Biostatistics

Chapter 7 Simple Linear Correlation and Regression

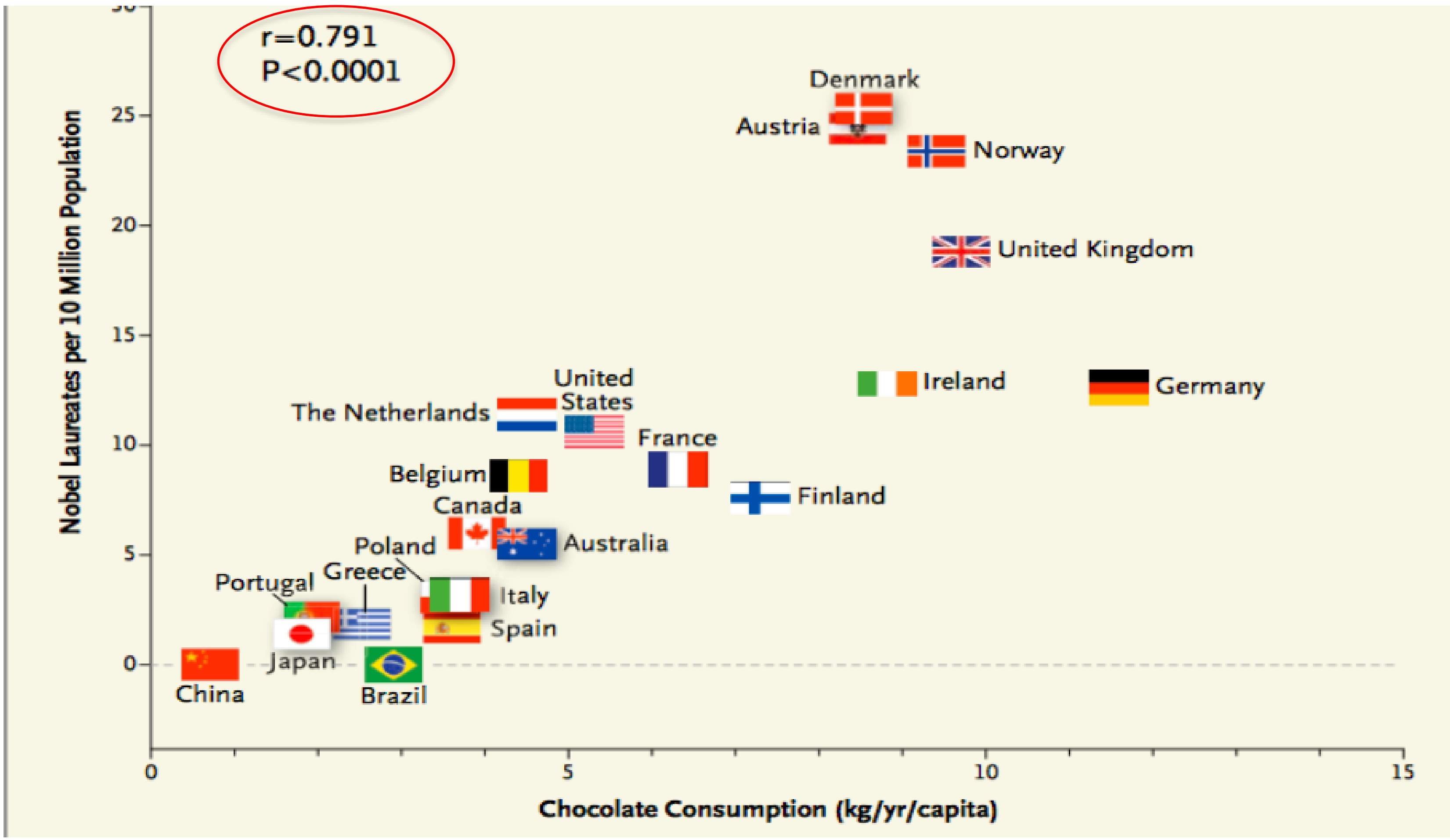
Jing Li

jing.li@sjtu.edu.cn

http://cbb.sjtu.edu.cn/~jingli/courses/2017fall/bi372/

Dept of Bioinformatics & Biostatistics, SJTU

Recall- eat chocolate



Covariance (协方差)

