Math 583 HW 6 2023 Due Wednesday, Oct. 11.

1) Let the ridge regression criterion $Q_R(\boldsymbol{\beta}) = RSS(\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta}$. Then $\bigtriangledown Q_R(\boldsymbol{\beta}) = \bigtriangledown RSS(\boldsymbol{\beta}) + \bigtriangledown \lambda \boldsymbol{\beta}^T \boldsymbol{\beta}$. In homework 5, you showed $\bigtriangledown RSS(\boldsymbol{\beta}) = -2\boldsymbol{X}^T \boldsymbol{Y} + 2\boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\beta}$. Find $\bigtriangledown Q_R(\boldsymbol{\beta})$, set $\bigtriangledown Q_R(\boldsymbol{\beta}) = \boldsymbol{0}$ and solve for $\boldsymbol{\beta}$. Show that the solution
$\hat{\boldsymbol{\beta}}_R = (\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^T \boldsymbol{Y}$.

2) The Hebbler (1847) data was collected from $n = 26$ districts in Prussia in 1843. Let $Y =$ the *number of women married to civilians* in the district with a constant and predictors $x_1 =$ the *population of the district in 1843*, $x_2 =$ the *number of married civilian men* in the district, $x_3 =$ the *number of married men in the military* in the district, and $x_4 =$ the *number of women married to husbands in the military* in the district. Sometimes the person conducting the survey would not count a spouse if the spouse was not at home. Hence Y and x_2 are highly correlated but not equal. Similarly, x_3 and x_4 are highly correlated but not equal. Then $\hat{\boldsymbol{\beta}}_{OLS} = (0.00035, 0.9995, -0.2328, 0.1531)^T$, forward selection with OLS and the C_p criterion used $\hat{\boldsymbol{\beta}}_{I,0} = (0, 1.0010, 0, 0)^T$, lasso had $\hat{\boldsymbol{\beta}}_L = (0.0015, 0.9605, 0, 0)^T$, lasso variable selection $\hat{\boldsymbol{\beta}}_{LVS} = (0.00007, 1.006, 0, 0)^T$, $\hat{\boldsymbol{\beta}}_{MMLE} = (0.1782, 1.0010, 48.5630, 51.5513)^T$, and $\hat{\boldsymbol{\beta}}_{OPLS} = (0.1727, 0.0311, 0.00018, 0.00018)^T$. Note that the last 5 estimators have $\hat{\boldsymbol{\beta}}_3 \approx \hat{\boldsymbol{\beta}}_4$, and all six estimators produce fitted values $\hat{Y}_i$ that are very highly correlated with the response Y_i. For OPLS, the largest $|\hat{\beta}_i|$ corresponds to the largest $|\widehat{Cov}(x_i, Y)|$, and $\hat{\beta}_i/\hat{\beta}_j = \widehat{Cov}(x_i, Y)/\widehat{Cov}(x_j, Y)$ does not depend on any other variables that may be in or out of the model. Similar properties hold for the OPLS population β_i. The MMLE did not use standardized predictors.

If $Y = \alpha_{OLS} + \boldsymbol{\beta}_{OLS}^T \boldsymbol{x} + e$, it seems likely that $\boldsymbol{\beta}_{OLS} = (0, 1, 0, 0)^T$. Then $\hat{\boldsymbol{\beta}}_{OLS}$ and the variable selection estimators appear to be estimating $\boldsymbol{\beta}_{OLS}$, as expected from low dimensional variable selection theory (although n = 26 is quite small).

a) Which value of x_i has the largest value of $\widehat{Cov}(x_i, Y)$?

b) Which two estimators $\hat{\boldsymbol{\beta}}_E$, do not appear to be estimating $(0, 1, 0, 0)^T$?

3) The data below are a sorted residuals from a lasso regression where $n = 1000$ and $p = 17$. Find shorth(997) of the residuals.

```
number      1     2     3     4   ...  997   998   999   1000
residual -3.28 -3.06 -3.04 -2.96 ...  2.66  2.71  2.81  3.62
```

4) Suppose $n = 15$ and 5-fold CV is used. Suppose observations are measured for the following people. Use the output below to determine which people are in the first fold.

```
folds:  4   1   4   3   5   3   5   1   1   3   2   5   2   4   2
```

1) Athapattu, 2) Azizi, 3) Cralley 4) Gallage, 5) Godbold, 6) Gunawardana, 7) Houmadi, 8) Mahappu, 9) Pathiravasan, 10) Rajapaksha, 11) Ranaweera, 12) Safari, 13) Senarathna, 14) Thakur, 15) Ziedzor

5) When doing a PI or CI simulation for a nominal $100(1-\delta)\% = 95\%$ interval, there are m runs. For each run, a data set and interval are generated, and for the ith run $Y_i = 1$ if μ or Y_f is in the interval, and $Y_i = 0$, otherwise. Hence the Y_i are iid Bernoulli$(1-\delta_n)$ random variables where $1-\delta_n$ is the true probability (true coverage) that the interval will contain μ or Y_f. The observed coverage (= coverage) in the simulation is $\overline{Y} = \sum_i Y_i/m$. The variance $V(\overline{Y}) = \sigma^2/m$ where $\sigma^2 = (1-\delta_n)\delta_n \approx (1-\delta)\delta \approx (0.95)0.05$ if $\delta_n \approx \delta = 0.05$. Hence

$$SD(\overline{Y}) \approx \sqrt{\frac{0.95(0.05)}{m}}.$$

If the (observed) coverage is within $0.95 \pm kSD(\overline{Y})$ the integer k is near 3, then there is no reason to doubt that the actual coverage $1-\delta_n$ differs from the nominal coverage $1-\delta = 0.95$ if $m \geq 1000$ (and as a crude benchmark, for $m \geq 100$). In the simulation, the length of each interval is computed, and the average length is computed. For intervals with coverage $\geq 0.95 - kSD(\overline{Y})$, intervals with shorter average length are better (have more precision).

a) If $m = 5000$ what is 3 SD$(\overline{Y})$, using the above approximation? Your answer should be close to 0.01.

b) If $m = 1000$ what is 3 SD$(\overline{Y})$, using the above approximation?

6) The smoothing spline simulation in Problem 4.7 compares the PI lengths and coverages of 3 large sample 95% PIs for $Y = m(x) + e$ and a single measurement x. Values for the first PI were denoted by scov and slen, values for 2nd PI were denoted by ocov and olen, and values for third PI by dcov and dlen. The average degrees of freedom of the smoothing spline was recorded as *adf*. The number of runs was 5000. The *len* was the average length of the PI and the *cov* was the observed coverage. One student got the following results.

Table 1: Results for 3 PIs

error type	n	95% slen	PI olen	95% dlen	PI scov	95% ocov	PI dcov	adf
5	100	18.028	17.300	18.741	0.9438	0.9382	0.9508	9.017

For the PIs with coverage ≥ 0.94, which PI was the most precise (best)?