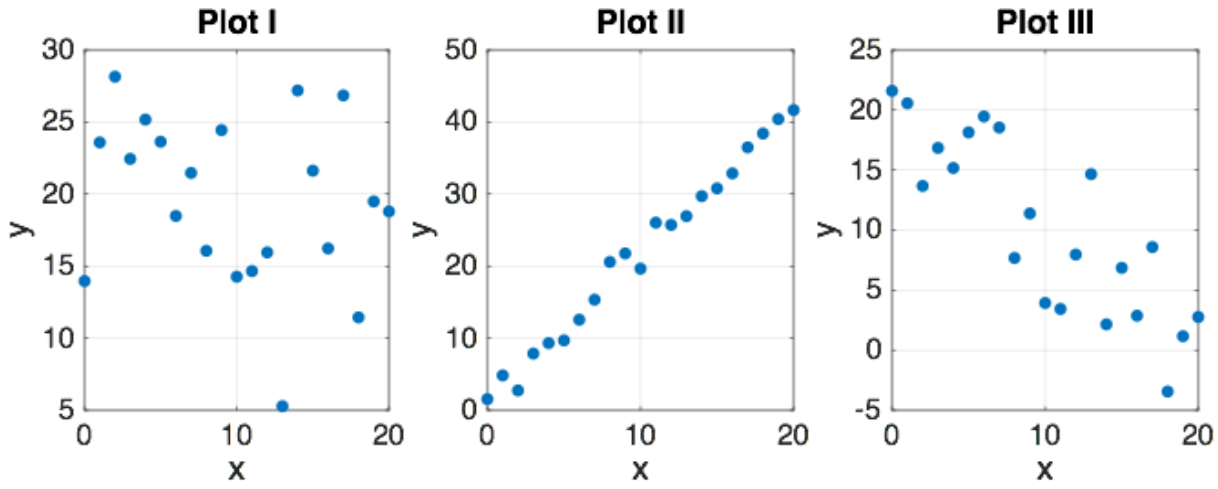
Cogs 109: Modeling and Data Analysis

Homework 2

Due Thursday 10/11 via gradescope



1. For each of the three data sets plotted above (I, II and III), answer the following:
   a. Does the data show a positive or negative correlation between x and y?
   b. Which function (equation) best describes each data set? In these equations, ε represents random noise, with mean value $\bar{\varepsilon} = 0$ and standard deviation $\sigma_\varepsilon = 1$.
      i. $f(x) = 1 + 2x + \varepsilon$
      ii. $f(x) = 20 + \varepsilon$
      iii. $f(x) = 20 - x + \varepsilon$
   c. Which regression table corresponds to each plot?

i.

	Estimate	SE	tStat	pValue
(Intercept)	1.2478	0.61327	2.0347	0.056077
x1	2.0417	0.052459	38.92	1.3891e-19

ii.

	Estimate	SE	tStat	pValue
(Intercept)	20.273	1.7491	11.591	4.6458e-10
x1	-1.0082	0.14962	-6.7383	1.9406e-06

iii.

	Estimate	SE	tStat	pValue
(Intercept)	21.808	2.4438	8.9236	3.1883e-08
x1	-0.2323	0.20905	-1.1112	0.28033

2. ISLR chapter 3, problem 3 (page 120)
3. ISLR chapter 3, problem 4 (pages 120-121)
4. In this problem, we will simulate a dataset and use multiple linear regression to investigate it. Imagine we conduct a survey of N=100 students and ask them how much time per week they spend on work (x_1) and how much time on play (x_2). We also ask them about their overall level of satisfaction (y), which we take to be the outcome. Download the dataset HW2.csv from the course website, which contains these data.
    a. Make a scatter plot showing y vs. x_1 . Comment on the relationship between these variables: do they appear correlated (positively or negatively)? Is their relationship linear or non-linear?
    b. Fit a simple linear regression of y vs. x_1 . In MATLAB, you could use the function `regress` or `fitlm`. Report the estimated intercept and slope, and make a plot showing the data points together with the regression line. Is there a statistically significant effect of x_1 on y ?
    c. What is the 95% confidence interval for the slope of x_1 ?
    d. Now fit a multiple linear regression with x_1 and x_2 as independent variables. Report a table with the regression results (similar to Table 3.9 on page 88 in ISLR). Which parameters have a statistically significant effect?
    e. Make a scatter plot showing y vs. $\hat{y}$, the predicted value of y .
    f. Create a categorical variable with 3 levels called WorkType, where WorkType="Idle" for $x_1 < 10$, WorkType="Diligent" for $10 \leq x_1 < 30$, and WorkType="Workaholic" for $x_1 \geq 30$. Fit a linear regression of y against WorkType and x_2 , and report the regression table.
    g. In part (f) you should have obtained two different coefficients for WorkType corresponding to different "levels" of this categorical variable. What is your interpretation of the term corresponding to WorkType=Workaholic?