Variance $\sigma^2 = \mathrm{Var}(x) = E(x-\mu)^2$

$$\mathrm{var}(x) = S^2 = \frac{\sum_{i}^{n}(x_i - \bar{X})^2}{n-1}$$

Covariance is a measure of how much two random variables change together

$$\mathrm{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{n-1}$$

Interpreting Covariance

$\text{cov}(X,Y) > 0 \longrightarrow$ X and Y are positively correlated

$\text{cov}(X,Y) < 0 \longrightarrow$ X and Y are inversely correlated

$\text{cov}(X,Y) = 0 \longrightarrow$ X and Y are independent

Correlation coefficient

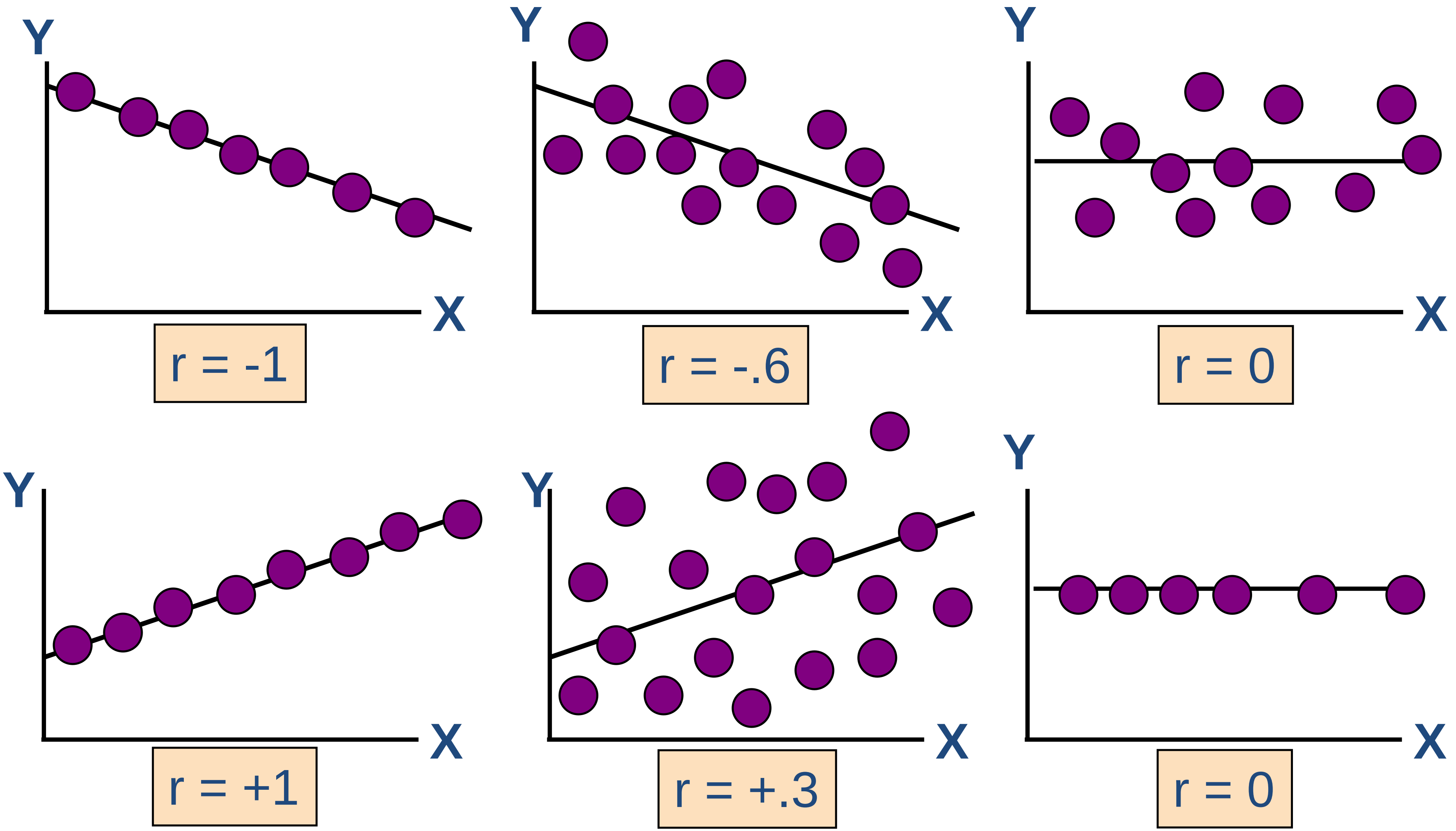
Pearson's Correlation Coefficient is standardized covariance (unitless):

