Math 583 HW 9 Fall 2020 Due Friday, Nov. 6.

Quiz 9 on Wednesday will have problems on this HW. Final: Monday, Dec. 7, 8-10 AM. Problem numbers are from Olive (2020). Do the source commands from homework 4. You need to install 2 *R* packages with the following *R* commands that are near the top of the *R* homework file. Some computers in the Math lab should have these packages. Type "library(glmnet)" and libraary(leaps) to check if packages were installed.

```
install.packages("glmnet")
install.packages("leaps")
```

7.16. The *R* program generates data satisfying the MLR model

$$Y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^T = (1, 1, 0, 0)$.

a) Copy and paste the commands for this part into *R*. The output gives $\hat{\boldsymbol{\beta}}_{OLS}$ for the OLS full model. Give $\hat{\boldsymbol{\beta}}_{OLS}$. Is $\hat{\boldsymbol{\beta}}_{OLS}$ close to $\boldsymbol{\beta} = 1, 1, 0, 0)^T$?

b) The commands for this part bootstrap the OLS full model using the residual bootstrap. Copy and paste the output into *Word*. The output shows $T_j^* = \hat{\boldsymbol{\beta}}_j^*$ for $j = 1, ..., 5$.

c) $B = 1000$ T_j^* were generated. The commands for this part compute the sample mean $\overline{T}^*$ of the T_j^*. Copy and paste the output into *Word*. Is $\overline{T}^*$ close to $\hat{\boldsymbol{\beta}}_{OLS}$ found in a)?

d) The commands for this part bootstrap the forward selection using the residual bootstrap. Copy and paste the output into *Word*. The output shows $T_j^* = \hat{\boldsymbol{\beta}}_{VS,j} = \hat{\boldsymbol{\beta}}^*_{I_{min},0,j}$ for $j = 1, ..., 5$. The last two variables may have a few 0s.

e) $B = 1000$ T_j^* were generated. The commands for this part compute the sample mean $\overline{T}^*$ of the T_j^* where T_j^* is as in d). Copy and paste the output into *Word*. Is $\overline{T}^*$ close to $\boldsymbol{\beta} = (1, 1, 0, 0)$?

7.19. 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.

7.20. This problem is like Problem 7.19, except elastic net is used instead of lasso.

a) Copy and paste the commands for this problem into *R*. Include the elastic net 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 elastic net response plot in *Word*. This did elastic net 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. (Problem 7.19 c) shows the DD plot for the data.)

Table 1: Bootstrapping Forward Selection, $n = 100, p = 5, \psi = 0, B = 1000$

		β_1	β_2	β_3	β_4	β_5	test
reg	cov	0.95	0.93	0.93	0.93	0.94	0.93
	len	0.658	0.672	0.673	0.674	0.674	2.861
vs	cov	0.95	0.94	0.998	0.998	0.999	0.993
	len	0.661	0.679	0.546	0.548	0.544	3.11

7.22. This simulation is similar to that used to form Table 1. Since 1000 runs are used, coverage in [0.93,0.97] suggests that the actual coverage is close to the nominal coverage of 0.95. **The simulation may take 20 minutes.**

The model is $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + e$ where $\boldsymbol{\beta}_S = (\beta_1, \beta_2, ..., \beta_{k+1})^T = (\beta_1, \beta_2)^T$ and $k = 1$ is the number of active nontrivial predictors in the population model. The output for *test* tests $H_0 : (\beta_{k+2}, ..., \beta_p)^T = (\beta_3, ..., \beta_5)^T = \mathbf{0}$ and H_0 is true. The output gives the proportion of times the prediction region method bootstrap test fails to reject H_0. The nominal proportion is 0.95.

After getting your output, make a table similar to Table 1 with 4 lines. Two lines are for reg (the OLS full model) and two lines are for vs (forward selection with I_{min}). The β_i columns give the coverage and lengths of the 95% CIs for β_i. If the coverage ≥ 0.93, then the shorter CI length is more precise. Were the CIs for forward selection more precise than the CIs for the OLS full model for β_3 and β_4? (In Table 1, forward selection was more precise for $\beta_i, i = 3,4,5$.)