

1. I want to ask you to be a scientist and try to make sense of data that link parents and offspring. Consider one plot (of your own choosing) from figures 1-4 depicting heights of 63 undergraduate college students and their parents. (I collected these data in the USA in the mid to late 1990s.)

What patterns can you discern?

What ideas or questions do you have about the causes producing those patterns?

What questions or reservations do you have about the process you go through in answering these questions?

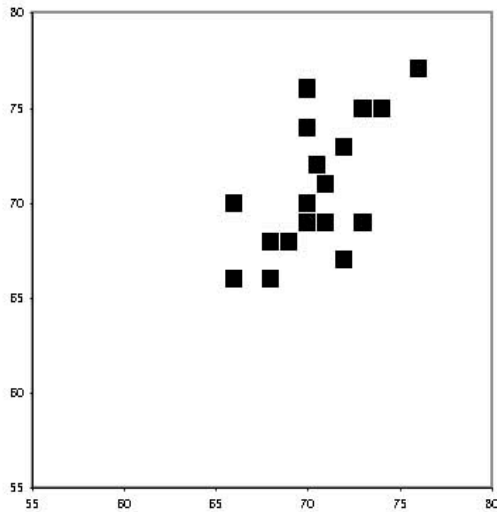


Figure 1:  
Son's vs. father's height (inches)

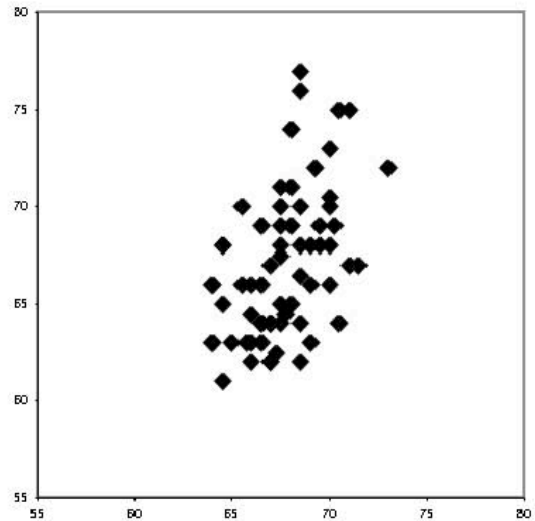


Figure 3:  
Student's vs. average of parents' height (inches)

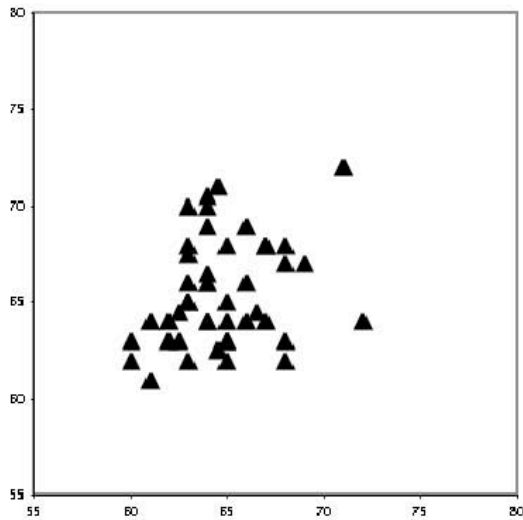


Figure 2:  
Daughter's vs. mother's height (inches)