$$r = \frac{\mathrm{cov}\,ariance(x, y)}{\sqrt{\mathrm{var}\, x}\sqrt{\mathrm{var}\, y}}$$

Correlation

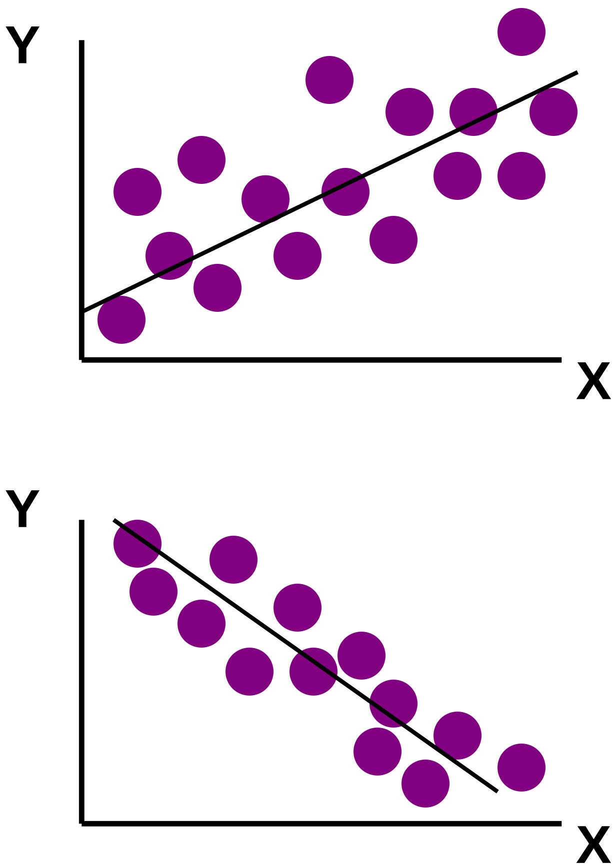
- Measures the relative strength of the *linear* relationship between two variables
- **Ranges between –1 and 1**
- The closer to –1, the stronger the **negative** linear relationship
- The closer to 1, the stronger the **positive** linear relationship
- The closer to 0, the weaker any positive linear relationship

