Math 583 HW 3 2023 Due Wednesday, Sept. 13.

A) The Gladstone (1905) brain data has response brain weight with predictors x_1 =age, x_2 = ageclass, x_3 = breadth of head, x_4 = cause of death, x_5 =cephalic index, x_6 = circum of head, x_7 = headht, x_8 = height of the person, x_9 = length of head, x_{10} = gender, and x_{11} = size of head. There are five infants and toddlers in the data set that appear to be good leverage points rather than outliers. The purpose of this problem is to show that there are often many MLR models for the response variable Y.

a) Copy and paste the commands for this part into *R*. Include the output after ls.print(outf) into *Word*. This output is for the full model.

b) Copy and paste the commands for this part into *R*. Include the output after ls.print(outf) into *Word*. This output is for an OLS submodel after backward elimination, and predictors x_1 =age, x_4 = cause of death, x_5 =cephalic index, x_7 = headht, x_9 = length of head, and x_{10} = gender. Include the output after ls.print(out2) into *Word*. Variable selection methods such as backward elimination attempt to eliminate unimportant predictors, given that the full model is good.

c) Copy and paste the commands for this part into *R*. Include the output after ls.print(outf) into *Word*. This output uses the predictors that were eliminated by backward elimination: x_2 = ageclass, x_3 = breadth of head, x_6 = circum of head, x_8 = height of the person, and x_{11} = size of head.

i) Include the response and residual plots for this model in *Word*.

ii) Is the model linear (do the plotted cases fall about the identity line)?

iii) Include the output after ls.print(out3) into *Word*.

B) 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.

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

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. (Now $\gamma = 0.4$ = default value.)

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. (Now $\gamma = 0.25$.)

C) 1.2. The table $\boldsymbol{W}$ shown below represents 4 measurements on 5 people. (See the example done in class.)

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($\boldsymbol{W}$).