1.4: Data Exploration

Dr. Bean - Stat 5100

Consider the linear model

$$Y = \beta_0 + \beta_1 X_i + \epsilon_i$$

In this example, estimates of the values of β_0 and β_1 are obtained using SAS for all four scenarios. The estimated models all have identical form, with identical measures of model goodness (which we will learn about in Handouts 2.2 and beyond).

$$\hat{Y} = 3 + 0.5X$$

Using the results of Figure 1, which models are appropriate, and which are inappropriate? Why?

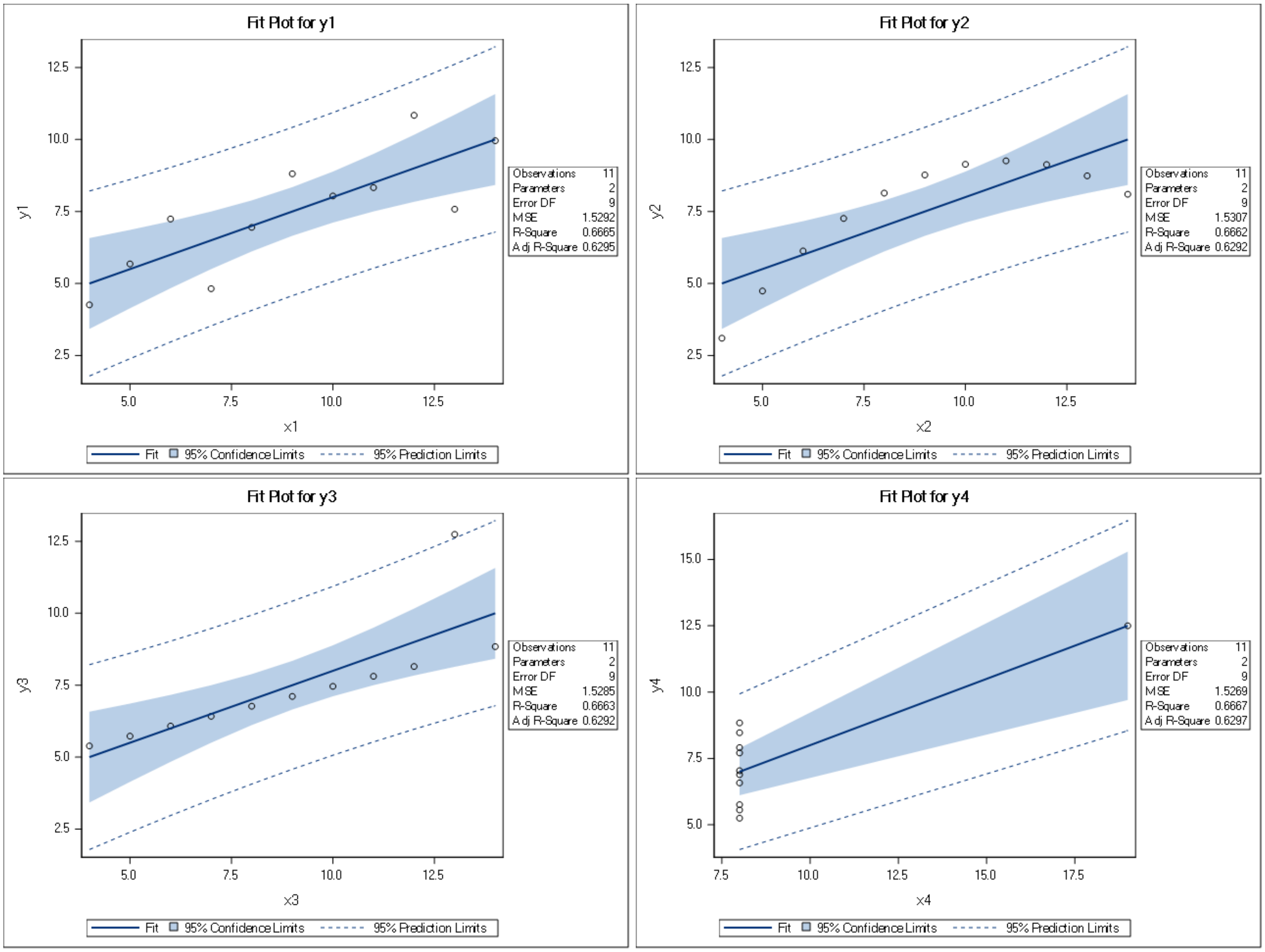

Figure 1: Plots of X vs Y, along with the estimated regression line, for Models 1-4.

Model 1 is the only appropriate model.

- Model 2 (upper right) shows a non-linear relationship between X and Y.
- Model 3 (lower left) shows an outlier point that is making the estimated slope larger than it would otherwise be.
- Model 4 (lower right) shows an influential point (far right) that completely dictates the estimated slope.