SECTION 3

Chapter 8, 9, 10, 11 and 12

CHAPTER 8 & 9

Correlation

- Scatter Plot
- Correlation Coefficient
- SD Line
- Correlation Coefficient Computation

Scatter Plot

• A scatterplot or scatter diagram is a two-dimensional plot of data. The horizontal dimension is called x, and the vertical dimension is called y.

• Each point on a scatterplot or scatter diagram shows two values, an x value and a y value. Each point represents a single case. A single case could be a single person or object, but a single case could be a matched pair (e.g. father-son, twins, husband-wife)

• Scatter diagrams only show association, but association does not mean causation

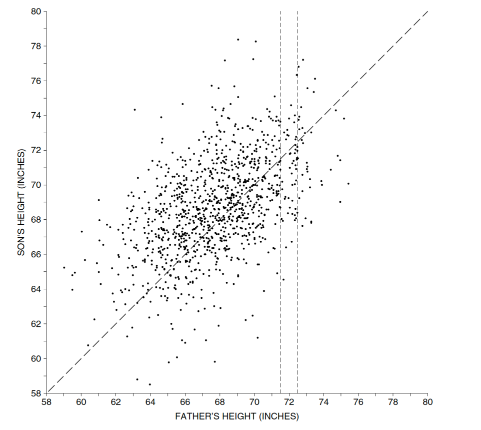
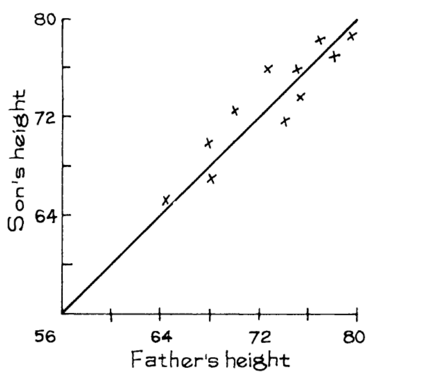


Figure 3. Son's height close to father's height.



A ***positive association***
between the heights of fathers and sons

The swarm of points slopes upward to the right, the y-coordinates of the points tending to increase with their x-coordinates.

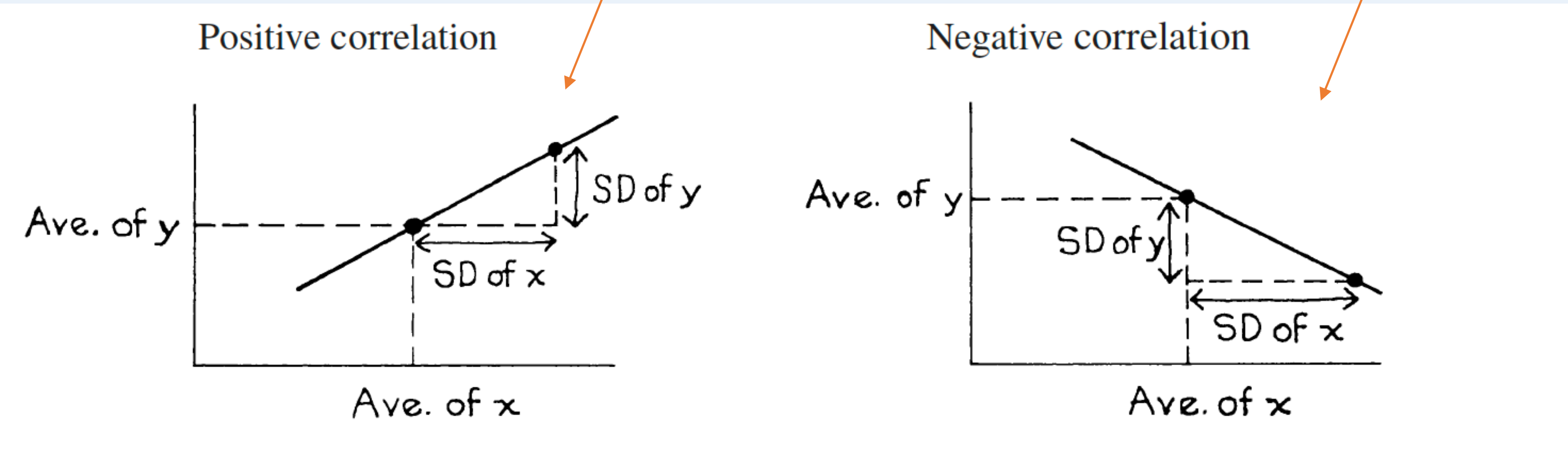
SD Line

The points in a scatter diagram generally seem to cluster around the SD line.

