HOMEWORK 2
BIOSTATISTICS 825
DUE SEPTEMBER 24TH, 2023

For questions 3–5, split the data into "learning" (50%), "validation" (25%) and "test" (25%) sets. Be sure to set the seed so your split can be reproduced. For these questions email me code that can be ran to reproduce all of your results. That is, email code that I can run to get exactly the results that you are reporting. All functions, packages (install.packages is not necessary) and data calls should be included.

For questions 1 and 2, you do not have to type up the solutions. A picture of a hand written solution is sufficient.

1. **(10 points)** Show that the ridge regression estimator $\beta_{RR}(k) = \{\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}\}^{-1}\boldsymbol{X}'\boldsymbol{Y}$ can be obtained by minimizing the loss function:

$$\phi(\beta) = \text{ESS}(\beta) + \lambda||\boldsymbol{\beta}||_2$$

where $\text{ESS}(\beta) = \sum_{i=1}^{n}(y_i - \boldsymbol{X}_i'\beta)^2$ and $||\boldsymbol{\beta}||_2 = \sum_{=1}^{r}\beta_j^2$. Assume the data are centered so that there is no intercept term.

2. **(10 points)** Complete exercise 3.6 in the online version of ESL.

3. **(25 points)** A wine connoisseur is interested in determine what factors impact expert opinion on the quality of red wine. He has collected information on acidity, citric acid levels, sugar content, chlorides, sulfur dioxide, density, pH, sulphate content, and alcohol content on 1000 Portuguese red wines and obtained and average quality rating from 5 different experts.

   We will use this "Wine" data set to complete the following questions.

   (a) **(5 points)** Using the learning set obtain the OLS estimates from the full model, and from 'forward' and 'backward' selection models. Calculate the EPE for full, forward and backward methods using the validation data.

   (b) **(5 points)** For all values k (there are 11 predictors in the data), find the subset of variables that minimizes the residual sums of squares using the training data. Then use the validation data to determine which k minimized the EPE.

   (c) **(5 points)** Analyze the training data using the forward stagewise algorithm. Use the `R` code from the example in class, not the built in function. For each step, calculate the EPE using the validation data. This may be time consuming if your step-size (i.e., `eps` in the example `R` code) is too small (might want to use `eps=0.01`).

   (d) **(5 points)** Using the best models from (a)–(c), reestimate the EPE using the test data. Make a table that has the regression coefficients for the **best** models from (a)–(c), and the EPE estimates (from validation and test data).

   (e) **(5 points)** Comment on the results in (d) and state which model would you recommend to use to predict wine quality.

4. **(25 points)** For this exercise we'll use the "Communities and Crime Data Set" (link to info). This data set contains variables related to violent crime. The goal is to predict the variable "ViolentCrimesPerPop". The first five variables in the dataset should not be used in the regression models.

   (a) **(5 points)** Analyze the data using PCR. Chose which K works the best based on the validation data.

   (b) **(5 points)** Analyze the data using PLS. Chose which K works the best based on the validation data.

   (c) **(5 points)** Compare the first linear combination for PCR and PLS. Plot the results and comment on the differences.

   (d) **(5 points)** Using the best models from (b) and (c), compare the results using the test data.

   (e) **(5 points)** Which model do you think should be used? You should base this on the results *and* interpretability.

5. **(30 points)** The dataset "`parkinsons_updrs.txt`" is composed of a range of biomedical voice measurements from 42 people with early-stage Parkinson's disease recruited to a six-month trial of a telemonitoring device for remote symptom progression monitoring. The recordings were automatically captured in the patient's homes. Columns in the table contain subject number, subject age, subject gender, time interval from baseline recruitment date, motor UPDRS, total UPDRS, and 16 biomedical voice measures. The attributes of the data are

   - subject number - Integer that uniquely identifies each subject
   - age - Subject age
   - sex - Subject gender '0' - male, '1' - female
   - test_time - Time since recruitment into the trial. The integer part is the number of days since recruitment.
   - motor_UPDRS - Clinician's motor UPDRS score, linearly interpolated
   - total_UPDRS - Clinician's total UPDRS score, linearly interpolated

   16 biomedical voice measures

   - Jitter(%), Jitter(Abs), Jitter:RAP, Jitter:PPQ5, Jitter:DDP - Several measures of variation in fundamental frequency
   - Shimmer, Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, Shimmer:APQ11, Shimmer:DDA - Several measures of variation in amplitude
   - NHR, HNR - Two measures of ratio of noise to tonal components in the voice
   - RPDE - A nonlinear dynamical complexity measure
   - DFA - Signal fractal scaling exponent
   - PPE - A nonlinear measure of fundamental frequency variation

Each row corresponds to one of 5,875 voice recording from these individuals. The main aim of the data is to predict the motor and total UPDRS scores ('motor_UPDRS' and 'total_UPDRS') from the 16 voice measures. Here we'll focus on total UPDRS score.

(a) **(5 points)** Center and scale the 16 voice measurements. Then for each voice measurement create a squared and cubic terms. Thus each variable will then have 3 columns (X, X^2, X^3). The total design matrix will have $48 = 3 \times 16$ columns.

(b) **(5 points)** Develop a ridge regression model of these data for six different values of λ. The six values of λ should be chosen so that $df(\lambda) \approx 8, 16, 24, 32, 40, 48$. Then chose the best value of λ based on the validation data.

(c) **(5 points)** Develop a LASSO model of these data for six different values of λ. The six values of λ should be chosen so that s (as defined in the notes) is $s \approx 0.15, 0.3, 0.45, 0.6, 0.75, 0.9$. Then chose the best value of λ based on the validation data.

(d) **(5 points)** Analyze the data using a grouped LASSO. Here you will have 16 groups based on the steps in (a). Describe how you identified the appropriate choice cutoff.

(e) **(5 points)** Compare the findings of all model to each other and the OLS estimates from the full model (using all 48 predictors). What are the similarities and differences.

(f) **(5 points)** What biomedical voice measures appear to have some impact on the outcome? Which model gives the 'best' answer to this question?