Scatter Plots of Data with Various Correlation Coefficients

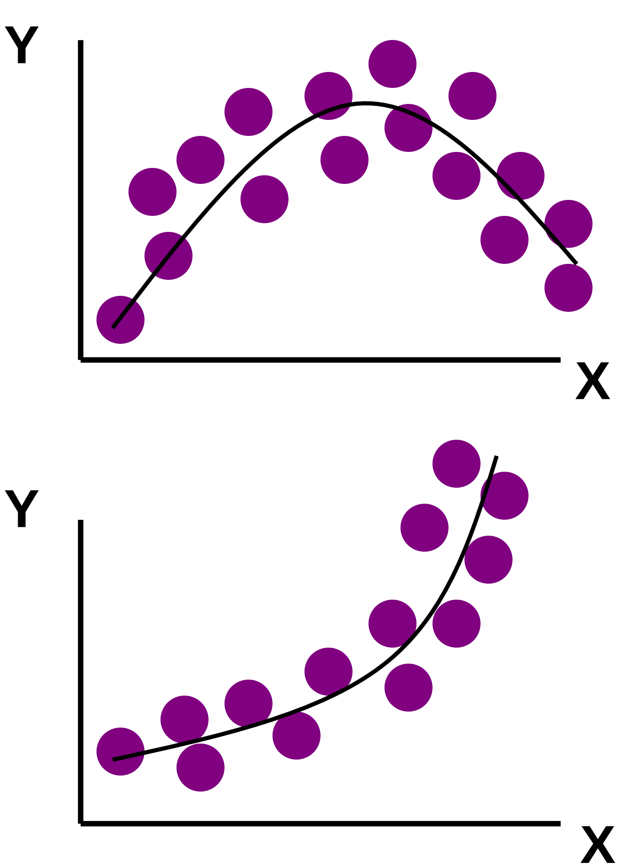


Linear Correlation

Linear relationships

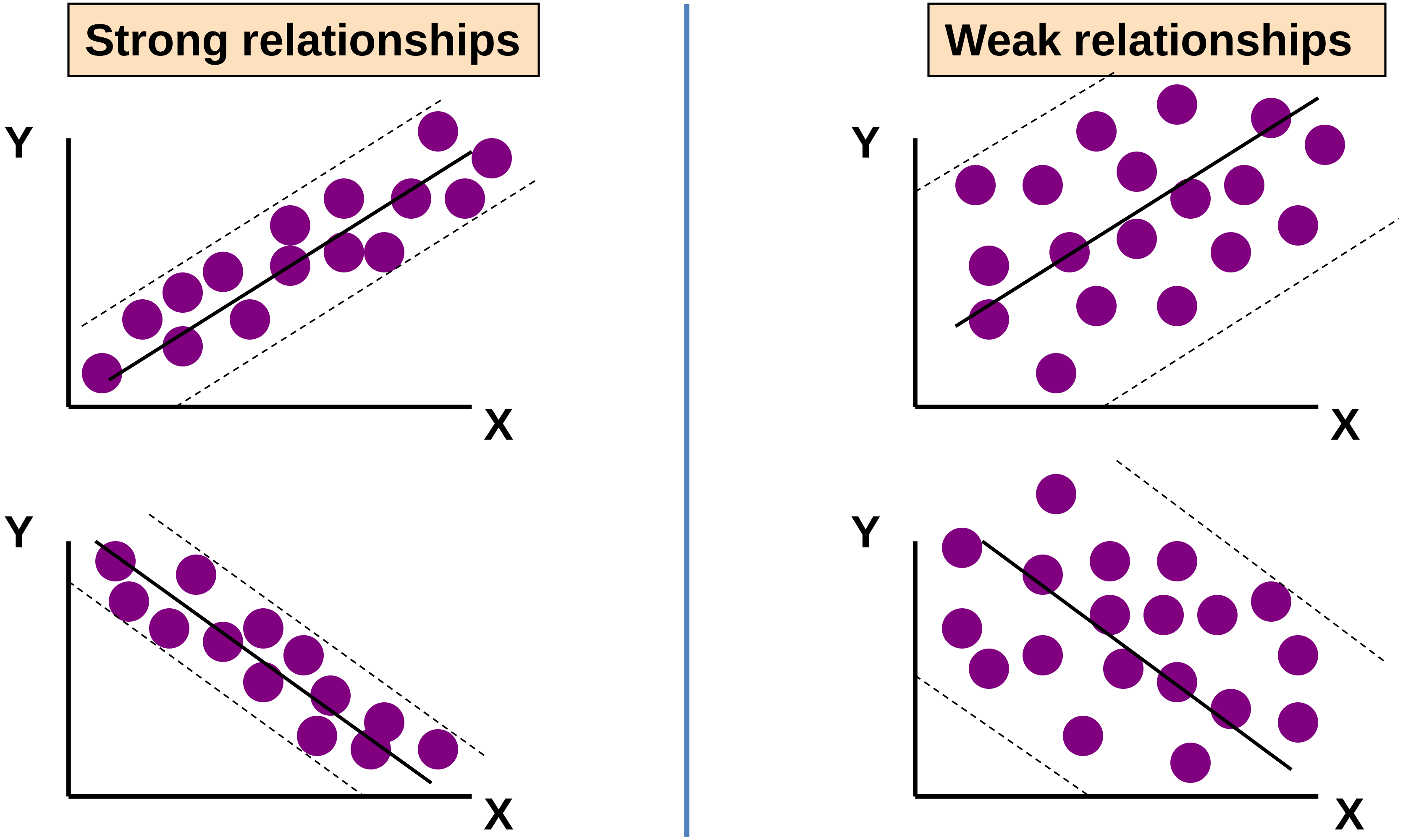


Curvilinear relationships

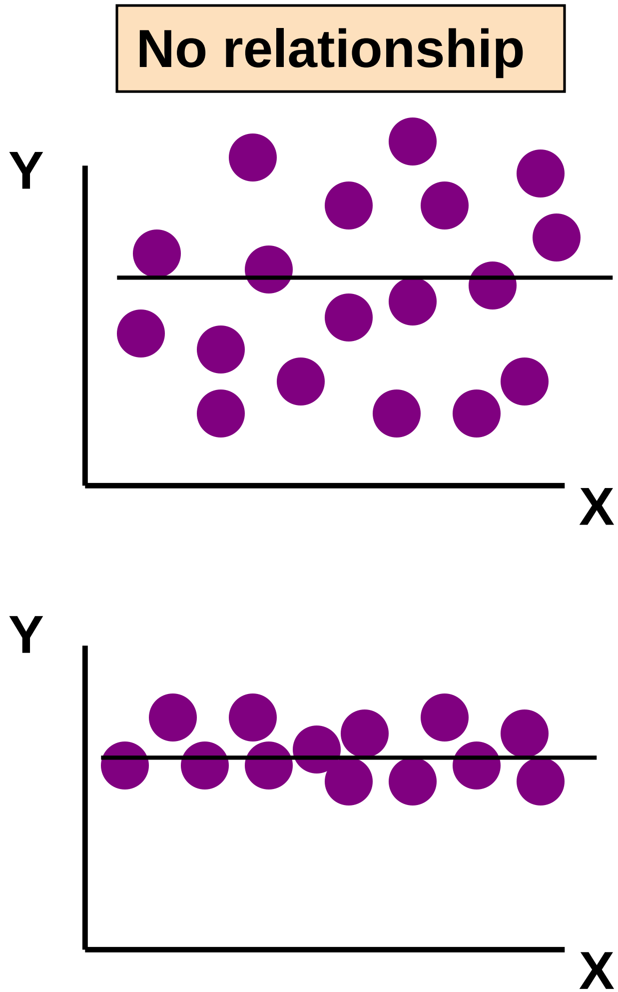


- Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Linear Correlation



Linear Correlation



Calculating by hand…

$$\hat{r} = \frac{\operatorname{cov} ariance(x, y)}{\sqrt{\operatorname{var} x}\sqrt{\operatorname{var} y}} = \frac{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}\sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}}$$

Simpler calculation formula…

$$\hat{r} = \frac{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\cancel{n-1}}}{\sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\cancel{n-1}}}\sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\cancel{n-1}}}} =$$

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

Numerator of covariance

$$\hat{r} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

Numerators of variance

Correlation Analysis… "$-1 \leq \rho < 1$"

- If the correlation coefficient is close to +1 that means you have a strong positive relationship.
- If the correlation coefficient is close to -1 that means you have a strong negative relationship.
- If the correlation coefficient is close to 0 that means you have no correlation.
- WE HAVE THE ABILITY TO TEST THE HYPOTHESIS
- H_0: $\rho = 0$