- Goes through the point of averages   Avg of y = Intercept + Slope × Avg of x
- Goes through all the points which are an equal number of SDs away from the average, for both variables

Slope: **(SD of y)/(SD of x)**

Slope: **-(SD of y)/(SD of x)**



Correlation Coefficient Computation

Method I:

$$r = \text{average of (x in standard units)} \times \text{(y in standard units)} \quad \text{where standard units} = \frac{\text{value} - \text{average}}{SD}$$

Method II:

$$r = \frac{cov(x, y)}{(SD\ of\ x)\times(SD\ of\ y)} \quad \text{where } \mathrm{cov}(x, y) = \text{average of products } xy - (\text{average of } x) \times (\text{average of } y)$$

Table 1. Data.

x	y
1	5
3	9
4	7
5	1
7	13

Method I:

Step 1. Convert the x-values to standard units

Step 2. Convert the y-values to standard units

Step 3. Work out the product for each (x, y) pair (x in standard units) ×(y in standard units)

Step 4. Take the average of the products

Method II:

Step 1. Calculate the average of products xy, avg of x, avg of y

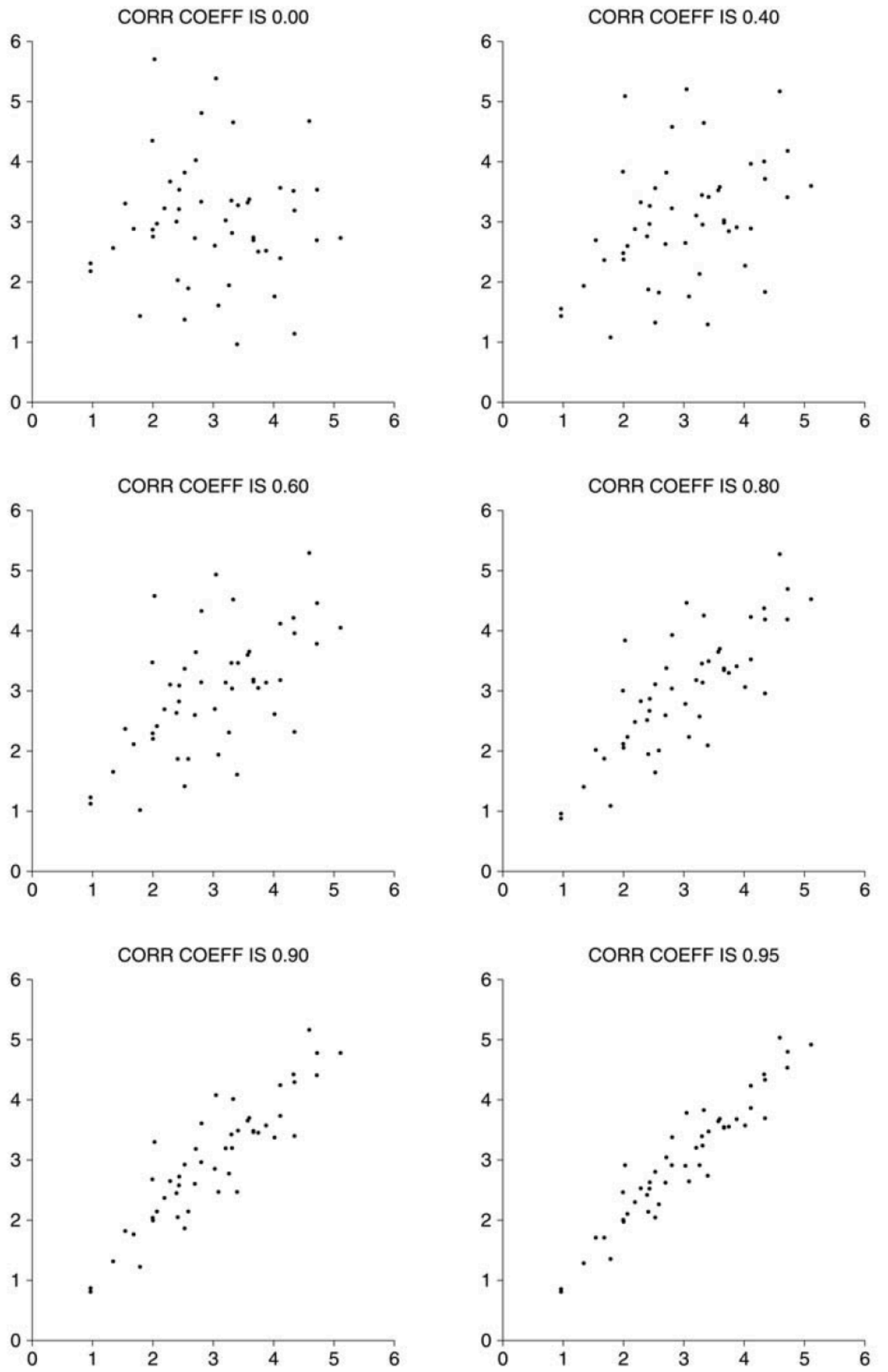
Step 2. Calculate covariance cov(x, y)

