# Homework 3 Functional Data Analysis (AMS 498B)

◇ This assignment should be handed in to gradescope.

◇ You can work together in groups but submit the homework separately.

◇ You should use RMarkdown to complete this assignment and render your work into pdf format. Be thrifty in what you include in your output and avoid listing extraneous matrices and vectors. RMarkdown provides much flexibility in what will and will not be included in its rendered report.

◇ All figures should be well-crafted and include labels for the axes and a title.

◇ Ten (10) points are given for each separate item, so some problems count for more than others.

◇ (GRAD) items are required by the 500 level students and count as extra credit for the 400 level students.

◇ *Show numerically* means just compute the specific case and verify the result. No general derivations are required.

400 1 (50), 2 (10) 60 total
500 1 (70), 2 (10) 80 total

1. Here are some exercises to become more familiar with ordinary least squares (OLS) fitting.
   For a linear model for fitting basis functions to data we use the notation:

   $$\mathbf{y} = \Phi\mathbf{c} + \mathbf{e}$$

   Let $\hat{\mathbf{c}}$ be the OLS estimate, given by $\hat{\mathbf{c}} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$. $\Phi\hat{\mathbf{c}}$ are the predicted values at the data points and $\mathbf{r} = \mathbf{y} - \Phi\hat{\mathbf{c}}$ are, of course, the residuals. The residuals are an estimate of $\mathbf{e}$.

   The Boulder daily minimum temperature measurements is the component `tmin` in the class R dataframe `Boulderdaily.rda`. See the script in `Lecture3.1OLSExample.R` for the R code and note `tmin` has been cleaned up to the vector `Y`. In this questions you should use the sine/cosine basis created as `Phi` used in getting `LSFit1`.

   (a) Recall that matrix-vector and matrix-matrix multiplication and in R uses `%*%`, the transpose of a matrix uses the function `t` , and the function

`solve` finds the inverse. Verify that the OLS estimates from using lm are the same as directly evaluating the formula for $\hat{\mathbf{c}}$ given above.

(b) For this OLS fit show numerically that $\mathbf{r}^T(\Phi\hat{\mathbf{c}}) = 0$ and also $\mathbf{r}$ has a mean of zero. Will the residuals always have a mean of zero for any basis?

(c) Show numerically that $\Phi^T\Phi$ is close to being a diagonal matrix.

(d) (GRAD) From HW 02 you showed that sin and cosine functions can be orthogonal based on an integral inner product. Explain how this analytical result is related to your finding in the previous question.

(e) Take a look at the two residual plots on page 5 of `Lecture3.1OLSExample.pdf`. Give an explanation for the patterns that you see in these.

(f) (GRAD) Explain how to model the patterns in the residuals from the previous question. (But you do not need to implement your idea just give a suggestion.)

(g) Make a plot that investigates whether the residual at a particular day depends on the residual from the previous day and comment on your results. This is a classic first step in a time series analysis – see the hint below. Since there are lots of points in this data set use the handy `bplot.xy` function to create boxplots instead of just a scatterplot.

*Hint:* To line up your residuals with the ones from the previous day.

```
# assume the residual vector is  res
N<- length( res)
res0<- res[ 2:N]
res1<- res[1:(N-1)]
# now res0 is today and res1 is previous day
```

2. This problem also refers to fitting the Boulder minimum temperatures. Let $K$ be the number of sine/cosine pairs used for the basis functions and also include the constant function in the basis set. Note in the example script $K = 6$ so there are $p = 2 \times 6 + 1 = 13$ total parameters. Vary fitting the seasonal cycle using OLS and $K$ from 1 to 20 and compute the (generalized) cross validation function:

$$GCV(K) = \frac{(1/n)\mathbf{r}^T\mathbf{r}}{(1 - p/n)^2}$$

where $\mathbf{r}$ are the residuals from the OLS fit, $p = 2K + 1$ are the number of parameters in the model (total number of basis functions/ columns of $\Phi$ ), and $n$ is the total number of observations.

(a) Make a plot of K against GCV(K) and indicate where this function has a minimum value.

(b) EXTRA CREDIT Based on the model where $K$ minimizes the GCV function add this fitted curve to the figure on the top of page 6 and comment on any differences between this case and when $K = 6$.