Distribution of the correlation coefficient

$$SE(\hat{r}) = \sqrt{\frac{1 - r^2}{n - 2}}$$

The sample correlation coefficient follows a T-distribution with n-2 degrees of freedom (since you have to estimate the standard error).

$$t = r / \sqrt{\frac{1 - r^2}{n - 2}}$$

History- Galton's Sweet Pea Data

- In *Natural Inheritance*, Galton (1894) provided a table, which contained a list of frequencies of daughter seeds of various sizes organized in rows according to the size of their parent seeds

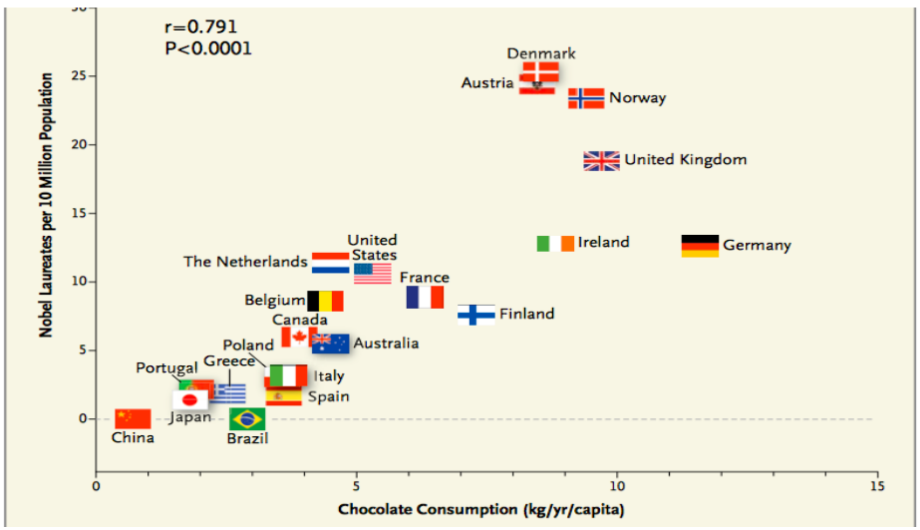
Diameter of Parent Seed (0.01 inch)	Diameter of Daughter Seed (0.01 inch)	Frequency
21.00	14.67	22
21.00	15.67	8
21.00	16.67	10
21.00	17.67	18
21.00	18.67	21
21.00	19.67	13
21.00	20.67	6
21.00	22.67	2
20.00	14.66	23
20.00	15.66	10
20.00	16.66	12
20.00	17.66	17
20.00	18.66	20
20.00	19.66	13

- In 1896, Pearson published his first rigorous treatment of correlation and regression
- A simpler proof than Pearson's for the product-moment method proposed by Ghiselli (1981)

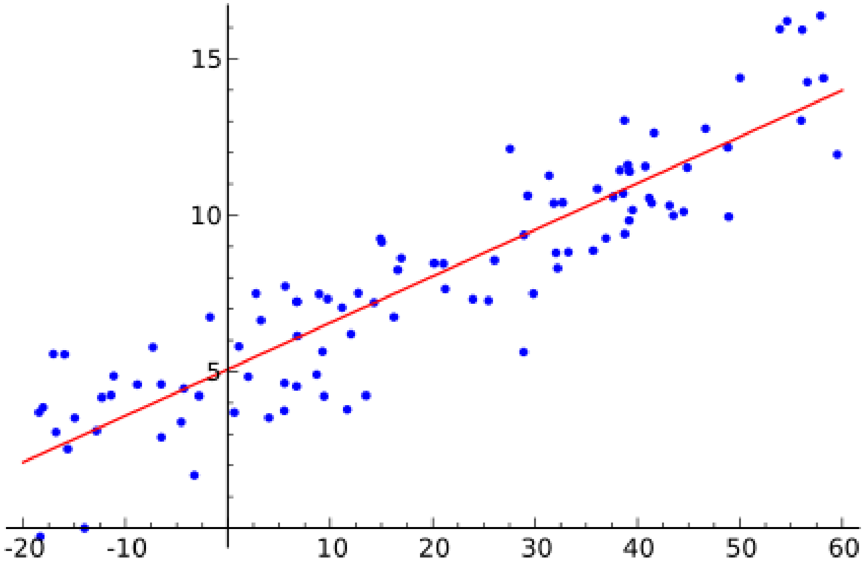
Linear Regression

Can we predict Novel Laureates per 10 million population using chocolate consumption?