Step 3. Calculate SD of x, SD of y

Step 4. Divide covariance by the product of SD of x and SD of y

Mean_x = 4, Mean_y = ; SD_x = 2, SD_y = 4

- **-1 <= r <= 1**
- The correlation r measures how close the data are to a line
- If r is close to 1 or -1, the data are close to a line
- If r is close to 0, the data are not close to a line
- r does **NOT** tell what percentage of the data fall on the line
- r = 0.80 does not indicate twice as much linearity as r = 0.40
- The correlation between x and y is the same as the correlation between y and x. r(x, y)=r(y, x)
- Invariant under addition: If some constant "a" is added to every one of the X or the Y values, the correlation is unchanged
- Invariant under multiplication: if all the x or the y values are multiplied by some positive constant "b", the correlation is unchanged. The correlation can change very dramatically if only ONE of the data points is changed



CHAPTER 10, 11, 12

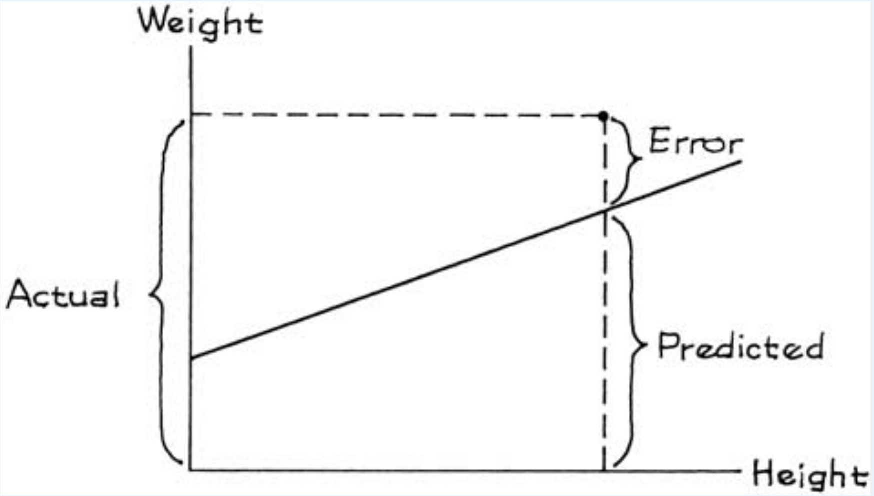
Regression

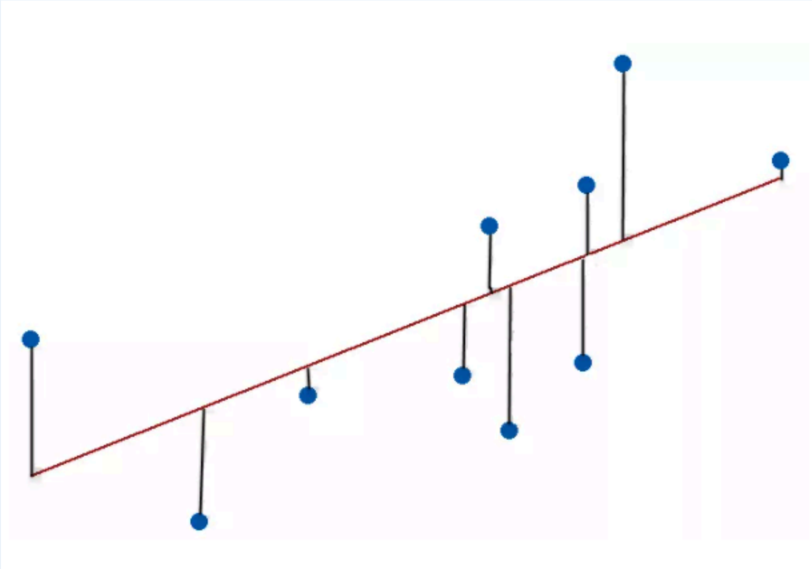
- R.M.S Error
- Regression Line Equation
- Least Squares Method
- Residuals Plot
- Normal Curve Inside a Vertical Strip

