STAT 636, Fall 2018 - Assignment 5

1. Consider the `Life_Expectancy` data. Let the response Y be `Life.expectancy`, and consider the numeric variables `Alcohol`, `percentage.expenditure`, `Total.expenditure`, `GDP`, `Income.composition.of.resources`, and `Schooling` as the predictor variables; call these $x_1, x_2, \ldots, x_6$, respectively. For the purpose of predicting the value of Y, we will use linear regression models of the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_6 x_{6i} + \epsilon_i$$

where the ϵ_i are IID $N(0, \sigma^2)$, $i = 1, 2, \ldots, n$.

   (a) Using a usual least squares linear regression model (the `lm` function in `R`)(Using **na.omit** function to remove the row that there are some missing data.):
      i. Use leave-one-out cross-validation to estimate the MSE of your model.
      ii. Use bootstrap (1000 times) to estimate the standard deviation of your MSE estimate. (set.seed(2) before sampling)

   (b) Make a diagnostics to check assumptions of the linear regression model in number 1. Specifically, please
      i. Make Normal QQ plot of residuals. Does the residuals appear normally distributed?
      ii. Create scatterplots for each covariate, with the covariate on x-axis and residuals on y-axis. Do you see any problematic patterns?

   (c) Using regularized (lasso-based) linear regression (the `glmnet` and `cv.glmnet` functions from `glmnet` package in `R`, with `family='gaussian'` and `alpha` $= 1$):
      i. Based on cross-validation, using the `cv.glmnet` function, what is the optimal value of the tuning parameter `lambda`?
      ii. Use leave-one-out cross-validation (code it up yourself) and `glmnet` to estimate the MSE of the lasso model using the optimal tuning parameter.
      iii. Use bootstrap (again, code yourself, 1000 times) to estimate the standard deviation of your MSE estimate.
      iv. Compare the estimated β coefficients from the lasso model (using λ you gotten from (i)) to the least-squares model, and confirm that the lasso model's coefficient estimates have been "shrunken" toward 0.

2. Consider the `HOF` data. Let the response Y be the indicator for whether players are in the Hall of Fame (1 for "yes", 0 for "no"), and consider the numerical variables `H`, `HR`, and `AVG` as the predictor variables (using na.omit before your job); call these x_1, x_2, x_3, respectively. Randomly split the $n = 1780$ training data sample observations into 2/3 for training and 1/3 for testing (Select 2/3 from data whose HOF is 1 and also 2/3 from data whose HOF is 0 and combine these two. Use set.seed(2) before sampling).

   Let p_i be the probability player i is in the Hall of Fame, conditional on that player's predictor variable values:

$$p_i = Pr\left(Y_i = 1 | x_{1i}, x_{2i}, x_{3i}\right)$$

and consider the logistic regression model

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$$

for $i = 1, 2, \ldots, n$. After fitting the model, obtaining parameter estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$, we will predict Hall of Fame status for an individual with predictor variable values of x_1^*, x_2^*, x_3^* based on his estimated probability $\hat{p}$, where

$$\hat{p} = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \hat{\beta}_2 x_2^* + \hat{\beta}_3 x_3^*\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \hat{\beta}_2 x_2^* + \hat{\beta}_3 x_3^*\right)}$$

Specifically, we will predict that $Y = 1$ if $\hat{p} > \kappa$, for some choice of $\kappa \in [0, 1]$. Fit the model to the training data:

(a) Using the default choice of $\kappa = 0.5$:

   i. Report the misclassification rate, sensitivity, and specificity of your model when applied to the training data.

   ii. Report the misclassification rate, sensitivity, and specificity of your model when applied to the test data. Comment of the relationship between the performance measures, testing compared to training.

(b) Now use leave-one-out (LOO) cross validation (CV) to "tune" the model with respect to κ on the training data, using misclassification rate as the guiding performance measure.

   i. Report an ROC curve.

   ii. What is the optimal choice of κ, and what are the CV-based estimates of misclassification rate, sensitivity, and specificity that correspond to this choice of κ?

(c) Now we will perform the lasson on the training data.

   i. Use cross-validation to choose the tuning parameter λ.

   ii. Fit a lasso regression model on the training set and computing the coefficients using λ above.

   iii. Evaluate its misclassification rate, sensitivity and specificity on the training data set using λ which you get from above.