# 3AS/3AS4: Applied Statistics (Spring Term)

ASSIGNMENT 3 Date: February 5, 2020

**To be submitted by 5pm, February 21, 2020**
**Submit in a single .pdf file**

1. 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.

   (b) Argue that in this setting, the ridge coefficient estimates satisfy $\hat{\beta}_1 = \hat{\beta}_2$.

   (c) Write out the lasso optimization problem in this setting.

   (d) 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 in (iii).

2. Consider a simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ for $i = 1, \ldots, n$ with $E(\epsilon_i) = 0$ and $E(\epsilon_i^2) = \sigma^2$. Compute the *leave-one-out-cross-validation* (LOOCV) error for this model.

3. Consider the `Default` data from the ISLR package. We will use a logistic regression model to predict the probability of `default` using `income` and `balance`.

   (a) Using `summary()` and `glm()` functions, determine the estimated standard errors for the coefficients associated with `income` and `balance` in a multiple logistic regression model that uses both predictors.

   (b) Write a function `boot.fn()`, that takes as input the `Default` data set as well as an index of the observations, and that outputs the coefficient estimates for `income` and `balance` in the multiple logistic regression model.

   (c) Use the `boot()` function together with your `boot.fn()` function to estimate the standard errors of the logistic regression coefficients for `income` and `balance`.

   (d) Comment on the estimated standard errors obtained using the `glm()` function and using your bootstrap function.

4. Consider the `Wage` data set from the ISLR package.

   (a) Perform polynomial regression to predict `wage` using `age`. Use cross-validation to select the optimal degree $d$ for the polynomial. What degree was chosen, and how does this compare to the results of hypothesis testing using ANOVA? Make a plot of the resulting polynomial fit to the data.

   (b) Fit a locally constant kernel regression estimator to predict `wage` using `age`, and perform cross-validation to choose the optimal bandwidth $h$. Make a plot of the fit obtained. Does a choice of the kernel function affect the fit?