

Tutorial 12
December 4/5, 2008

1. An electric utility company tries to estimate the relation between the daily amount of electricity used by its customers, and the daily summer temperature. It has collected the following data:

| | | | | | | | | | | |
|-------------|----|----|----|----|----|----|----|----|----|----|
| Temperature | 96 | 89 | 81 | 86 | 83 | 73 | 78 | 74 | 76 | 78 |
| Electricity | 24 | 20 | 22 | 23 | 21 | 18 | 19 | 20 | 18 | 18 |

- (a) We wish to build a linear regression model of the form $y \approx \theta_0 + \theta_1 x$, where y denotes the electricity consumption, and x denotes the temperature. Derive the parameters θ_0 and θ_1 that minimize the sum of the squared residuals

$$\sum_{i=0}^n (y_i - \theta_0 - \theta_1 x_i)^2$$

over all θ_0 and θ_1 .

- (b) We calculate $\theta_0 = 1.21$ and $\theta_1 = 0.23$. If the temperature on a given day is 90 degrees, predict the amount of electricity consumed on that day.
2. The sample mean is by far the most commonly computed statistic, and the Weak Law of Large Numbers gives one of its key properties. Here we explore its performance as an estimator of an unknown deterministic parameter.

Let $Y_i = \theta + N_i$ for $i = 1, 2, \dots, k$. Assume that the N_i s are i.i.d, $\mathbf{E}[N_i] = 0$ and $\text{var}(N_i) = \sigma^2$ for each i . Think of θ as an unknown parameter of interest and of each Y_i as a noisy measurement of θ .

- (a) Let $M_k = \frac{1}{k} \sum_{i=1}^k Y_i$ denote the sample mean of the n measurements. Show that M_k converges in probability to θ .
- (b) $\mathbf{E}[(M_k - \theta)^2]$ is the mean-squared error (MSE) of the sample mean estimator. Express this MSE in terms of the variances of the N_i s.

As background for the remainder of this problem, recall that in the Bayesian case the LMS estimator of X given Y is necessarily the conditional mean of X , given Y . However, in classical statistics there is no comparably simple rule for finding the estimator with the smallest mean-squared error. In this example with zero-mean additive noise, we began by using the sample mean as our estimator for θ , but we will soon discover a better estimator for some noise types, i.e., we will develop an unbiased estimator with lower mean-squared error than the sample mean.

For the remainder of this problem we consider the interesting case when the noise is uniformly distributed over the interval $[-\delta, \delta]$

- (c) Calculate the mean squared error of the sample mean for estimating θ with this noise model.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2008)

- (d) Given a single measurement $Y_1 = y_1$ what is the smallest interval that the parameter θ must lie with certainty? Given two measurements $Y_1 = y_1$ and $Y_2 = y_2$ what is the smallest interval that parameter θ must lie with certainty ?
- (e) Generalize this result to find the smallest interval where θ must lie with certainty, given an arbitrary number k of measurements. Let's use the midpoint of this interval as our new nonlinear estimator for θ , i.e., $\hat{\Theta}_k$.
- (f) Is $\hat{\Theta}_k$ unbiased?
- (g) As a step toward finding the mean squared error of $\hat{\Theta}_k$, write the error

$$\hat{\Theta}_k - \theta$$

as a function of the values of the noise terms $N_1 = n_1, \dots, N_k = n_k$. Then calculate the joint probability density $f_{W,Z}(w, z)$ for the largest and the smallest noise terms:

$$W = \max_{1 \leq j \leq k} N_j \text{ and } Z = \min_{1 \leq j \leq k} N_j.$$

- (h) Calculate the mean squared error of $\hat{\Theta}_k$. Compare your answer to the one you found in part (c) for the linear estimator. You may find the following definite integral helpful:

$$\int_{-\delta}^{\delta} \int_{-\delta}^w (w+z)^2 (w-z)^{k-2} dz dw = \frac{2(2\delta)^{k+2}}{(k-1)(k)(k+1)(k+2)}, \quad k \geq 2.$$