R.M.S Error (Root Mean Square error)

The r.m.s. error for regression says how far typical points are above or below the regression line.

r.m.s. error = $\sqrt{\dfrac{(error\ \#1)^2 + (error\ \#2)^2 + \cdots + (error\ \#n)^2}{n}} = \sqrt{1 - r^2} \times SD_{\ predictor}$





Slope and Intercept

Slope: the rate at which y increases with x, along the line. slope = rise/run

Intercept: the height when x = 0.

<u>Slope Calculation Method:</u>
Find two points in the line, A and B. Find the rise and run, then do the calculation

Exercises:

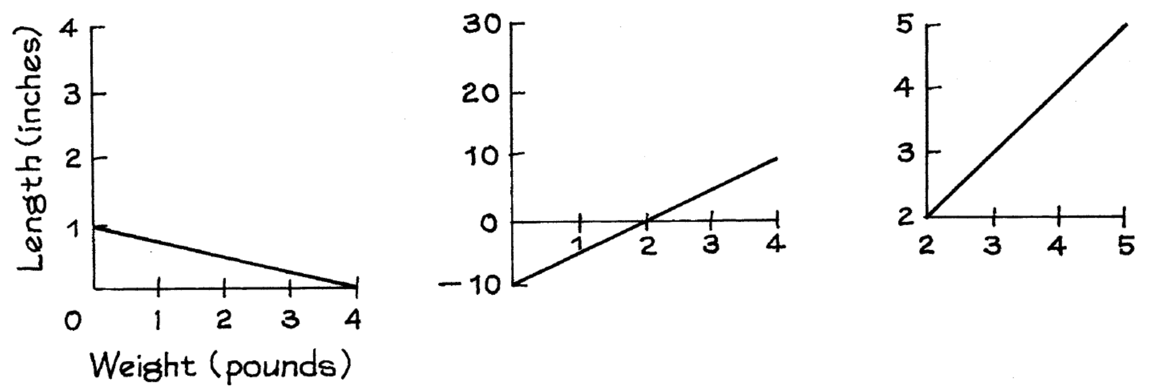


Figure 16. Figure 17. Figure 18.

NOTE:

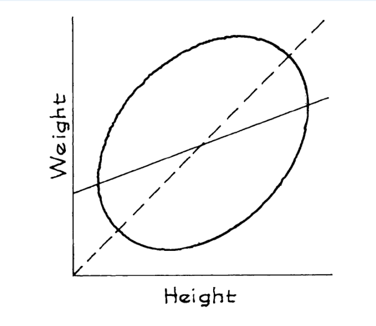
- For some lines with same slopes, they are parellel with each other
- With the positive slope, the line shows the increasing trend
- With the negative slope, the line shows the decreasing trend
- The larger the absolute value of slope, the steeper the line

Regression Line Equation

Slope of the regression line: estimate a unit increase in x result in the average change in y.

Y = mX + b
Y = m(X+1) +b = mX + m + b

Slope: $\frac{r \times SD\ of\ y}{SD\ of\ x}$ *Difference between the slope of SD line?*



Intercept of the regression line: the predicted value for y when x is 0

Y = intercept + slope × X

Steps to get the regression line equation from the data (x, y)s:

1. Calculate means and SDs for x and y
2. Calculate correlation coefficient r
3. Find the slope
4. Plug in a point to find the intercept

The you can find the predicted value for a new X

Plug in the new X into the equation

Least Squares Method

We can fit a lot of different lines for one data set. Among all lines, the one that makes the **smallest r.m.s. error** in predicting y from x is the regression line (least squares line).

The errors are squared to compute the r.m.s. error, and the regression line makes the r.m.s. error as small as possible.

