JHU Engineering for Professionals
Applied and Computational Mathematics
Data Mining: 625.740

Homework for Module on Linear Regression

1. Show that, if x, y are jointly Gaussian, the regression of y on x is given by

$$E(y|x) = \frac{\alpha\sigma_y x}{\sigma_x} + \mu_y - \frac{\alpha\sigma_y \mu_x}{\sigma_x}, \quad \text{where } \Sigma = \begin{pmatrix} \sigma_x^2 & \alpha\sigma_x\sigma_y \\ \alpha\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}.$$

2. Given data $(X_1, Y_1), \ldots, (X_n, Y_n)$, consider the regression through the origin model

$$Y_i = \beta X_i + \nu_i, \quad \text{where } E(\nu_i|X_i) = 0 \text{ and } \mathrm{Var}(\nu_i|X_i) = \sigma^2.$$

   (a) Find $\hat{\beta}$, the least squares estimate for β.

   (b) Find the standard error of the estimate, $\sqrt{\mathrm{Var}(\hat{\beta})}$.

   (c) Find conditions that guarantee that the estimate is consistent:

$$\forall \varepsilon > 0, \quad P(|\hat{\beta} - \beta| > \varepsilon) \to 0 \text{ as } n \to \infty.$$

3. The columns in the file `polynomial_data.txt` are the X and Y values of a polynomial function $y = \sum_{k=0}^{n} a_k x^k$ with added Gaussian noise.

   (a) For each $n \in \{3, 4, 5\}$, fit an n^{th} degree polynomial to the data. What would you say is the most likely value of n?

   (b) Estimate the level of the noise.