Chapter 8. Correlation

- How to read (and draw) scatter diagrams. (A1,A4)
- The spread in a scatter diagram shows the weakness of the relationship between two variables.
- If there is a strong association between two variables, then knowing one helps a lot in predicting the other. But when there is a weak association, information about one variable does not help much in guessing the other.
- Graphically interpretation of the correlation coefficient: measuring clustering around a line. ($r$ goes up to 1 [or -1] as clustering increases to a line that slopes up [or down])
- The point of averages (ave. of $x$, ave. of $y$) picks out the center of the scatter diagram.
- Scatter diagrams are summarized by
  - the average of the $x$-values, the SD of the $x$-values,
  - the average of the $y$-values, the SD of the $y$-values,
  - the correlation coefficient $r$.
- Correlations are always between $-1$ and 1. A positive correlation means that the cloud slopes up; as one variable increases, so does the other. A negative correlation means that the cloud slopes down; as one variable increases, the other decreases. (R2,R5,R11)
- The SD line indicates the drift of the scatter diagram. This line goes through the point of averages, and its slope is $(\text{SD of } y) / (\text{SD of } x)$; the sign is the same as that of $r$. (If $r$ is 0, either sign can be used.)
- Algorithm for computing the correlation coefficient:
  Convert each variable to standard units. The average of the products gives the correlation coefficient. (R9(a))

Chapter 9. More about Correlation

- $r$ is a pure number, invariant under scaling, symmetric in $x$ and $y$.
- Since $r$ is invariant under change of scale, “clustering” must be interpreted relative to the SDs. (figure 3)
- $r$ measures linear association, not association in general. So $r$ may not be useful if there is a strong nonlinear association, or outliers.
- Ecological correlations are based on rates or averages. They are often used in political science and sociology. And they tend to overstate the strength of an association. The idea that individual behavior can be inferred from group behavior is called the ecological fallacy. (figure 6) (R11)
- Correlation measures association. But association is not the same as causation. (example 1 on p.150)
- Attenuation: the reduction of $r$ due to restriction of range. (A9,B2)

Chapter 10. Regression

- Estimating the average of $y$ from $x$: if $x$ goes up by one SD, on the average, $y$ does not go up by a whole SD, but only by part of an SD, namely, $r \times \text{SD of } y$.
- The regression line for $y$ on $x$ estimates the average value for $y$ corresponding to each value of $x$.
- A graph which displays the average of $y$ against $x$ is called a graph of averages.
- The regression line is a smoothed version of the graph of averages. If the graph of averages follows a straight line, that line is the regression line. Regression line should not be used when there is a non-linear association between the variables. (Figure 4) (B1)
- Regression estimates for individuals. (R2,9)
- When $x$ goes up by one SD, most people want $y$ to go up by a full SD too. The fact that it doesn’t is the regression effect. It is due to the spread of the scatter diagram around the SD line. People resist this statistical explanation, and want some real cause for the regression effect: that is the regression fallacy. (D1,D2)
- There are two regression lines, one for $y$ on $x$, another for $x$ on $y$. (figure 8)