Stat 13 lecture 23
correlation and regression

- A cartoon from handout
- The taller the father, the taller the son
- Tall father's son is taller than short farther's son
- But tall father's son is not as tall as father; short father's son is not as short as father
- Galton's data

4. The set of points in the plane for which $f_{X,Y}(x, y) = c$ is an ellipse for each constant $c > 0$. Furthermore, these ellipses are concentric.[1] [See Figure 10.3.4, reproduced from (122). Also, see (189) for a fuller discussion of the smoothing procedure.]

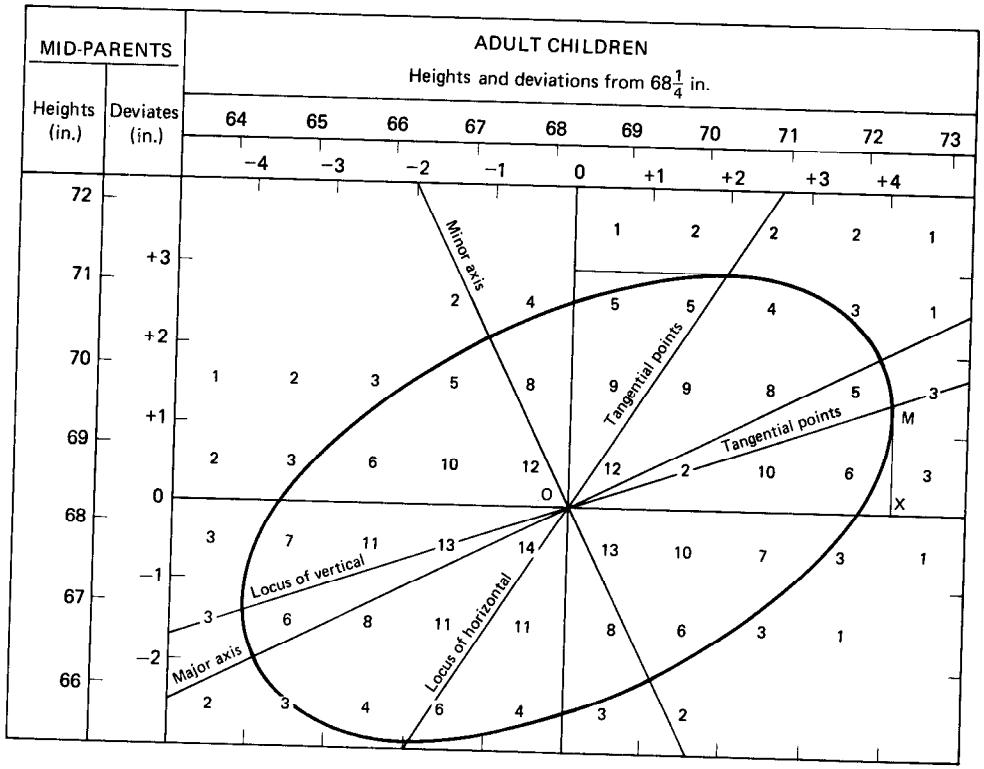


Figure 10.3.4 Galton's contour lines.

Galton (51) gave the following recollection of the birth of property 4:

> At length, one morning, while waiting at a roadside station near Ramsgate for a train, and poring over the diagram in my notebook, it struck me that the lines of equal frequency ran in concentric ellipses. The cases were too few for my certainty, but my eye, being accustomed to such things, satisfied me that I was approaching the solution. More careful drawings strongly corroborated the first impression.

[1] Of this discovery, Pearson said, "That Galton should have evolved all this from his observations is to my mind one of the most noteworthy scientific discoveries arising from pure analysis of observations" (122).

