

Correlation and Regression Exercise

1. Statistician Frank Anscombe created four pairs of x and y variables to illustrate the importance of plotting your data first. The data are formatted for Stata in *anscombe.dta*. The first three pairs are formed by matching $x1-3$ with each of the first three y variables. The fourth pair is formed by $(x4, y4)$.

Data set	1-3	1	2	3	4	4
Variable	x	y	y	y	x	y
Obs. no. 1 :	10.0	8.04	9.14	7.46	8.0	6.58
2 :	8.0	6.95	8.14	6.77	8.0	5.76
3 :	13.0	7.58	8.74	12.74	8.0	7.71
4 :	9.0	8.81	8.77	7.11	8.0	8.84
5 :	11.0	8.33	9.26	7.81	8.0	8.47
6 :	14.0	9.96	8.10	8.84	8.0	7.04
7 :	6.0	7.24	6.13	6.08	8.0	5.25
8 :	4.0	4.26	3.10	5.39	19.0	12.50
9 :	12.0	10.84	9.13	8.15	8.0	5.56
10 :	7.0	4.82	7.26	6.42	8.0	7.91
11 :	5.0	5.68	4.74	5.73	8.0	6.89

TABLE. Four data sets, each comprising 11 (x, y) pairs.

Source: Anscombe, F. J. 1973. Graphs in statistical analysis. *American Statistician* 27:17-21.

- a. Without making scatterplots, find the correlation and the least-squares regression line for all four pairs [i.e., $(x1_3, y1)$, $(x1_3, y2)$, $(x1_3, y3)$, $(x4, y4)$]. What do you notice? Use the regression line to predict y for $x = 10$.
- b. Make a scatterplot for each of the data sets and add the regression line to each plot.
- c. In which of the four cases would you be willing to use the regression line to describe the relationship between y and x ? Explain your answer in each case.

2. It is important for archaeological research and for cultural resources management to predict how well human skeletal material will be preserved at mortuary sites. Claire Gordon and Jane Buikstra analyzed the relationship between bone preservation and soil acidity, as measured by pH, in the Lower Illinois River Valley. They used regression equations to predict how much skeletal material should be recoverable from mortuary sites. Gordon and Buikstra argue that researchers can use such predictions to choose the appropriate excavation strategy and to estimate how much bias in the samples can be attributed to imperfect preservation.

- a. Gordon and Buikstra analyzed burials for 63 adults and 32 children. Soil pH was associated with bone preservation in both adults ($r = -0.92$) and children under 15 years of age ($r = -0.48$). Note that the variable preservation is scored such that higher values indicate great destruction of bone. Describe the strength and direction of the relationship for each subsample. How much variation in mature

and subadult bone preservation is accounted for by soil pH? Why do you think the strength of the association varies between children and adults?

- b. Here are the results Gordon and Buikstra report for the regression of bone preservation on soil pH.

Adult's Simple Regression

$$\text{PRESERVATION} = -1.3\text{pH} + 12.5$$

Slope $p < .00001$, Intercept $p < .00001$

Children's Simple Regression

$$\text{PRESERVATION} = -1.5 \text{ pH} + 14.9$$

Slope $p < .005$, Intercept $p < .002$

Source: Gordon, C. C., and J. E. Buikstra. 1981. Soil pH, bone preservation, and sampling bias at mortuary sites. *American Antiquity* 46:566-571.

How would you interpret the results of each regression equation? What do the slope and intercept of the regression equations tell us about the relative rate of bone destruction for children versus adults as pH decreases?

3. How well does the number of beers a student drinks predict his or her blood alcohol content (BAC)? Researchers at Ohio State University conducted an experiment in a student dormitory to answer this question. Sixteen volunteers were randomly assigned to drink between one and nine 12-ounce beers. Thirty minutes after drinking the last beer, a police officer measured their BAC by breathalyzer. Students were equally divided between men and women and differed in weight and usual drinking habit. Because of this variation, many students don't believe that number of drinks predicts blood alcohol well. What do the data say (*bloodalc.dta*)?

- Make a scatterplot of blood alcohol content against number of beers consumed. Describe the form, direction, and strength of the relationship.
- What is the correlation (r) between BAC and beers consumed? What does it mean?
- Calculate the regression of BAC on number of beers consumed, and create a new scatterplot that shows the least-squares regression line. How much variation in BAC is explained by the number of beers consumed?
- What is the regression equation? What does the coefficient for beers mean? How would you interpret the 95% confidence interval for this coefficient?
- On average, what BAC would you predict for someone who drinks five beers?
- Given the data available, how might you modify the analysis to explain additional variation in BAC?

4. What accounts for cross-cultural variation in total fertility rates? We can take a first stab at this problem using data on national health indicators and sociodemographic variables for 192 member countries in the World Health Organization (see *who2005.dta*).
- Make a scatterplot of total fertility rate (TFR) against percent of the population living below the poverty line ($< \$1$ per day). Describe the form, direction, and strength of the relationship, and summarize with Pearson's r .
 - Find the least-squares regression line for TFR and percent of the population below the poverty line. How much variation in TFR does percent poverty explain?
 - On average, for each 1% increase in the percentage of people living in poverty, how does TFR change?
 - One assumption of the least-squares model is that the scatter of points around the regression line should be roughly the same over the entire range of the data. We can check this assumption by plotting the residuals against the predictor, x . If the assumption holds, we should see more or less random scatter. If the residuals show any pattern in relation to x , then our conclusions may be biased. Use `rvpplot` (residuals versus predictor) in Stata to make a scatterplot of the residuals against the percentage of people living in poverty. Does the plot provide evidence of a problem in the model?
 - Propose an explanation for why TFR is associated with percent of people living below the poverty line.
5. Some people use mean SAT scores to rank state or local school systems. This approach is flawed, because the percent of high school students who take the SAT varies from place to place. Let's examine the relationship between the percent of a state's high school graduates who took the exam in 2002 and the state average SAT verbal score that year. The data are in *states.dta*.
- Make a scatter plot of the mean verbal SAT score against the percent of high school students to take the exam. Describe the form, direction, and strength of the relationship.
 - Find the regression of verbal SAT score on percent taking the exam. Interpret the coefficient.
 - Make a residual-versus-predictor plot. Does the plot provide any evidence of a problem in the model?
6. One way to study the brain's response to sounds is to compare responses to "pure tones" with responses to recognizable sounds. To compare responses, researchers anesthetized macaque monkeys. They fed pure tones and also monkey calls directly to their brains by inserting electrodes. Response to the stimulus was measured by the ring rate (electrical spikes per second) of neurons in various areas of the brain. The data are in *monkeysounds.dta*.

- a. One notable finding is that responses to monkey calls are generally stronger than are responses to pure tones. Give a numerical measure or measures that supports this finding.
- b. Make a scatterplot of the brain's response to monkey call against response to pure tone. Find the least-squares line and add it to your plot. Describe the scatterplot. Which point has the largest residual? Which point is an outlier in the x direction?
- c. How influential are each of these points for the correlation r ?
- d. How influential are each of these points for the regression line?