DS 100/200: Principles and Techniques of Data Science **Date: October 16, 2019**

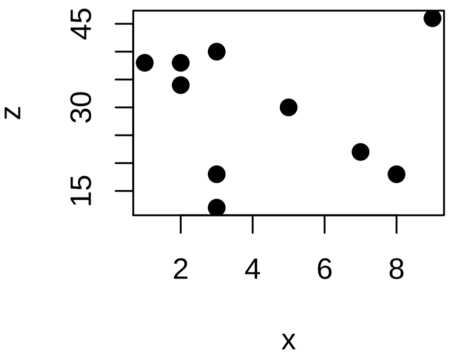
Discussion #8 Solutions

Name:

Regression Notions

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 above, sketch a scatter plot to show that there is a useful linear relationship that involves x.

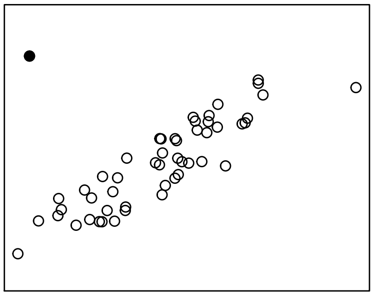
Solution: Although there is no relationship between x and z, if we know y then we can use x to perfectly predict z. That is,

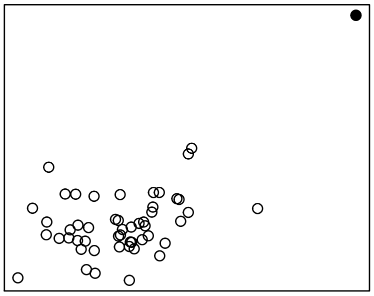
$$z = 2x + 2y$$

A scatter plot of $(2x + 2y, z)$ shows that all points fall on a line.

Scatter plots are limited in that they can't always reveal a linear relationship between three or more variables.

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?





Solution: When we drop the point in the left scatter plot, the correlation will get stronger. The correlation increases from 0.60 to 0.90, when the point is dropped.

When we drop the point in the right scatter plot, the correlation will get weaker. For these data, the correlation drops from 0.60 to 0.30.

3. The following are excerpts from https://prospect.org/features/roe-v.-wade-abort-crime/, which discusses a 2001 study by economists Donahue and Levitt.

> Looking at state-by-state and year-by-year figures, the two professors found a remarkable correlation between abortion rates and crime rates 15 to 18 years later.
>
> "According to our estimates," they boldly asserted, "legalized abortion is a primary explanation, accounting for at least one-half of the overall crime reduction... . The social benefit to reduced crime as a result of abortion may be on the order of $30 billion annually."

> "What's odd about our study," Levitt now reflects as he prepares for publication of the work and, presumably, renewed assaults on its authors, "is it manages to offend just about everybody. [But] our worldview is an economic worldview–that people respond to incentives. I view it as being apolitical."

Are their findings evidence that get-tough anti-crime policies have less effect on crime than most people think and that allowing women to choose when to have children has more?

Do their findings advocate for forced abortion against select elements of the American population?

Does this study argue that before Roe V Wade, more unwanted children were being born, often into difficult, non-nurturing, impoverished environments, and such children would be more likely than others to grow up to commit crimes as troubled, angry, gang-affiliated teenagers and young adults?

Solution: All three of these statements attach causal implications to the observed relationship.

The first statement attempts to attribute the decrease in crime rates to the increase in abortion rates rather than tougher crime laws.

The second statement is a hysterical view point.

The third statement is a pro-choice view about the positive impact of a women's decision.

The passing of Roe v Wade was viewed as a natural experiment. Are you aware of other things going on at the time, which may complicate the findings? The crack epidemic was in the 1980s and early 1990s.

Ibser, in his 2002 thesis, studied the data from Donoho and Levitt. He found that New York state's data point looked like the dark circle in the above right plot. How might this impact the findings?

Solution: Note that the relationship the Donho and Levitt found had a strong negative correlation between abortion rate and crime.

Ibser found that New York state was driving the relationship between crime rate and abortion rate. NY was a state that allowed abortion well before RvW and so had high abortion rate, and it also had very low crime rates. With out NY, the relationship was much weaker.

The Bootstrap

4. We can use the bootstrap to carry out inference on the slope of a simple linear regression. Below is a simple linear regression model

$$\alpha + \theta x$$

where (x, y) are observed continuous values, y is the response, and x is the explanatory variable, aka the feature. We can use the data to estimate the intercept and the slope, we arrive at the following equation:

$$\hat{y}_i = \hat{\alpha} + \hat{\theta} x_i$$

Suppose we want to test the hypothesis that $\theta = 9$. Consider the following diagram of the bootstrap process to test this hypothesis. Fill in the 9 blanks below the diagram using the phrases below:

(A) Population

(B) Bootstrap population

(C) Observed sample

(D) Expected sample

(E) Bootstrap sample

(F) Sampling distribution

(G) Sampling

(H) Bootstrapping

(I) Bootstrap sampling distribution

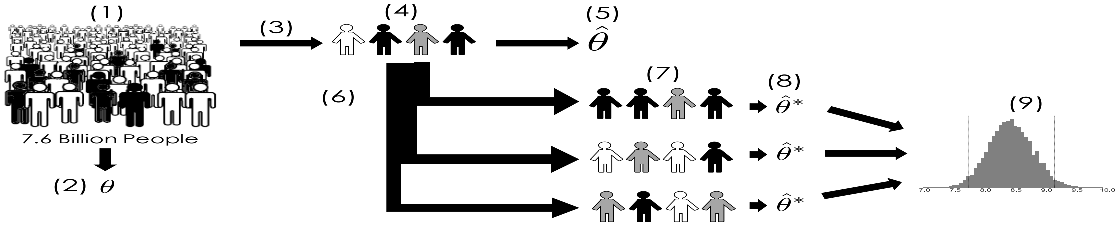
(J) Empirical distribution

(K) True distribution

(L) Population parameter

(M) Sample Statistic

(N) Bootstrap Statistic



1. ______(A)______

2. ______(L)______

3. ______(G)______

4. ______(C)______

5. ______(M)______

6. ______(H)______

7. ______(E)______

8. ______(N)______

9. ______(I)______

5. Describe how you would test the hypothesis that the population value for θ is 9 at the 95%-level. Fill in the blanks below:

   1. Null Hypothesis: _______________

   2. Alternative Hypothesis: _______________

   3. We _______________ the null hypothesis.

Solution: Null Hypothesis: $\theta = 9$
Alternative Hypothesis: $\theta \neq 9$

We can use the variability in the bootstrap estimates, $\hat{\theta}^*$ to construct a 95% confidence interval for θ, the slope of the true line. That is we use the bootstrap percentile method to create a confidence interval that extends from the 2.5th percentile to the 97.5th percentile of the bootstrapped slopes.

We see that the value 9 is within the lower and upper endpoints of the 95% confidence interval so we do not reject the null hypothesis.

Explain the reasoning behind your conclusion.

Solution: We need to know more about the sampling process. We have assumed that the sampling is done at random and is representative of the population. If this probability model is true and if the population $\theta = 9$, then we would expect the sample estimate of θ to have a sampling distribution like the bootstrap histogram above.