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

Discussion #8

Name: Raguvir Kunani

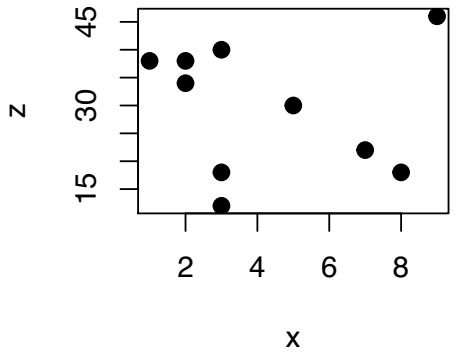
Regression Notions

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

Even though x and z are not correlated, there still is a linear relationship involving x and z: $2x + 2y = 2z$

Takeaway: Just because 2 variables have a low correlation does not mean there is no linear relationship involving those 2 variables!

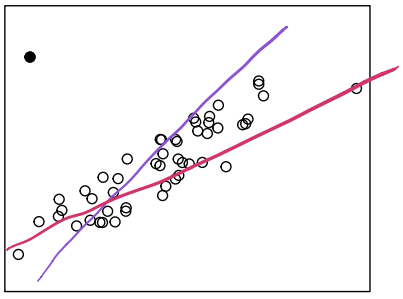


The correlation between x and z is -0.07. The scatter plot reflects this weak relationship. It appears that we should not bother to include x in a linear model for predicting z. Examine x, y and z carefully, and in the space above, sketch a scatter plot to show that there is a useful linear relationship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

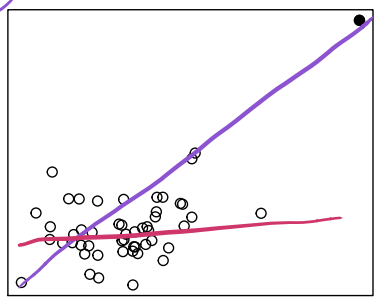
stronger correlation ⇒ |r| increases

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?

(A) stronger



(B) weaker



Takeaway: Correlation very sensitive to outliers

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?

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?

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

$$\alpha + \theta x$$

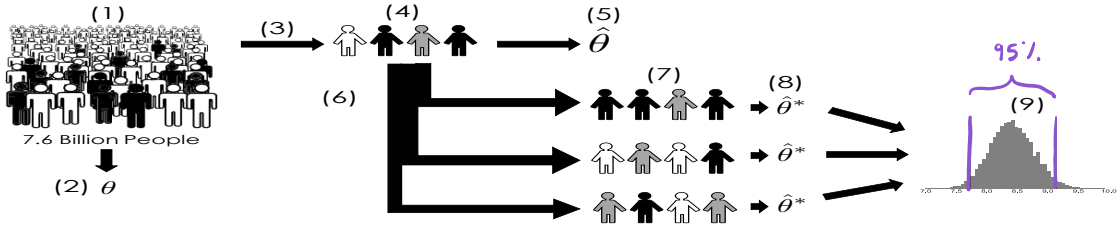
where (x, y) are observed continuous values, y is the response, and x is the explanatory variable, aka the feature. We can use the data to estimate the intercept and the slope, we arrive at the following equation:

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

"^" means estimate

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. Population
2. Population parameter
3. Sampling
4. observed & bootstrap Sample population
5. Sample statistic
6. Bootstrapping
7. Bootstrap sample
8. Bootstrap statistic
9. Bootstrap sampling distribution

We use a statistic to estimate a parameter

Parameters are often treated as truth.

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: $\theta = 9$ can happen by chance

   2. Alternative Hypothesis: $\theta = 9$ does not happen by chance

   3. We fail to reject the null hypothesis.

   Explain the reasoning behind your conclusion.

The 95% confidence interval of the bootstrap sampling distribution contains $\theta = 9$, so we can say that it's possible for $\theta = 9$ to have occurred by chance.

Note: When we say in the null hypothesis that $\theta = 9$ could have occurred by chance, there is an implicit set of assumptions we are making about the world (this is called the data generation model). The null hypothesis is saying that, given our assumptions, it is not improbable that $\theta = 9$ occurs. Also, we never accept the null hypothesis because we can never be 100% confident.