Chocolate ~ Nobel laureates



Simple Linear Regression



Linear Regression

- Regression analysis is used to predict the value of one variable (the dependent variable, 因变量) on the basis of other variables (the *independent variables*, *自变量*).
- Dependent variable: denoted **Y**
- Independent variables: denoted $\mathbf{X_1, X_2, \ldots, X_k}$
- If we only have ONE independent variable, the model is

$$y = \beta_0 + \beta_1 x + \varepsilon$$

which is referred to as simple linear regression. We would be interested in estimating β_0 and β_1 from the data we collect.

Linear Regression

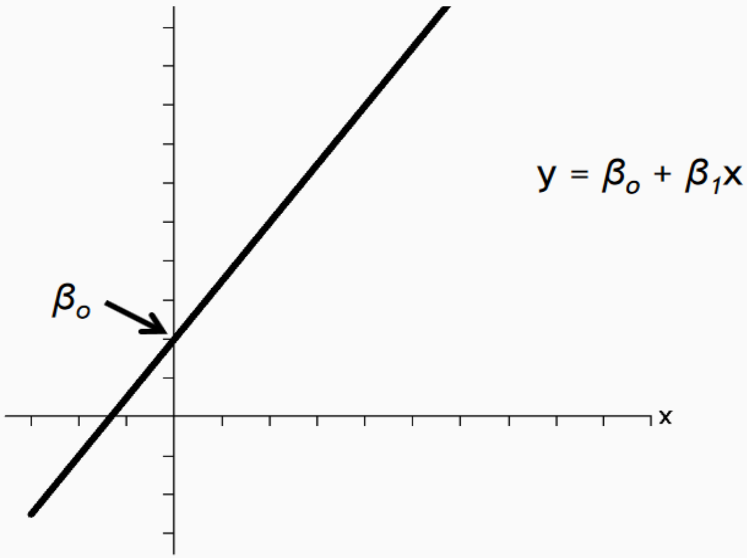
- Variables: $$y = \beta_0 + \beta_1 x + \varepsilon$$
  X = Independent Variable (we provide this)
  Y = Dependent Variable (we observe this)

- Parameters:
  β_0 = Y-Intercept
  β_1 = Slope
  $\varepsilon \sim$ Normal Random Variable ($\mu_\varepsilon = 0$, $\sigma_\varepsilon = ???$) [Noise]

The Intercept, β_0

$$y = \beta_0 + \beta_1 x + \varepsilon$$

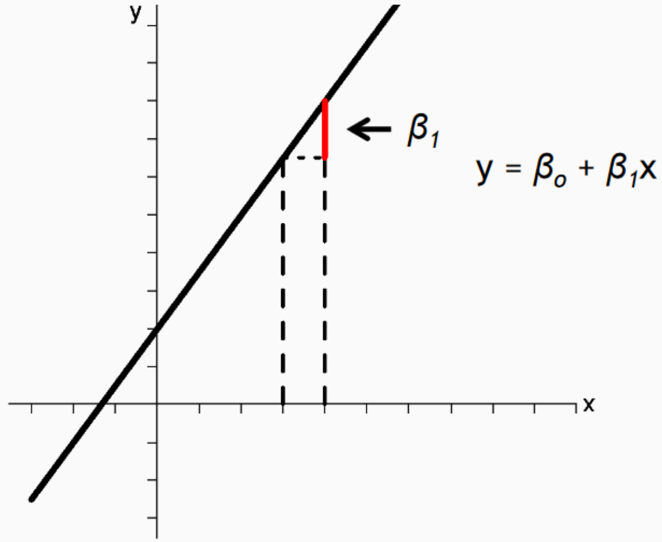
- The intercept β_o is the value of y when x is 0
  - It is the point on the graph where the line crosses the y (vertical) axis, at the coordinate $(0, \beta_o)$



