these values, complete a table like those above that gives the four estimated means for this data set.

(d) Estimate the difference between the mean for $(a, b) = (1, 1)$ and the mean for $(a, b) = (2, 2)$.

(e) Test whether the mean difference estimated in part (d) is significantly different from 0. To do so, you will need more information. First, let $\hat{\underline{\theta}}$ denote the vector given by `coef(out)`. Let V denote the estimated variance matrix of this vector, which is given in R by `vcov(out)`. Then, for any vector $\underline{m}$, $\underline{m}'\hat{\underline{\theta}}/\sqrt{\underline{m}'V\underline{m}}$ has a t distribution under the null hypothesis $H_0 : \underline{m}'\underline{\theta} = 0$. The degrees of freedom associated with this t distribution are the error degrees of freedom from the ANOVA table. Note that the R code `2*(1-pt(abs(x),df))` will provide a two-sided p-value if x is the test statistic and df is the correct degrees of freedom.

9. Write an R function, that will return a p-values for testing effects of interest when given a vector containing the data for one of the genes in the data that you downloaded. Explain what tests your function carries out, and report the p-values for the first gene in the data set.

10. Use the apply function to analyze the data for all genes in the data set. Make histograms of the p-values for each test of interest and comment on their shapes.