(d) unknown

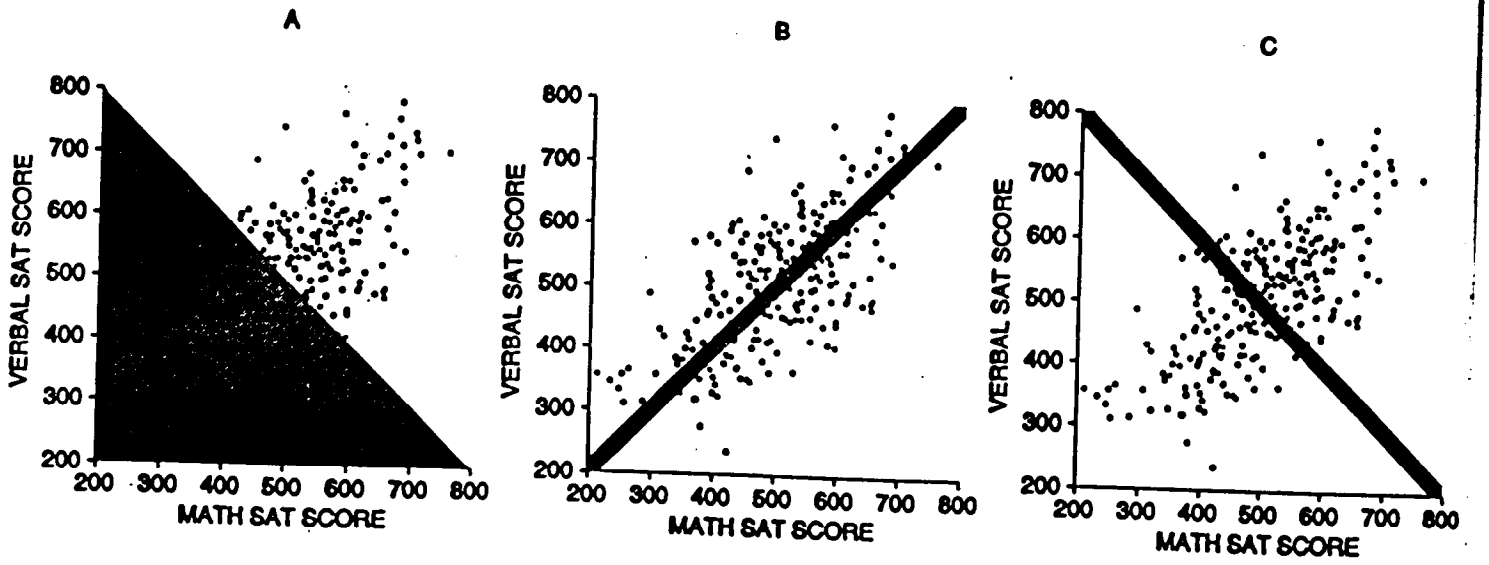
10. True or false: A student who is at the 40th percentile of first-year GPAs is also likely to be at the 40th percentile of second-year GPAs. Explain briefly. (The scatter diagram is football-shaped.)

7. SUMMARY

1. Associated with an increase of one SD in x, there is an increase of only r SDs in y, on the average. Plotting these *regression estimates* gives the *regression line* for y on x.

2. The *graph of averages* is often close to a straight line, but may be a little bumpy. The regression line smooths out the bumps. If the graph of averages is a straight line, then it coincides with the regression line. If the graph of averages has a strong non-linear pattern, regression may be inappropriate.

3. The regression line can be used to make predictions for individuals. But if you have to extrapolate far from the data, or to a different group of subjects, be careful.



2. In a study of the stability of IQ scores, a large group of individuals is tested once at age 18 and again at age 35. The following results are obtained.

   age 18: average score $\approx$ 100, SD $\approx$ 15
   age 35: average score $\approx$ 100, SD $\approx$ 15, $r \approx 0.80$

   (a) Estimate the average score at age 35 for all the individuals who scored 115 at age 18.
   (b) Predict the score at age 35 for an individual who scored 115 at age 18.

3. Pearson and Lee obtained the following results in a study of about 1,000 families:

   average height of husband $\approx$ 68 inches, SD $\approx$ 2.7 inches
   average height of wife $\approx$ 63 inches, SD $\approx$ 2.5 inches, $r \approx 0.25$

   Predict the height of a wife when the height of her husband is
   (a) 72 inches (b) 64 inches (c) 68 inches (d) unknown

