Math 583 HW 8 2023 Due Wednesday, Oct. 25.

**A)** Let $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$. Then there are several estimators $\hat{\boldsymbol{\beta}}_E$ such that, under regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\beta}}_E - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2 \boldsymbol{V})$ where $\boldsymbol{X}^T\boldsymbol{X}/n \xrightarrow{P} \boldsymbol{V}^{-1}$. These estimators include OLS, ridge regression, lasso, elastic net, and Liu type estimators. Often we want to test $H_0 : \boldsymbol{A\beta} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{A\beta} \neq \boldsymbol{0}$.

a) What is $\boldsymbol{A}$ for testing $H_0 : \beta_2 = 0$?

b) What is $\boldsymbol{A}$ for testing $H_0 : \beta_{p-2} = \beta_{p-1} = \beta_p = 0$? Assume $p > 3$.

**B)** Let

$$\tilde{\sigma}^2_X = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

be the method of moments estimator of the variance of $X$. Let the standardized data

$$W_i = \frac{X_i - \overline{X}}{\tilde{\sigma}_X}$$

be the estimated z-score of each $X_i$ using the method of moments estimator of the standard deviation $\tilde{\sigma}_X = \sqrt{\tilde{\sigma}^2_X}$.

a) Show $\sum_{i=1}^n W_i = 0$.

b) Show $\sum_{i=1}^n W_i^2 = n$.

**C)** Consider choosing $\hat{\boldsymbol{\eta}}$ to minimize the criterion

$$Q(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W\eta})^T(\boldsymbol{Z} - \boldsymbol{W\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1}|\eta_i|^j$$

where $\lambda_{1,n} \geq 0$, $a > 0$, and $j > 0$ are known constants.

a) Which value of $j$ corresponds to lasso?

b) Which value of $j$ corresponds to ridge regression?

c) Which value of $j$ corresponds to OLS (in that $\hat{\boldsymbol{\eta}}_{OLS}$ minimizes the criterion)? Hint: See Exam 2 review 90).

**D)** This problem generates MLR models with heterogeneity, and then fits OLS (if $n > p + 5$) and OPLS models. The response plots are made and the OPLS fitted values are plotted versus the OLS fitted values.

a) i) Copy and paste the commands for this part into *R*. Right click the mouse and pick Stop for the first two plots to see the three plots.

ii) Were the OLS fitted values and OPLS fitted values highly correlated?

b) i) Copy and paste the commands for this part into *R*. Right click the mouse and pick Stop for the first two plots to see the three plots.

ii) Were the OLS fitted values and OPLS fitted values highly correlated?

c) i) Copy and paste the commands for this part into *R*. This is high dimensional data so only the OPLS response plot is generated.

ii) Is the response plot linear (the variance is not constant)?

d) i) Copy and paste the commands for this part into *R*. This is high dimensional data so only the OPLS response plot is generated.

ii) Is the response plot linear (the variance is not constant)?

**E)** This problem is like homework 7 E), except lasso and lasso variable selection are used for Poisson regression instead of MLR.

For the homework, we will use 100 runs instead of 5000 runs, but the simulation still takes a few minutes. With 100 runs, PI coverage between 0.89 and 1.0 gives no reason to believe that the actual coverage is not close to the nominal coverage of 0.95.

a) Copy and paste the commands for this part into *R*. Then make a table similar to the table in HW7. Here $n = 100, p = 100, k = 1$.

b) Copy and paste the commands for this part into *R*. Then make a table similar to the table in HW7. Here $n = 100, p = 100, k = 10$. Now there is much more underfitting, but lasso picks models good for prediction, so some of the PIs have adequate coverage.