Stat 516 Homework 2 Due Tuesday, February 8

1. Refer to the expression for $E(X_i|D_i)$ on slide 15 of the course notes on preprocessing of two-color microarray data. Is this expression correct? Provide a clear derivation to support your answer.

2. Suppose the following are the pixel intensities associated with an Affymetrix probe cell. Determine the single value that would be assigned to this probe cell assuming the layout of the values matches the layout of the pixels in the probe cell.

```
30610 31343 26855 34881 26215 29016 35762 37741
31046 32474 27517 29427 24026 33187 30683 36603
31988 36411 29184 34416 32092 24087 32175 33624
29232 19270 33989 35218 30082 35236 28004 38053
23515 32037 17145 26992 29316 28634 25706 26930
26896 24833 29453 37174 28360 30951 31682 35808
25509 32228 26776 30999 27335 23802 26478 40689
29476 34116 29046 35431 31898 29921 28882 33119
```

3. Write an R function that will accept a vector and return the one-step Tukey bi-weight of the numbers in the vector as implemented by Affymetrix.

4. Suppose the values below are PM and MM values that have been background adjusted using the MAS 5.0 background adjustment method.

	Probe Pair						
	1	2	3	4	5	6	7
PM	2234	3945	4232	2645	1787	1100	2345
MM	5895	2238	1887	300	1800	154	156

Complete the following parts using the exact Affymetrix algorithms as presented in class.

   (a) Determine the ideal mismatch for each probe pair.

   (b) Determine the *probe value* for each probe pair.

   (c) Find the *signal log value* for the probe set.

5. Suppose the matrix below contains background-adjusted (RMA method), base-2-logged, and quantile-normalized PM values from 10 GeneChips for an Affymetrix probe set with 11 probe pairs. There is one row for each GeneChip and one column for each probe.

```
7.58  8.23 8.87  9.80 10.85 10.44  9.23 10.52  9.06  9.92 10.67
9.67 11.52 9.85 11.64 12.46 12.24 11.39 12.20 11.33 12.39 12.70
8.67 10.63 9.20 10.67 11.80 11.37 10.28 11.37 10.12 11.35 11.52
8.55 10.61 9.16 10.80 11.68 11.28 10.43 11.48 10.37 11.64 11.85
8.59 10.72 9.19 11.11 11.93 11.39 10.75 11.73 10.49 11.76 12.05
7.76  9.74 8.64  9.69 10.72 10.17  9.18 10.33  8.90 10.20 10.48
8.25 10.37 8.94 10.60 11.44 11.12 10.27 11.25  9.92 11.37 11.72
8.02 10.17 8.87 10.06 11.33 10.82  9.67 10.93  9.40 10.70 11.04
8.79 10.98 9.12 10.76 11.88 11.37 10.39 11.30 10.20 11.40 11.60
9.26 11.09 9.52 11.26 11.86 11.59 10.97 11.70 10.68 11.92 12.07
```

Determine the GeneChip-specific RMA expression measures for this probe set.

6. Suppose an estimate of location is desired for a data set $x_1, x_2, \ldots, x_n$. An M-estimate for a function $f : \boldsymbol{R}^2 \to \boldsymbol{R}$ is defined as any value of $m \in \boldsymbol{R}$ that minimizes $\sum_{i=1}^{n} f(x_i, m)$.

   (a) The sample mean is one example of an M-estimate. Give the function f that corresponds to the sample mean.

   (b) The sample median is another example of an M-estimate. Give the function f that corresponds to the sample median.

   Tukey's biweight estimate is defined as the M-estimate corresponding to the function

$$f(x,m) = \begin{cases} \frac{1}{6}\left[1 - \left\{1 - \left(\frac{x-m}{k\text{MAD}}\right)^2\right\}^3\right] & \text{if } \left|\frac{x-m}{k\text{MAD}}\right| \le 1 \\ \frac{1}{6} & \text{otherwise;} \end{cases}$$

   where k is a positive constant and MAD is the median of the absolute deviations from the median for $x_1, x_2, \ldots, x_n$.

   (c) Write an R function with three arguments: m, x, and k. The m argument should be the first argument. The function should return the value of $\sum_{i=1}^{n} f(x_i, m)$ where f is as defined for Tukey's biweight estimate.

   (d) Write a function that will return a Tukey biweight estimate when given a vector of data points x and a constant k. You will need to numerically minimize $\sum_{i=1}^{n} f(x_i, m)$ as a function of m. One R function for numerically minimizing a function of one variable is *optimize*. You may want to include a call to *optimize* inside your function. You should also make use of the function that you wrote in part (c). Note that *optimize* will require you to specify lower and upper endpoints of an interval to be searched for the minimum. The key to getting *optimize* to work well is to carefully select the interval over which *optimize* will search for a solution. If that interval is too wide, *optimize* will not find the proper minimizer. You should come up with some method for specifying an interval as a function of the input data. It doesn't have to be completely fool proof for all data sets, but an overlying simplistic approach may not work for many data sets.

   (e) The one-step Tukey biweight estimate that we learned about in class is an approximation to the value that should be computed by the function you wrote for part (d) with $k = 5$. Write an R function that will compute the one-step Tukey biweight estimate as defined by Affymetrix (see question 3) except that for this problem you should ignore the factor of 0.0001 used by Affymetrix to avoid division by 0. Compare the one-step function with the function that you wrote in part (d) for the two data sets below.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Data Set 1	1	4	2	9	12	14	14	55
Data Set 2	-11.00	-1.00	0.50	0.25	1.00	2.00	2.40	45.00

   (f) The one-step estimator in part (e) computes the first step of an iterative procedure. An initial approximation to the Tukey biweight estimate is computed (the median), and then a weighted average is computed using this initial approximation to determine weights for each x_i. These weights are used to get a new approximation for the Tukey biweight estimator. Rather than stopping after this one step, it should be straightforward to adjust the function that you wrote in part (e) so that it will iterate for a user-defined number of iterations. In the second iteration, for example, the one-step estimate will play the role of the median in the numerator of each t_i in order to get new weights and a

new approximation of the Tukey biweight estimator. (Note that the median absolute deviation from the median (MAD) is not to be adjusted from iteration to iteration.) At each step of the algorithm, the previous approximation to the Tukey biweight estimator should be used to obtain a weight for each data point. The corresponding weighted average should serve as the new approximation for the Tukey biweight estimator. Adjust your function from part (e) accordingly, and try it on the two test data sets used in part (e). For each data set, how many iterations are needed before the two approaches agree through the first 5 decimal places?