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

8 Predictive Modeling: Correlation and Regression

In the course so far, we have focused on analyzing a single continuous random variable, typically denoted as X, which follows some probability distribution. Our objectives have been two-fold: first, to conduct hypothesis tests about X, and second, to identify an appropriate probability distribution that characterizes X, enabling us to estimate probabilities for both observed and unobserved values of the variable. From now on, we will instead focus on **two continuous random variables simultaneously**, which we call Y and X, respectively. Hence, we will now deal with **pairs of measurements** of both variables which we denote as $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$, where n is the sample size (common for both variables). The aim of **predictive modeling** is to establish a robust statistical relationship between the two continuous random variables so that we can predict Y from measurements of X. Predictive modeling can be performed with statistical methods such as **regression**. However, before constructing a regression model, we need to examine the **correlation** between Y and X, i.e., their strength of association.

8.1 Pearson's Correlation Coefficient

Imagine now that we have a set of pairs of measurements of storm surge (Y) and surface pressure (X). From a theoretical physics point of view, we expect that storm surge should increase as surface pressure decreases. We can graphically represent this pattern by creating a scatterplot, where we visualize these pairs of measurements (as shown in Figure 1). The "negative" relationship shown in Figure 1 can be summarized numerically by computing the so-called **covariance**. The covariance is a sample statistic that can be calculated from the pairs of measurements of two variables and describes whether the two variables "covary". The covariance for continuous random variables Y and X (cov_{yx}) can be computed as follows:

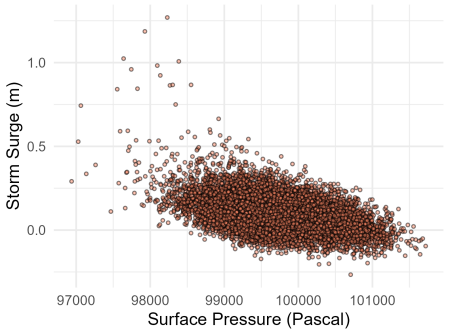

Figure 1: Scatterplot of storm surge vs. surface pressure.

$$cov_{yx} = \frac{1}{n-1}\Sigma_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}), \tag{1}$$

where $\bar{y}$ and $\bar{x}$ are the sample means of Y and X, respectively. The sign of cov_{yx} tells us the direction of the relationship between Y and X, while its absolute value indicates the strength of this relationship; the greater the absolute value, the stronger the relationship. For the data shown in Figure 1 the covariance is $cov_{yx} = -26.90$.

It is obvious that the absolute value of cov_{yx} is dependent on the absolute values of y and x, meaning that for different data the scale of cov_{yx} will be much different. Even for the same data but with different units, the scale of cov_{yx} can change dramatically. For example, the covariance of the data shown in Figure 1, but with storm surge now measured in centimeters and surface pressure in kilopascals, is $cov_{yx} = -2.69$. A standardized way to measure covariation, which does not depend on the scale of variables Y and X, is the so-called **Pearson's correlation coefficient (r)**. The Pearson's correlation coefficient takes values between -1 and 1 $(-1 \leq r \leq 1)$ and can be estimated from the data as follows:

$$cor_{yx} = r = \frac{\Sigma_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{cov_{yx}}{s_y s_x}, \tag{2}$$

where s_y and s_x are the standard deviations of Y and X, respectively. Values of r close to 1 indicate a strong positive correlation, while values of r close to -1 suggest a strong negative correlation. *It is important to remember that the

Pearson's correlation coefficient measures the strength of **linear** relationship between Y and X.* Therefore, if $Y = b + wX$ (perfect linear relationship), where b and w are constants, then $|r| = 1$. The Pearson's correlation coefficient for the data shown in Figure 1 is $r = -0.44$.

8.2 Linear Regression

The most popular statistical method for modeling a relationship between a **response** variable Y and a **predictor** variable X is to fit a linear equation to the pair of measurements (y_i, x_i). This model is called **linear regression** and can be written formally as below:

$$y_i = b + w \times x_i + \epsilon_i, \tag{3}$$

where y_i is the i^{th} measurement of the response variable, x_i is the i^{th} measurement of the predictor variable, b is the **bias** parameter, w is the **weight** parameter, and ϵ_i is the error term (deviation from line). Parameters b and w are also called **intercept** and **slope**, respectively, and need to be estimated from the data. The way to estimate parameters of a linear regression model is by minimizing the **sum of squared errors (SSE)**:

$$\text{SSE} = \Sigma_{i=1}^{n} \epsilon_i^2 = \Sigma_{i=1}^{n} (y_i - b - w \times x_i)^2. \tag{4}$$

