STAT 8330: **Homework 2**

Due on September 21, 2020 (10:01pm)

1. JWHT (2013), Chap 5, Prob. 8

2. JWHT (2013), Chap 5, Prob. 9

3. **Youtube analysis:** When a video is uploaded to youtube, the uploader sees the status of the video as "processing." During this time, the video is being converted (transcoded) from the original format to a format picked by youtube for display. The *time* it takes to do this conversion is important for both computational reasons (lots of videos added to youtube every day) as well as uploader convenience. The dataset `youtube.csv` consists of information contained in 8 variables which include input and output video characteristics along with their transcoding time and memory resource requirements while transcoding videos to different but valid formats.
   In particular, we have:

   - utime - total transcoding time in minutes (response variable)
   - duration - The length of the video, in seconds
   - size - Size of the input file, in MB
   - umem -Computer memory allocated for transcoding in MB
   - OutputPixels - Number of pixels in the output image in millions. For example a 1080p video has the pixel resolution 1920x1080, and therefore has 2,073,600 pixels

   Suppose you would like to be able to predict the time it will take you to upload videos to youtube for future reference. Considering a normal error regression model with an intercept, evaluate each possible 1, 2, and 3 variable model through 5-fold cross-validation (with MSE as the predictive evaluation). Use the `cv.glm` routine in R as described in the JWHT (2013) Chapter 5 lab. Report the CV value for each model and select the "best" model. Does the residuals from this model suggest any problems with the noraml error regression assumptions?

4. In the library `MASS` is a data set on a population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, who were tested for diabetes according to World Health Organization criteria. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases. There is a training data set (`Pima.tr`) and test data set (`Pima.te`) that each contain 8 variables, with the response being coded `Yes` (diabetic) or `No` (not diabetic) in the variable `type`. We are interested in classifying the women given the other 7 variables (`npreg` - number of pregnancies; `glu` - plasma glucose concentration; `bp` -diastolic blood pressure; `skin` - triceps skin fold thickness; `bmi` - body mass index; `ped` - diabetes pedigree function; `age`).

   Consider a logistic regression classification model to predict whether the subject is diabetic.

   (a) Consider a model with an intercept and all possible **2 variable models**. Use 5-fold cross-validation and report results from these models along with your choice for the best model. Note, if you use the `cv.glm` function, you have to specify the cost function! Use the one given in the binary example in the R help associated with the `cv.glm` function (repeat, don't use the default cost function here!). Interpret this cost function in terms of classification and logistic regression.

   (b) Use your best model to report the confusion matrix using the test data.

(c) Find the best logistic regression classification model possible using any combination of variables. Discuss if your model is better (or not) than the best 2-variable model you identified above.

(d) Find the best logistic regression classifiction model using deviance loss. Describe your procedure and compare your results to the loss function used in the previous parts of this problem.