

# Scatterplots, Correlation, & Linear Regression

Recitation Handout

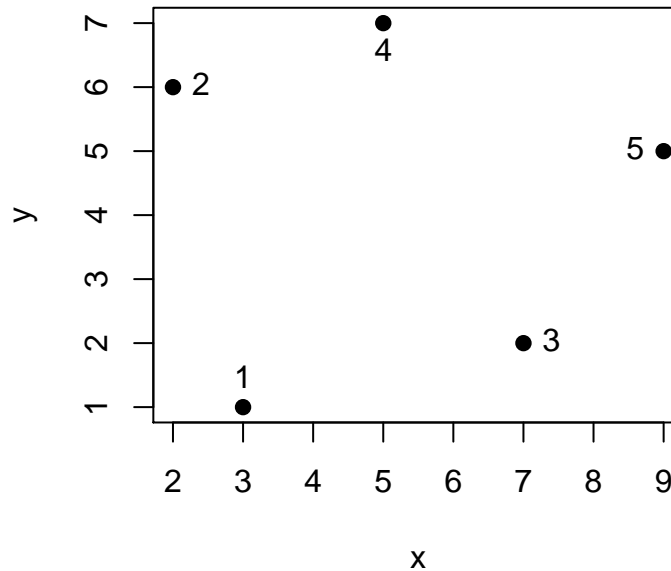
TF: Michael Gill

## 1 Scatterplot

A scatterplot is a way to visually display the relationship between two \_\_\_\_\_-level variables.

Quickly fill in the data table on the left using the scatter plot below:

Obs.	x	y
1		
2		
3		
4		
5		



### 1.1 Scatterplots in SPSS

To do a scatterplot in SPSS, you simply need to use the Graphs → Chart Builder command. Pick the Simple Scatter and the just drag your independent variable to the x-axis and the dependent variable to the y-axis. Then hit OK. Once you see the scatterplot in your output, you can edit it to add lines or change colors by double-clicking on it.

### 1.2 Examples

- 1) You work get a job at a fancy newspaper and they ask you to write a story about the recent political unrest with regard to organized labor. To support your argument about the relationship between organized labor and Democrats, they ask you to include a scatterplot showing the relationship between the percent of workers who are union members in a state (union) and the percent of the state legislators who are Democrats (demstate). What direction is the relationship? Does it appear linear?

## 2 Correlation

Pearson's correlation coefficient is a measure of the \_\_\_\_\_ and \_\_\_\_\_ of the relationship between two interval-level variables. If the two variables are  $X$  and  $Y$  with averages  $\bar{X}$  and  $\bar{Y}$ , and estimated standard deviations  $s_x$  and  $s_y$ , then the correlation coefficient is:

$$r = \frac{\sum_i \left( \frac{X_i - \bar{X}}{s_x} \right) \left( \frac{Y_i - \bar{Y}}{s_y} \right)}{n - 1} \quad (1)$$

which we could write as:

$$r = \frac{\sum_i ( \quad ) ( \quad )}{n - 1} \quad (2)$$

That is, we want to know how often observations fall in same place in their distribution for both variables. For instance, if wealth and education have a high positive correlation this means that people who are on the upper end of the wealth distribution will be on the upper end of the education distribution.

If a country's GDP per capita has a high negative correlation with its levels of political violence (in, say, number of political violence deaths per 1,000 citizens), then those countries on the low end of the GDP distribution will be on the \_\_\_\_\_ end of the political violence distribution.

### 2.1 Example

Let's say you wanted to calculate the Pearson correlation coefficient for the following two variables:

x	y
1	17
2	14
3	16
4	10
5	8

Note that  $\bar{X} = 3.0$ ,  $\bar{Y} = 13.0$ ,  $s_x = 1.6$ , and  $s_y = 3.9$ .

### 2.2 Correlation in SPSS

To get a correlation matrix in SPSS, simply use the Analyze → Correlate → Bivariate command.

- 2) After your article hits the front page, you get a letter from a large number of readers that want more detail than just the scatterplot and what they call your "impressions." Calculate the correlation between union and

demstate to bolster your claims. Is the relationship significant?

### 3 Linear Regression

Linear regression is a way to see how our dependent variables changes as our independent variable changes. We are going to assume that we can write the relationship down as a linear function:

$$Y = a + bX \quad (3)$$

which you might remember from algebra looks suspiciously like  $y = mx + b$ . All we are doing here is trying to estimate a line that summarizes the relationship between variables  $Y$  and  $X$ . In this class, the properties of the line will be parameters in the population that we want to estimate.

$$Y = \alpha + \beta * X \quad (4)$$

$$\alpha = \quad (5)$$

$$\beta = \quad (6)$$

POPQUIZ!!!! Could we ever know  $\alpha$  and  $\beta$ ? Why or why not?

#### 3.1 Linear Regression in SPSS

To run a linear regression in SPSS, simply use the Analyze → Regression → Linear. Input your independent and dependent variables and hit OK.

- 3) The last letter you get is an angry letter from your Gov50 TF asking you why you didn't run a linear regression to explain Democratic performance in the state legislature as a function of union membership. What would be the independent and dependent variables for this regression? What are the estimated coefficient and intercept?