MATH221, Week 7

Regression Notation

Introduction

There is a variety of notation one must know to make sense of correlation and regression.

Notation

m = slope of regression or best-fit line

b = y-intercept of regression or best fit line

$\hat{y}$ = predicted or estimated y-value from the regression or best fit-line for a given x-value

$\hat{y} = mx + b$ is the equation of the regression line or line of best fit

Note: The regression line or line of best – fit is the linear equation of the form y = mx + b that most closely approximates the y-values in the data set from the x-values in the data set. One uses the linear correlation coefficient and the coefficient of determination to understand how good the approximates will be so that one can decide if the equation of the regression line will be useful in approximating values of y for values of x not in the given data set but within the domain of the given data set. The equation of the regression line can never be used to find meaningful approximations of y-values for x values outside the domain of the data set

Domain is the independent or x – values for which the given data set is defined. If the smallest x value in the data set were 10 and the largest were 57. We would say that the domain of the data set is all x-values between 10 and 57 inclusive. The equation of the regression line would provide meaningful estimates of the dependent or y variable using values of x within this domain.

x_1 = the first value the x-variable takes on from the data set.

y_1 = the y-value of a data point corresponding to x_1

(x_1, y_1)

x_i = the ith value the x-variable takes on from the data set.

y_i = The y-value of a data point corresponding to x_i

(x_i, y_i)

$\bar{x}$ = mean of the x-values in the given data set = $\frac{\sum x}{n}$.

$\bar{y}$ = mean of the y-values in the given data set = $\frac{\sum y}{n}$