

Homework 2: Linear Regression

INSTRUCTOR: DANIEL L. PIMENTEL-ALARCÓN

DUE 02/18/2022

DO NOT POLLUTE! AVOID PRINTING, OR PRINT 2-SIDED MULTIPAGE.

In class we studied the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^* + \boldsymbol{\epsilon}$, under the *homoskedastic* assumption $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$. In this homework you will derive the same results for the slightly more general *heteroskedastic* model where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^*)$. Each subproblem is worth 10 points.

Problem 2.1. Derive an expression for the coefficient vector $\boldsymbol{\theta}$ that minimizes the mean squared error, i.e.,

$$\arg \min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2.$$

Problem 2.2. Derive an expression for the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}^*$, i.e.,

$$\arg \max_{\boldsymbol{\theta}} \mathbb{P}(\mathbf{y}, \mathbf{X} | \boldsymbol{\theta}, \boldsymbol{\Sigma}^*)$$

Problem 2.3. What is the distribution of the MLE of $\boldsymbol{\theta}^*$?

Problem 2.4. Given a new sample with feature vector \mathbf{x} , what is the MLE of the response, \hat{y} ?

Problem 2.5. Given a new sample with feature vector \mathbf{x} , what is the distribution of the MLE \hat{y} ?

Problem 2.6. Derive an expression for the MLE of $\boldsymbol{\Sigma}^*$, i.e.,

$$\arg \max_{\boldsymbol{\Sigma}} \mathbb{P}(\mathbf{y}, \mathbf{X} | \boldsymbol{\theta}^*, \boldsymbol{\Sigma})$$

Problem 2.7. Consider the following vector \mathbf{y} , containing information about glucose level of four individuals, and the following data matrix \mathbf{X} containing information about height and weight of the corresponding individuals:

$$\mathbf{y} = \begin{bmatrix} 110 \\ 140 \\ 180 \\ 190 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 180 & 150 \\ 150 & 175 \\ 170 & 165 \\ 185 & 210 \end{bmatrix}.$$

Given these data

- What are your maximum likelihood estimates of $\boldsymbol{\Sigma}^*$ and $\boldsymbol{\theta}^*$?
- Given a new sample with feature vector $\mathbf{x} = [175 \ 170]^T$, what is the maximum likelihood estimate of its response \hat{y} ?
- Derive a 95% confidence interval for \hat{y} .
- Would you conclude that height is a significant feature for this model? Why?
- Would you conclude that weight is a significant feature for this model? Why?