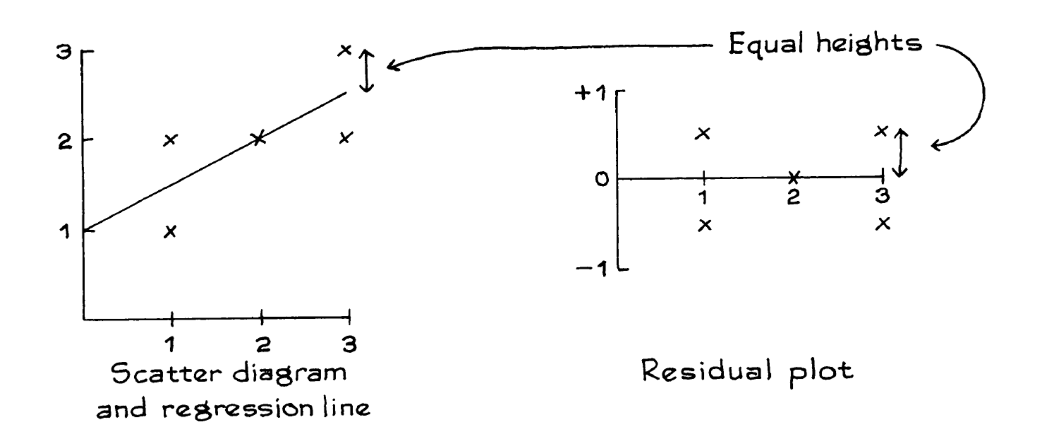
NOTE:

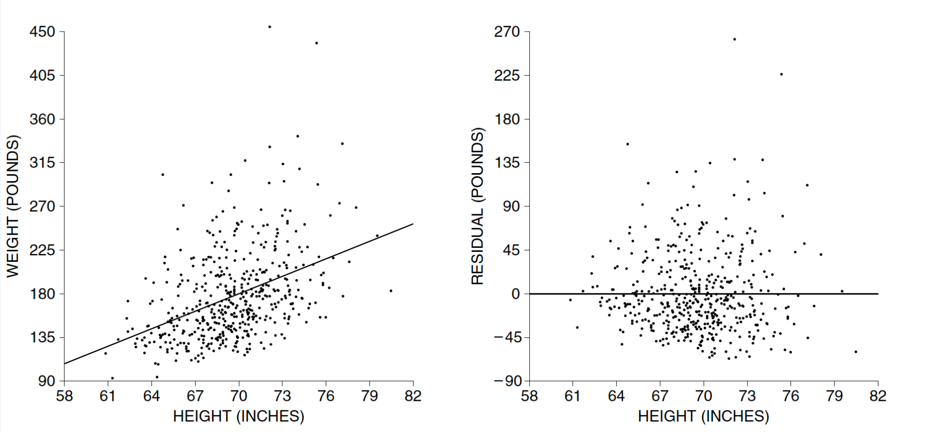
- Outliers can be very influencial to the least squares regression
- Linear regression might not be a good and reasonable fit to the data, we could check the residuals plot

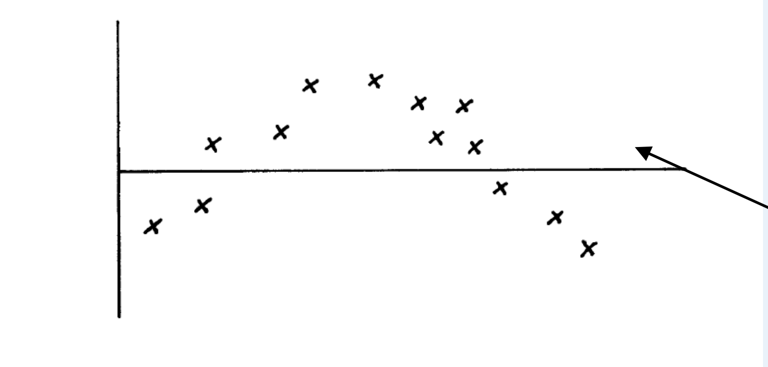
Residuals Plot

Residuals: prediction errors. Residual = measured value (actual value) - predicted value

Regression Diagnostic: Residuals should **randomly** lie around the line of **y = 0. NO STRONG PATTERN!**



