Math 583 HW 2 Fall 2017. Due Wednesday, Sept. 6.

Quiz 2 on Friday, Sept. 8 is similar to HW 2. Use 2 sheets of notes. Place your solutions on a separate sheet of paper. DO NOT place solutions side by side. YOU ARE BEING GRADED FOR WORK NOT JUST THE FINAL ANSWER. As a rule of thumb, you should have some idea of what you were doing, even without the book or notes. You are encouraged to form groups to discuss ideas and HW problems, but do not copy.

Problem numbers are from the Olive text but usually are not yet in the handout text.

A) 1.2. The table $\boldsymbol{W}$ shown below represents 4 measurements on 5 people.

age	breadth	cephalic	size
39.00	149.5	81.9	3738
35.00	152.5	75.9	4261
35.00	145.5	75.4	3777
19.00	146.0	78.1	3904
0.06	88.5	77.6	933

a) Find the sample mean $\overline{\boldsymbol{x}}$.

b) Find the coordinatewise median MED($\boldsymbol{W}$).

Hint: See example done in class.)

Copy and paste the two source commands and the three library commands from near the top of *slrhw.txt* for the following *R* problems.

B) 1.13. For the Buxton (1920) data with multiple linear regression, *height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, cases 61–65, were reported to be about 0.75 inches tall with head lengths well over five feet!

a) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. The identity line passes right through the outliers which are obvious because of the large gap. Prediction interval (PI) bands are also included in the plot.

b) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. This did lasso for the cases in the `covmb2` set B applied to the predictors which included all of the clean cases and omitted the 5 outliers. The response plot was made for all of the data, including the outliers.

c) Copy and paste the commands for this problem into *R*. Include the DD plot in *Word*. The outliers are in the upper right corner of the plot.

C) 1.14. Consider the Gladstone (1905) data set that has 12 variables on 267 persons after death. There are 5 infants in the data set. The response variable was *brain weight*. Head measurements were *breadth, circumference, head height, length,* and *size* as well as *cephalic index* and *brain weight*. *Age*, *height*, and three categorical variables *cause*, *ageclass* (0: under 20, 1: 20-45, 2: over 45) and *sex* were also given. The constant x_1 was the first variable. The variables *cause* and *ageclass* were not coded as factors. Coding as factors might improve the fit.

a) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. The identity line passes right through the infants which are obvious because of the large gap. Prediction interval (PI) bands are also included in the plot.

b) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. This did lasso for the cases in the `covmb2` set B applied to the nontrivial predictors which are not categorical (omit the *constant, cause, ageclass* and *sex*) which omitted 8 cases, including the 5 infants. The response plot was made for all of the data.

c) Copy and paste the commands for this problem into *R*. Include the DD plot in *Word*. The infants are in the upper right corner of the plot.

D) 1.15. The *slpack* function `mldsim6` compares 7 estimators: FCH, RFCH, CMVE, RCMVE, RMVN, `covmb2`, and MB described in Olive (2017c, ch. 4). Most of these estimators need $n > 2p$, need a nonsingular dispersion matrix, and work best with $n > 10p$. The function generates data sets and counts how many times the minimum Mahalanobis distance $D_i(T, \boldsymbol{C})$ of the outliers is larger than the maximum distance of the clean data. The value *pm* controls how far the outliers need to be from the bulk of the data, and *pm* roughly needs to increase with $\sqrt{p}$.

For data sets with $p > n$ possible, the function `mldsim7` used the Euclidean distances $D_i(T, \boldsymbol{I}_p)$ and the Mahalanobis distances $D_i(T, \boldsymbol{C}_d)$ where $\boldsymbol{C}_d$ is the diagonal matrix with the same diagonal entries as $\boldsymbol{C}$ where $(T, \boldsymbol{C})$ is the `covmb2` estimator using j concentration type steps. Dispersion matrices are effected more by outliers than good robust location estimators, so when the outlier proportion is high, it is expected that the Euclidean distances $D_i(T, \boldsymbol{I}_p)$ will outperform the Mahalanobis distance $D_i(T, \boldsymbol{C}_d)$. Again the function counts the number of times the minimum outlier distance is larger than the maximum distance of the clean data.

Both functions used several outlier types. The simulations generated 100 data sets. The clean data had $\boldsymbol{x}_i \sim N_p(\boldsymbol{0}, diag(1, ..., p))$. Type 1 had outliers in a tight cluster (near point mass) at the major axis $(0, ..., 0, pm)^T$. Type 2 had outliers in a tight cluster at the minor axis $(pm, 0, ..., 0)^T$. Type 3 had mean shift outliers $\boldsymbol{x}_i \sim N_p((pm, ..., pm)^T, diag(1, ..., p))$. Type 4 changed the pth coordinate of the outliers to *pm*. Type 5 changed the 1st coordinate of the outliers to *pm*. (If the outlier $\boldsymbol{x}_i = (x_{1i}, ..., x_{pi})^T$, then $x_{i1} = pm$.)

Table 1: Number of Times All Outlier Distances > Clean Distances, otype=1

n	p	γ	osteps	pm	FCH	RFCH	CMVE	RCMVE	RMVN	covmb2	MB
100	10	0.25	0	20	85	85	85	85	86	67	89

a) Table 1 suggests with osteps = 0, `covmb2` had the worst count. When *pm* is increased to 25, all counts become 100. Copy and paste the commands for this part into *R* and make a table similar to Table 1, but now osteps=9 and $p = 45$ is close to $n/2$ for the second line where $pm = 60$. Your table should have 2 lines from output.

b) Copy and paste the commands for this part into *R* and make a table similar to Table 2, but type 2 outliers are used.

c) Copy and paste the commands for this part into *R* and make a table similar to Table 2, but type 3 outliers are used.

Table 2: Number of Times All Outlier Distances > Clean Distances, otype=1

n	p	γ	osteps	pm	covmb2	diag
100	1000	0.4	0	1000	100	41
100	1000	0.4	9	600	100	42