Data 100, Fall 2020 **October 7, 2020**

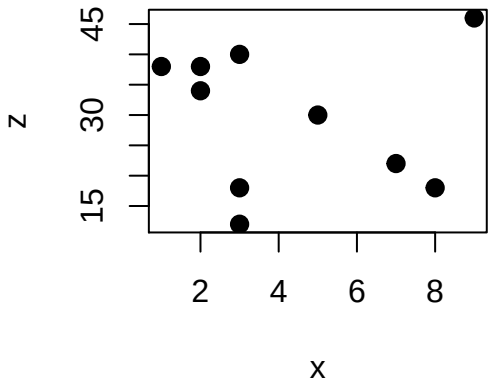
Discussion #6

Name:

Scatterplots and Correlation

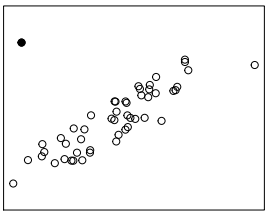
1. When we have more than two variables, it can be difficult to discern relationships from pairwise plots. Here is an example. Consider the 3 variables x, y, and z. We have 10 observations. Suppose we are interested in predicting z.

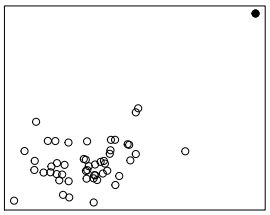
x	y	z
2	17	38
1	18	38
9	14	46
7	4	22
8	1	18
2	15	34
3	17	40
3	3	12
5	10	30
3	6	18



The correlation between x and z is -0.07. The scatter plot reflects this weak relationship. It appears that we should not bother to include x in a linear model for predicting z. Examine x, y and z carefully, and in the space below, sketch a scatter plot to show that there is a useful linear relationship that involves x.

2. Consider the two scatter plots below. For each scatter plot consider what happens to the correlation when the specially marked point is removed. Does the correlation get weaker, stronger, or stay about the same?





Linear Regression Fundamentals

3. In this problem, we will review some of the core concepts in linear regression.

   Suppose we create a linear model with parameters $\hat{\theta} = [\hat{\theta_0}, \ldots, \hat{\theta_p}]$. As we saw in lecture, given an observation $\vec{x}$, such a model makes predictions $\hat{y} = \hat{\theta} \cdot \vec{x} = \hat{\theta_0} + \hat{\theta_1} x_1 + \hat{\theta_2} x_2 + \ldots + \hat{\theta_p} x_p$.

   (a) Suppose $\hat{\theta} = [2, 0, 1]$ and we receive an observation $\vec{x_1} = [1, 2, 3]$. What $\hat{y_1}$ value will this model predict for the given observation?

   (b) Suppose the true y_1 was 3.5. What will be the L_2 loss for our prediction $\hat{y_1}$ from the previous part?

   (c) Suppose we receive another observation $\vec{x_2} = [1, 5, 1]$. What $\hat{y_2}$ value will this model predict for the given observation?

(d) Suppose the true y_2 was 4. What will be the mean squared error of our model, given the two observations?

Modeling

4. We wish to model exam grades for DS100 students. We collect various information about student habits, such as how many hours they studied, how many hours they slept before the exam, and how many lectures they attended and observe how well they did on the exam. Propose a model to predict exam grades and a loss function to measure the performance of your model on a single student.

5. Suppose we collected even more information about each student, such as their eye color, height, and favorite food. Do you think adding these variables as features would improve our model?