Homework 1
Stat 597a: Spatial Models
Due Thursday, September 28 in class

Hand in a paper copy of your solutions. You don't need to typeset them, but make sure the problems are presented in order. Any code you write that is not specifically requested in the questions should be included as an Appendix at the end.

1. Suppose we want to simulate a random vector $Y \sim N(\mu, \Sigma)$. If Σ is symmetric and positive definite, it can be represented using the Cholesky decomposition $\Sigma = LL^T$, where L is a lower triangular matrix. Consider the following algorithm for simulating Y:

   - Calculate the matrix L.
   - Sample $Z \sim N(0, I)$, where I is the $n \times n$ identity matrix.
   - Let $Y = \mu + LZ$.

   (a) Show that Y generated in this way has the correct distribution. You may use the fact that a linear function of a multivariate normal random variable is again multivariate normal; just show the mean and variance are correct.

   (b) Write a function or a few lines of code in `R` to implement this method for arguments `mu` and `Sigma`. You may use the built-in function `chol` for the Cholesky decomposition and `rnorm` to generate Z.

   (c) For a mean and covariance function of your choosing, use your code from (b) and make a few plots illustrating realizations of a Gaussian process on $[0, 1]$, but changing the different parameters in the model. These differences will be easier to see if you keep the same Z sample but just change `mu` and `Sigma`.

2. The file `CAtemps.RData` contains two `R` objects of class `SpatialPointsDataFrame`, called `CAtemp` and `CAgrid`. `CAtemp` contains average temperatures from 1961-1990 at 200 locations (latitude and longitude) in California in degrees Fahrenheit, along with their elevations in meters. `CAgrid` contains elevations in meters over a grid of locations. I've given you some code to get started with this data in `HW1.R`.

Consider the following model for the temperature data.

$$Y_i = \mu(s_i; \beta) + e(s_i; \sigma^2, \rho, \tau)$$

where $\mu(s;\beta) = \beta_0 + \beta_1 Longitude(s) + \beta_2 Latitude(s) + \beta_3 Elevation(s)$ and $e(s_i; \sigma^2, \rho, \tau)$ is a zero mean stationary Gaussian process with exponential covariance function.

Another way of writing this is as

$$Y_i = \mu(s_i; \beta) + Z(s_i; \sigma^2, \rho) + \epsilon_i$$

where now Z is a mean zero Gaussian process like e but without the nugget term, and the ϵ_i are iid $N(0, \tau^2)$, independent of Z. This is important because we want to predict $\mu(s_i; \beta) + Z(s_i; \sigma^2, \rho)$ *without* the measurement error.

(a) Using the `CAtemp` data, form a preliminary estimate of β using ordinary least squares and make a color plot of the residuals. Include your estimates and plot.

(b) Estimate the variogram nonparametrically and then fit the exponential variogram to it using weighted least squares. Make and include a plot of the nonparametric and parametric variogram functions. Also store your parameter estimates and report them.

(c) We will now form the GLS estimate of β by hand, rather than using the `gls` function. (This function doesn't handle longitude and latitude well, and I also want to give you some practice with matrix calculations in `R`.)

- Use the `rdist.earth` function in `fields` to create a matrix of distances (in miles) between pairs of locations in `CAtemp`.
- Create the covariance matrix, plugging in your estimates from the fitted variogram. *Hint: Sum two matrices, one without a nugget and one using the* `diag` *function to create the matrix* $\tau^2 I$.
- Invert the covariance matrix and store it for later reference.
- Create the X matrix. *Hint: Use* `cbind`.
- Put all the pieces together to form $\hat{\beta}_{GLS}$.

(d) Calculate and plot the EBLUP of $\mu + Z$ at the locations in `CAgrid`, plugging in your estimates from (b) and (c). Calculate and plot the (estimated) standard error of Z at each prediction location.