```
Math 585    HW 11 Spring 2024       Due Friday, April 26
Quiz 11 on Wed. April 24, covers HW11.
Exam 3 is Wed. May 1.   2 pages, problems A)-E)
Final is  May .
```

For the following R problems perform the perform the *source("J:/mpack.txt")* and *source("J:/mrobdata.txt")* commands as described in homework 3. Also copy and paste commands from (http://parker.ad.siu.edu/Olive/mrsashw.txt) for the relevant problem into R.

A), 12.3 Using the Searle (1982, p. 333) identity
$tr(\boldsymbol{A}\boldsymbol{G}^T\boldsymbol{D}\boldsymbol{G}\boldsymbol{C}) = [vec(\boldsymbol{G})]^T[\boldsymbol{C}\boldsymbol{A} \otimes \boldsymbol{D}^T][vec(\boldsymbol{G})]$, show
$(n-p)U(\boldsymbol{L}) = tr[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\hat{\boldsymbol{B}}^T\boldsymbol{L}^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}]$
$= [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]$ by identifying $\boldsymbol{A}, \boldsymbol{G}, \boldsymbol{D}$, and $\boldsymbol{C}$.

B), 12.8 This problem examines multivariate linear regression on the Cook and Weisberg (1999a) mussels data with $Y_1 = \log(S)$ and $Y_2 = \log(M)$ where S is the shell mass and M is the muscle mass. The predictors are $X_2 = L$, $X_3 = \log(W)$ and $X_4 = H$: the shell length, log(width) and height.

a) The R command for this part make the response and residual plots for each of the three variables. Click the rightmost mouse button and highlight *Stop* to advance the plot. When you have the response and residual plots for one variable on the screen, copy and paste the two plots into *Word*. Do this two times, once for each response variable. The plotted points fall in roughly evenly populated bands about the identity or $r = 0$ line.

b) Copy and paste the output produced from the R command for this part from $partial on. This gives the output needed to do the MANOVA F test, MANOVA partial F test and the F_j tests.

c) The R command for this plot makes a DD plot of the residuals and adds the lines corresponding to the three prediction regions of Section 5.2. The robust cutoff is larger than the semiparametric cutoff. Place the plot in *Word*. Do the residuals appear to follow a multivariate normal distribution?

d) Do the MANOVA partial F test where the reduced model deletes X_3 and X_4.

e) Do the F_2 test.

f) Do the MANOVA F test.

C), 12.9 This problem examines multivariate linear regression on SAS Institute (1985, p. 146) Fitness Club Data data with $Y_1 = chinups$, $Y_2 = situps$ and $Y_3 = jumps$. The predictors are $X_2 = weight$, $X_3 = waist$ and $X_4 = pulse$.

a) The R command for this part make the response and residual plots for each of the three variables. Click the rightmost mouse button and highlight *Stop* to advance the plot. When you have the response and residual plots for one variable on the screen, copy and paste the three plots into *Word*. Do this three times, once for each response variable. Are there any outliers?

b) The R command for this plot makes a DD plot of the residuals and adds the lines corresponding to the three prediction regions of Section 5.2. The robust cutoff is larger than the semiparametric cutoff. Place the plot in *Word*. Are there any outliers?

D), 12.10 This problem uses the *mpack* function `mregsim` to simulate the Wilk's Lambda test, Pillai's trace test, Hotelling Lawley trace test, and Roy's largest root test for the F_j tests and the MANOVA F test for multivariate linear regression. When `mnull = T` the first row of $\boldsymbol{B}$ is $\mathbf{1}^T$ while the remaining rows are equal to $\mathbf{0}$. Hence the null hypothesis for the MANOVA F test is true. When `mnull = F` the null hypothesis is true for $p = 2$, but false for $p > 2$. Now the first row of $\boldsymbol{B}$ is $\mathbf{1}^T$ and the last row of $\boldsymbol{B}$ is $\mathbf{0}$. If $p > 2$, then the second to last row of $\boldsymbol{B}$ is (1, 0, ..., 0), the third to last row is (1, 1, 0, ..., 0) etcetera as long as the first row is not changed from $\mathbf{1}^T$. First m iid errors $\boldsymbol{z}_i$ are generated such that the m errors are iid with variance σ^2. Then $\boldsymbol{\epsilon}_i = \boldsymbol{A}\boldsymbol{z}_i$ so that $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T = ((\sigma_{ij}))$ where the diagonal entries $\sigma_{ii} = \sigma^2[1 + (m-1)\rho^2]$ and the off diagonal entries $\sigma_{ij} = \sigma^2[2\rho + (m-2)\rho^2]$ where $\rho = 0.10$. Terms like *Wilkcov* give the percentage of times the Wilk's test rejected the $F_1, F_2, ..., F_p$ tests. The $mancv wcv pcv hlcv rcv fcv output gives the percentage of times the 4 test statistics reject the MANOVA F test. Here hlcov and fcov both correspond to the Hotelling Lawley test using the formulas in problem A).

5000 runs will be used so the simulation will take several minutes. Sample sizes $n = 10\min(m,p), n = 10\max(m,p)$ and $n = 10mp$ were interesting. Want coverage near 0.05 when H_0 is true and coverage close to 1 for good power when H_0 is false. Multivariate normal errors were used in a) and b) below.

a) Copy the coverage parts of the output produced by the *R* commands for this part. Used $n = 20, m = 2, p = 4$. Here H_0 is true except for the F_1 test. Wilk's and Pillai's tests had low coverage < 0.05 when H_0 was false. Roy's test was good for the F_j tests but why was Roy's test bad for the MANOVA F test?

b) Copy the coverage parts of the output produced by the *R* commands for this part. Used $n = 20, m = 2, p = 4$. Here H_0 is false except for the F_4 test. Which two tests seem to be the best for this part?

E), 12.11 This problem uses the *mpack* function `mpredsim` to simulate the prediction regions for $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ for multivariate regression. With 5000 runs this simulation takes several minutes. The *R* command for this problem generate iid lognormal errors then subtract the mean producing $\boldsymbol{z}_i$. Then the $\boldsymbol{\epsilon}_i = \boldsymbol{A}\boldsymbol{z}_i$ are generated as in problem D). Used n=100, m=2, and p=4. The nominal coverage of the prediction region is 90%, and 92% of the training data is covered. The ncvr output gives the coverage of the nonparametric region. What was ncvr?