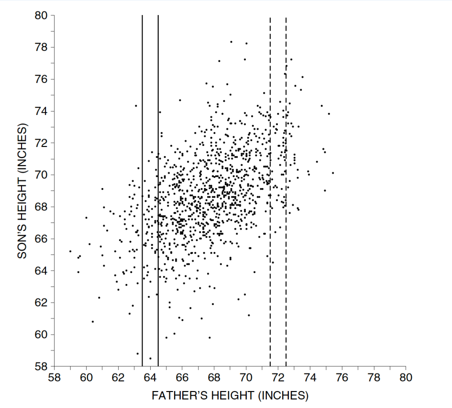


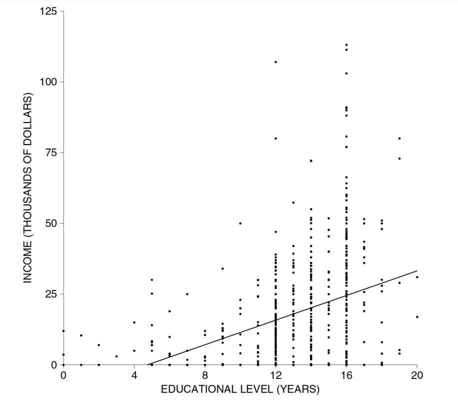


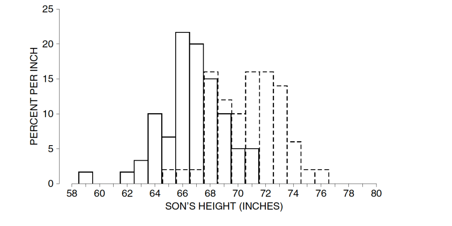
You could see a curve pattern here: it is probably a mistake to use a regression line

Normal Curve Inside a Vertical Strip

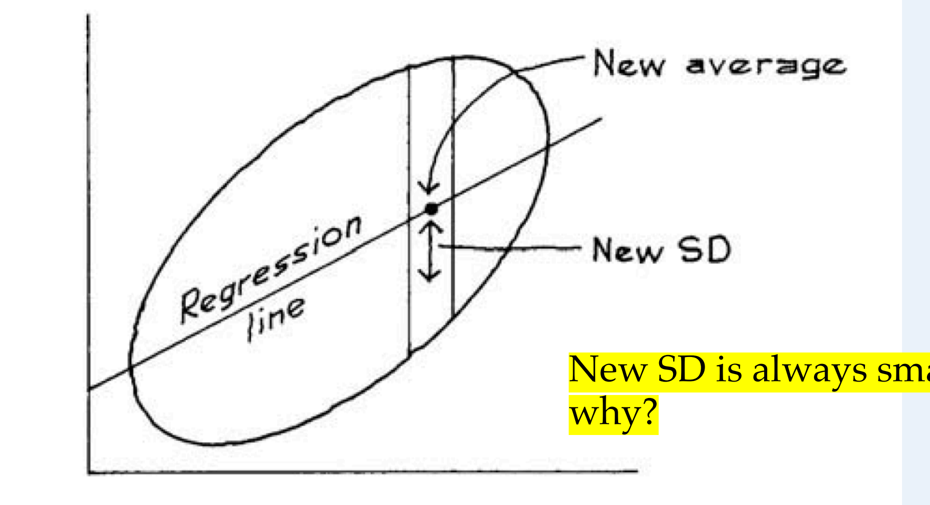
Homoscedasticity (football-shaped) vs Heteroscedasticity







Take the points in a narrow vertical strip. They will be off the regression line (up or down) by amounts similar in size to the r.m.s. error.



New SD is always smaller, why?

Take the points inside a narrow vertical strip.
Their y-values are a new data set.

The **new average** is given by the regression method.
The **new SD** is given by the **r.m.s. error** of the regression line.

Inside the strip, a typical y-value is around the new average—give or take the new SD.

Exercise I

- Average LSAT score = 162, SD = 6
- Average first year score = 68, SD = 10, r= 0.60

- Among the student who scored 174 on the LSAT, about what percentage had first year scores over 88?

- Slope, intercept, r.m.s error
- New average 80
- New SD
- New Z score

Exercise II

- Average SAT score = 550, SD = 80
- Average first year GPA = 2.6, SD = 0.6, r= 0.40

- Suppose the percentile rank of one student on the SAT is 90th, among the first-year students. Predict his percentile rank on first-year GPA. The scatter plot is foot-ball shaped.