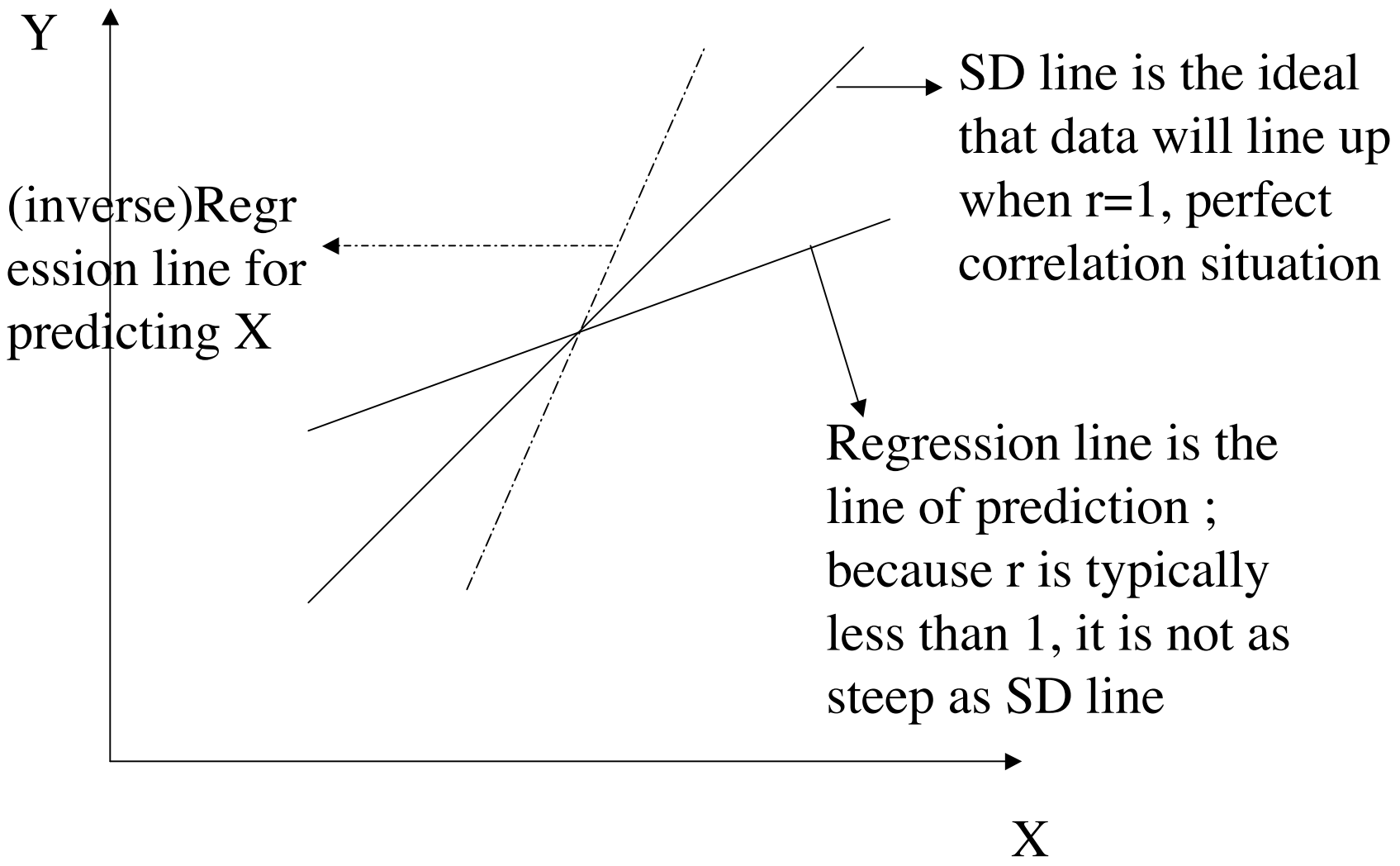
4. In one study, the correlation between the educational level of husbands and wives in a certain town was about 0.50; both averaged 12 years of schooling completed, with an SD of 3 years.[7]

Regression line

- The formula :
- $y=\mu_y+r[SD(Y)/SD(X)](x-\mu_x)$ where μ_y is mean of Y and μ_x is mean of X
- Application : (a) predict IQ at age 35 for someone with IQ of 110 at age 18.
- (b) predict IQ at age 35 for someone with IQ 90 of at age 18

Answer

- (a) y= 100 + .8 (15/15) (110-100) =108,
- Observe that this is greater than average but less than x=110
- (b) y=100+ .8 (15/15) (90-100)=92,
- This is smaller than average but is greater than x=90.
- This is consistent with the cartoon.



Slope of SD line = SD(Y)/SD(X)

Slope of regression line = r [SD(Y)/SD(X)]