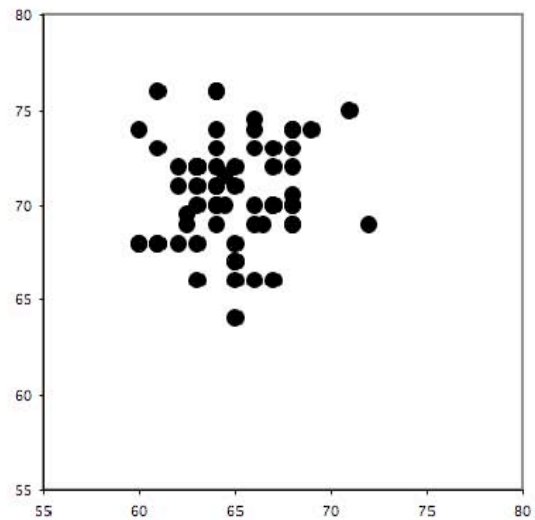
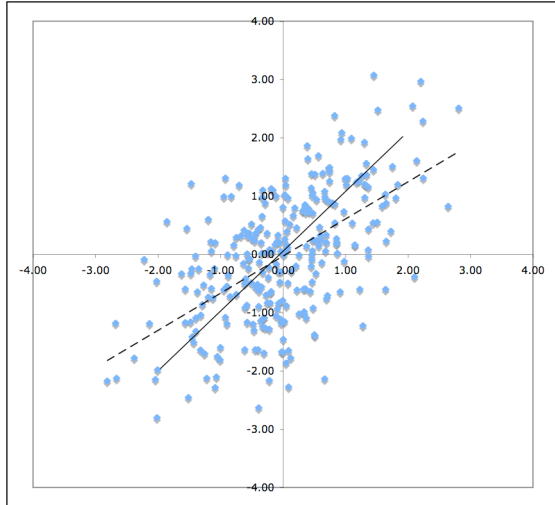


Figure 4:  
Father's vs. mother's height (inches)

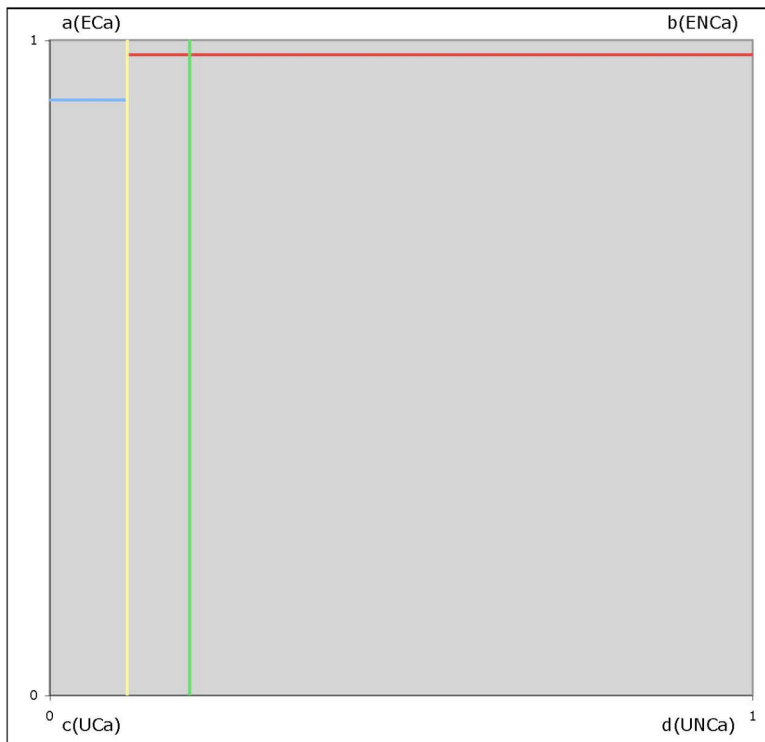


Association of continuous variables.

Scatter plot of set of two standardized variables with correlation 0.5. The solid line is  $y = x$ , which is the best fit by eye, and the dashed line is  $y = .5x$ , that is, the linear regression line. The average mean squared vertical distance from the regression line is the minimum of all possible lines (i.e., least squares solution or best predictor or  $y$  for any  $x$ ). The average perpendicular distance from the solid line is  $1 - \text{correlation}$ . That is, the larger the correlation, the tighter the packing of the points. This remains the case even if  $y$  and  $x$  are switched. It is also the case that  $1 + \text{correlation} = \text{size of } 1^{\text{st}}$  principal component and  $1 - \text{correlation} = \text{size of } 2^{\text{nd}}$  principal component (or vice versa if correlation is negative). This can be generalized to associations among multiple variables.

---

Association of categorical variables



a Exposed cases	0.010
b Exposed non-cases	0.020
c Unexposed cases	0.100
d Unexposed non-cases	0.870
Controls as fraction of non-	
f cases	0.100
fb Exposed controls	0.002
fd Unexposed controls	0.087
Relative Risk (full popn.)	3.233
Odds Ratio (full popn.)	4.350
Odds Ratio (case-control)	4.350
"Rel. risk" (case-control)	1.558