Homework 3

NAME: Your Name
DUE DATE: March 15th, 11:59 pm

Problem 1 [5 points]

Prove that the probability that an observation appears in a specific bootstrap sample is approximately 0.632. *Hint*: Calculate the probability that the first (or any other) bootstrap observation is a specific observation (e.g., the jth) from the original dataset.

Problem 2 (ISL 5.9) [8 points]

This question will concern the Boston housing data set, from the MASS library.

a) Based on this data set, provide an estimate for the population mean of medv. Call this estimate $\hat{\mu}$.

b) Provide an estimate of the standard error of $\hat{\mu}$. Interpret this result.

c) Now estimate the standard error of $\hat{\mu}$ using the bootstrap. How does this compare to your answer from (b)?

d) Based on your bootstrap estimate from (c), provide a 95 % confidence interval for the mean of medv. Compare it to the results obtained using t.test(Boston$medv).

e) Based on this data set, provide an estimate, $\hat{\mu}_{med}$, for the median value of medv in the population.

f) We now would like to estimate the standard error of $\hat{\mu}_{med}$. Unfortunately, there is no simple formula for computing the standard error of the median. Instead, estimate the standard error of the median using the bootstrap. Comment on your findings.

g) Based on this data set, provide an estimate for the tenth percentile of medv in Boston suburbs. Call this quantity $\hat{\mu}_{0.1}$. (You can use the **quantile()** function.)

h) Use the bootstrap to estimate the standard error of $\hat{\mu}_{0.1}$. Comment on your findings.

Problem 3 [5 points]

In this problem, you will explore potential bias when estimating test error using cross-validation.

a) Simulate a random vector y of length $n = 100$ in which each element follows a Bernoulli distribution with probability 0.5. Then simulate a random vector X as a matrix with n rows and $p = 1000$ columns in which each element also follows a Bernoulli distribution with probability 0.5.

b) For each predictor i, fit a simple linear regression model $Y = \beta_0 + \beta_1 X_i + \epsilon$ and record the corresponding R^2 value. Let i^* be the index with highest R^2 value.

c) Use 5-fold cross-validation to estimate the test error of the model $Y = \beta_0 + \beta_1 X_{i^*} + \epsilon$.

d) Is this reflective of the true test error? Explain your answer.

Problem 4 (ISL 6.5) [6 points]

It is well-known that ridge regression tends to give similar coefficient values to correlated variables, whereas the lasso may give quite different coefficient values to correlated variables. We will now explore this property in a very simple setting.

Suppose that $n = 2$, $p = 2$, $x_{11} = x_{12}$, $x_{21} = x_{22}$. Furthermore, suppose that $y_1 + y_2 = 0$ and $x_{11} + x_{21} = 0$ and $x_{12} + x_{22} = 0$, so that the estimate for the intercept in a least squares, ridge regression, or lasso model is zero: $\hat{\beta}_0 = 0$.

a) Write out the ridge regression optimization problem in this setting. Argue that in this setting, the ridge coefficient estimates satisfy $\hat{\beta}_1 = \hat{\beta}_2$.

b) Write out the lasso optimization problem in this setting. Argue that in this setting, the lasso coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are not unique. In other words, there are many possible solutions to the optimization problem. Describe these solutions.

Problem 5 [13 points]

The iod.csv file in the (Data Sets folder on Canvas) contains a large data set from two kidney disease studies (MDRD Modificiation of Diet in Renal Disease) and AASK African-American Study of Kidney Disease). The variables included are the following:

- **WEIGHT**: weight in kg
- **BMI**: body mass index
- **GFR**: glomerular filtration rate
- **UCRE**: urine creatinine
- **UUN**: urine urea nitrogen
- **UPHO**: urine phosphorus
- **SUN**: serum urea nitrogen
- **SCR**: serum creatinine
- **TCHOL**: total cholesterol
- **ALB**: albumin
- **HBA1C**: hemoglobin AIC
- **PHOS**: serum phosphorus
- **TRIG**: triglycerides
- **LDL**: low density lipoprotein (cholesterol)
- **HDL**: high density lipoprotein (cholesterol)
- **HB**: hemoglobin
- **MAP**: mean arterial pressure
- **UPRO**: urine protein
- **BSA**: body surface area
- **SODIUM**: sodium
- **GLUC**: glucose

- **BLACK**: black race (0/1)
- **HEIGHT**: height in cm
- **AGE**: age
- **FEMALE**: female (0/1)
- **CYS**: serum cystatin
- **DBP**: diastolic blood pressure
- **SBP**: systolic blood pressure
- **CRP**: C-reactive protein
- **DIAB**: diabetes
- **HBPSTATUS** high blood pressure (0/1)

Use GFR as your outcome and construct a predictive model for it using a) stepwise regression; b) ridge regression; c) lasso regression using cross-validation to choose your model form. For this exercise, just use the variables as they are given.

Describe your findings in clearly written text, tables, and figures discussing both the differences between the model findings and consistencies. Which factors are predictive? How well do the models predict the outcomes? Consider how you can get a good estimate of test error using cross-validation. At the end, refit using the best fitting model of each type on the whole dataset.