Taking the partial derivatives of SSE with respect to the parameters $(\frac{\partial \text{SSE}}{\partial b}, \frac{\partial \text{SSE}}{\partial w})$ and setting them equal to zero yields a 2×2 system of equations which can be solved algebraically. The estimators for b and w are given as follows:

$$\hat{w} = \frac{\Sigma_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r \times \frac{s_y}{s_x}, \tag{5}$$

$$\hat{b} = \bar{y} - \hat{w} \times \bar{x}. \tag{6}$$

The linear regression estimate for the i^{th} measurement (i.e., $\hat{b} + \hat{w} \times x_i$) is usually denoted as $\hat{y}_i$. The goodness-of-fit of a linear regression model can be evaluated by computing the so-called **coefficient of determination R^2**. This coefficient measures the proportion of variance in Y explained by X and can be calculated as follows:

$$\text{R}^2 = 1 - \frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{\Sigma_{i=1}^{n}(y_i - \bar{y})^2}. \tag{7}$$

The coefficient of determination is between 0 and 1 ($0 \leq \text{R}^2 \leq 1$), with values close to 1 indicating that all points fall on the line (perfect fit). Figure 2 shows a linear regression line fit to the pair of measurements of storm surge and surface pressure from Figure 1. For these data, the linear regression parameters are found to be $\hat{b} = 6.974$ and $\hat{w} = -6.899\text{e-}05$, while the coefficient of determination is $\text{R}^2 = 0.19$.

How do we know that surface pressure has indeed a statistically significant effect on storm surge?

When fitting a linear regression model, many statistical computing environments, including R, will automatically run a hypothesis test to check if the effect of the predictor variable on the response variable is statistically significant. In other words, they will check the null hypothesis that $w = 0$, against the alternative hypothesis that $w \neq 0$. *This hypothesis test is based on the fundamental assumptions of **normality** and **homoscedasticity** of the residuals ϵ_i. Essentially, we want to make sure that $\epsilon_i \sim N(0, \sigma^2)$, or equivalently, that $\frac{\epsilon_i}{\sigma} \sim N(0, 1)$.* Thus, after we fit a linear regression model, it is wise to check

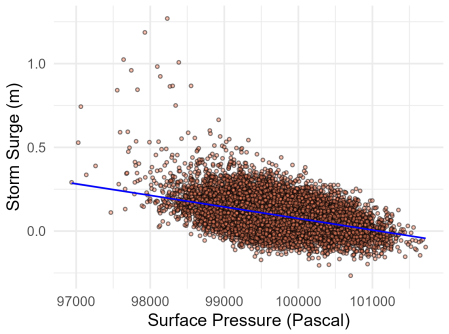

Figure 2: Scatterplot of storm surge vs. surface pressure. The blue line shows the linear regression fit.

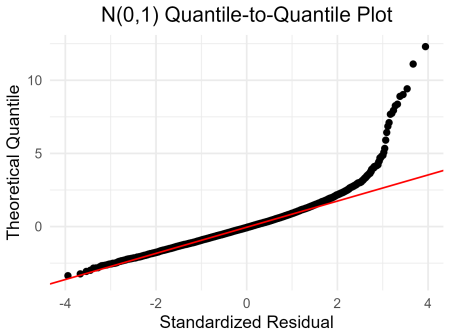

Figure 3: Quantile-to-quantile plot of standardized residuals obtained after fitting a linear regression model to the pair of measurements of storm surge and surface pressure from Figure 1.

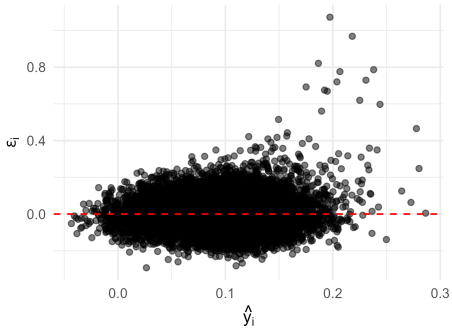

Figure 4: Scatterplot of residuals vs. estimates obtained after fitting a linear regression model to the pair of measurements of storm surge and surface pressure from Figure 1.

if these assumptions hold. Figures 3 and 4 illustrate visual checks of these assumptions for the linear regression model fit to the pair of measurements of storm surge and surface pressure from Figure 1. If these assumptions hold, we can obtain a test statistic as shown below:

$$t = \frac{\hat{w}}{\sqrt{\frac{\text{SSE}}{(n-2)\times\Sigma_{i=1}^{n}(x_i-\bar{x})^2}}}, \tag{8}$$

which follows the t-distribution with $n-2$ degrees of freedom. The effect is statistically significant if $2 \times$ p-value ≤ 0.05.