M647 Spring 2012 Assignment 3, due Friday Feb. 10

1. [10 pts] One natural observation ecologists have made is that the number of species in a reasonably isolated region typically depends on the area of the region. In particular, the relationship between number of species S and region area A often has the form

$$S = kA^{\gamma},$$

where k and γ are parameters. In the M-file *bwidata.m* (available on the course web site), you will find data for bird species on islands in the West Indies. Convert this equation into a form in which linear regression can be applied, and use this form to compute regression values for k and γ.

2a. [5 pts] An alternative approach to the analysis we carried out in class for predicting a son's height based on the heights of his parents would be to use a multivariate fit with

$$S = p_1 + p_2 M + p_3 F,$$

where S denotes the son's M denotes mother's height, and F denotes father's height. Use data stored in the M-file *heights.m* to find values for p_1, p_2, and p_3 for this fit. According to this model, which height is more significant for a son's height, the mother's or the father's? Write a MATLAB *anonymous* function for your model and evaluate it at $(M, F) = (60, 70)$; i.e., the case in which the mother is five feet tall and the father is five feet, ten inches. Estimate the standard deviation for your fit.

2b. [5 pts] For the same data as in Part (a) find parameter values for a multidimensional polynomial fit of the form

$$S = p_1 + p_2 M + p_3 F + p_4 M^2 + p_5 F M + p_6 F^2.$$

Write a MATLAB *anonymous* function for your model and evaluate it at $(M, F) = (60, 70)$. Estimate the standard deviation for your fit, compare it with the standard deviation from Part (a), and discuss which model you find preferable.

3. [10 pts] In this problem we'll collect a few straightforward observations about PDFs that we've found useful in class.

3a. Show that if X is a random variable with PDF $f_X(x)$, and we set $Y = cX$ for any constant $c > 0$, then Y has PDF $f_Y(y) = \frac{1}{c} f_X(\frac{y}{c})$.

3b. Show that if X and Y are any two independent continuous random variables with respective PDFs f_X and f_Y, then the PDF for $Z = X + Y$ is

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(z - x) f_Y(x) dx.$$

Note. This integral is a *convolution*, and we often write

$$f_X * f_Y(z) = \int_{-\infty}^{+\infty} f_X(z - x) f_Y(x) dx.$$

You can take as your starting point the observation

$$\int_a^b f_Z(z)dz = P(a \leq Z \leq b) = P(a \leq X + Y \leq b)$$
$$= \iint_{a \leq x+y \leq b} f_X(x) f_Y(y) dx dy.$$

3c. Show that if X and Y are independent Gaussian $N(\mu, \sigma^2)$ random variables, then $Z = X + Y$ is $N(2\mu, 2\sigma^2)$.

4a. [5 pts] Verify the identity

$$\mathrm{Cov}(\sum_{k=1}^{n} A_{ik} X_k, \sum_{l=1}^{n} A_{jl} X_l) = \sum_{k=1}^{n} \sum_{l=1}^{n} A_{ik} A_{jl} \mathrm{Cov}(X_k, X_l)),$$

for any $n \times n$ matrix A and any vector of random variables $\vec{X}$.

4b.[5 pts] Verify the identity

$$\mathrm{Cov}(A\vec{X}) = A\mathrm{Cov}(\vec{X})A^{tr},$$

for any $n \times n$ matrix A and any vector of random variables $\vec{X}$.

5. [10 pts] A random variable X is said to be distributed according to a *Rayleigh* distribution provided its PDF has the form

$$f(x; \theta) \quad = \quad \begin{cases} \frac{x}{\theta^2} e^{-\frac{x^2}{2\theta^2}} & x > 0 \\ 0 & x \leq 0 \end{cases},$$

where we generally take $\theta > 0$. (Since it only appears squared, its sign is irrelevant.)

5a. Compute $E[X]$ and $\mathrm{Var}[X]$. In these calculations, you should feel free to use the standard integral

$$\int_0^{+\infty} e^{-ax^2} dx = \frac{1}{2}\sqrt{\frac{\pi}{a}}, \quad a > 0$$

without justification. (Though it's well worth looking up how to compute this.)

5b. Given data $\{x_k\}_{k=1}^N$, find a maximum likelihood estimate for θ^2.

5c. Write down the MLE estimator for θ^2, and compute its expected value.