Data Analytics for Coastal Systems

Georgios Boumis, Ph.D.
Assistant Professor
Department of Civil and Environmental Engineering
The University of Maine

Abstract. This course introduces statistical methods for analyzing environmental data, where students develop R programming skills through hands-on work on processing real-world data from Maine's coastal environments. Emphasis is given in datasets such as sea level observations, temperature and sea-surface pressure measurements, as well as seasonal wind speed records. The course advances systematically from basic data visualization and descriptive statistics to predictive modeling, probability distribution fitting via maximum likelihood estimation, and parametric uncertainty quantification. It culminates with the application of extreme-value theory and trend detection for evaluating rare coastal events, particularly within the context of non-stationarity due climate change-driven sea-level rise.

Contents

9 Predictive Modeling: Multiple Linear Regression **1**

9 Predictive Modeling: Multiple Linear Regression

An extension of simple linear regression, where in this case there are **more than one predictor variables**, is called **multiple linear regression (MLR)**. The MLR model can be written formally as below:

$$y_i = b + w_1 \times x_{1_i} + w_2 \times x_{2_i} + \cdots + w_k \times x_{k_i} + \epsilon_i, \tag{1}$$

where k is the number of predictor variables. The same model in matrix form can be rewritten as:

$$Y = X\beta + \epsilon, \tag{2}$$

where Y is an $n \times 1$ vector of responses, β is a $(k+1) \times 1$ parameter vector, X is an $n \times (k+1)$ matrix of predictors, and ϵ is an $n \times 1$ vector of residuals. The estimator of β is then given by minimizing the sum of squared errors in matrix form:

$$\hat{\beta} = \underset{\beta}{\text{argmin}}[(Y - X\beta)^T(Y - X\beta)] = (X^TX)^{-1}X^TY, \tag{3}$$

where A^T denotes the transpose of matrix A, while A^{-1} denotes the inverse of matrix A. Note that for $\hat{\beta}$ to be identifiable, the matrix X^TX must invertible. When predictor variables are highly correlated with each other, X^TX becomes close to non-invertible and $\hat{\beta}$ becomes unstable, meaning that small changes in the data can cause large changes in $\hat{\beta}$. This issue is called **multicollinearity**. A quick way to diagnose multicollinearity is through the so-called **variance inflation factor (VIF)**. The VIF for predictor j can be computed as:

$$\text{VIF}_j = \frac{1}{1 - \text{R}_j^2}, \tag{4}$$

where R_j^2 is the coefficient of determination (see **Chapter 8**) obtained from regressing the j^{th} predictor on all other $k-1$ predictors. *A value of VIF ≥ 5 indicates a linearly dependent predictor variable on all other predictors.*

As with simple linear regression, the goodness-of-fit of a MLR model can still be evaluated by computing the coefficient of determination. However, when we add more than $k = 1$ predictor variables, the coefficient of determination always increases (or, at worst, remains the same), even if these variables are irrelevant or noise. *Hence, this property makes R^2 unreliable for model comparison when models have different numbers of predictors.* The **adjusted coefficient of determination R^2_{adj}** provides a solution to this problem, which can be used for model intercomparison with different number of predictor variables. It penalizes model complexity and its value can decrease when we add uninformative predictors - its formula is shown below:

$$\text{R}^2_{\text{adj}} = 1 - \frac{\frac{\Sigma_{i=1}^n (y_i - \hat{y}_i)^2}{n-k-1}}{\frac{\Sigma_{i=1}^n (y_i - \bar{y})^2}{n-1}}. \tag{5}$$

Figure 1 shows a 3D scatterplot where we visualize the same storm surge vs. surface pressure data from Chapter 8, but here we add an extra dimension by also showing the respective wind speed measurements. A MLR model for these data can thus be expressed as:

$$\text{Storm\_Surge} = b + w_1 \times \text{Surface\_Pressure} + w_2 \times \text{Wind\_Speed}. \tag{6}$$

For these data, the multiple linear regression parameters are found to be $\hat{b} = 6.77$, $w_1 = -6.71\text{e-}05$, and $w_2 = 0.0089$, while the adjusted coefficient of

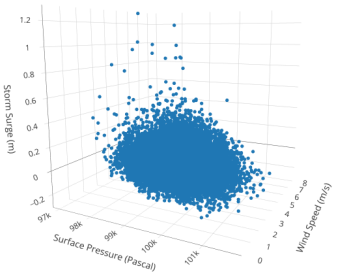

Figure 1: 3D Scatterplot of storm surge vs. surface pressure vs. wind speed.

determination is $R^2_{adj} = 0.21$ (vs. $R^2_{adj} = 0.19$ for the simple linear regression with surface pressure being the only predictor). The VIF factor is found to be ~ 1, meaning that there is no collinearity between these two predictors.

In the case of MLR, obtaining a test statistic (t) for conducting a hypothesis test about the effect of the k^{th} predictor (w_k) on the response variable is not straightforward. We will exclusively rely on software for this. *However, we still need to check that the assumptions of normality and homoscedasticity of residuals hold, i.e, that $e_i \sim N(0, \sigma^2)$.*