The Slope, β1

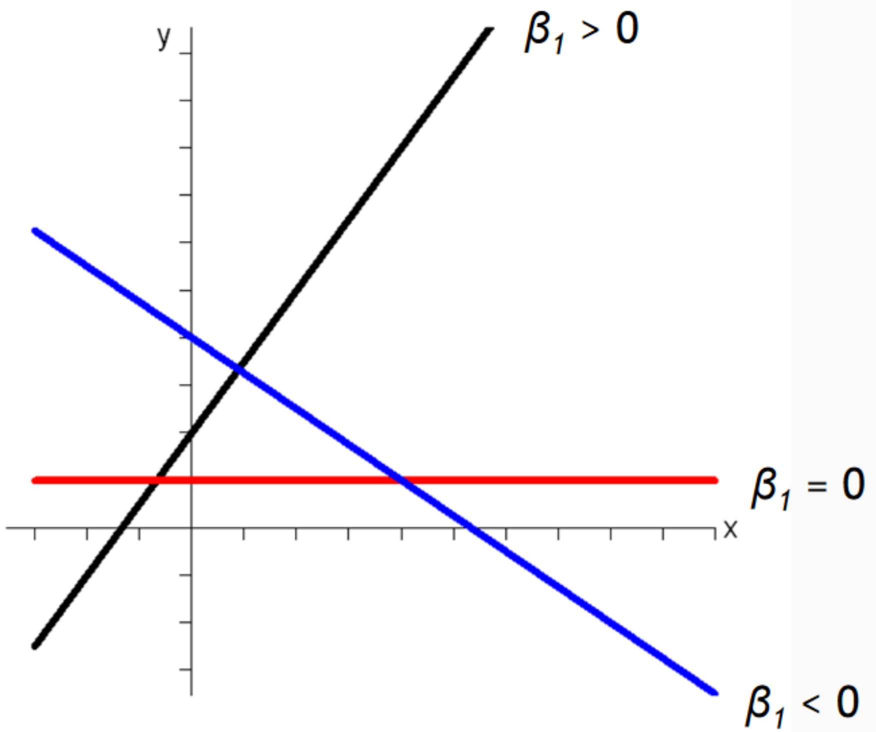
$$y = \beta_0 + \beta_1 x + \varepsilon$$

- The slope β_1 is the change in y corresponding to a unit increase in x



The Slope, β1

- The slope β_1 is the change in y corresponding to a unit increase in x: β_1 is difference in y-values for x+1 compared to x



Building the Model – Collect Data

- Test 2 Grade = β_0 +β1*(Test 1 Grade)

- From Data:

  Estimate β_0

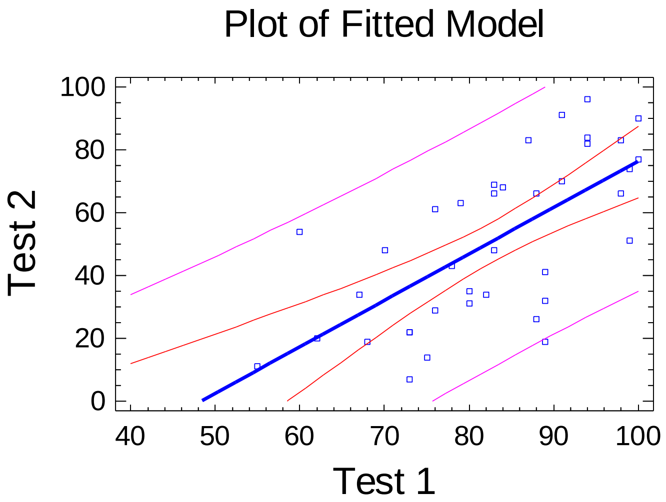
  Estimate β_1

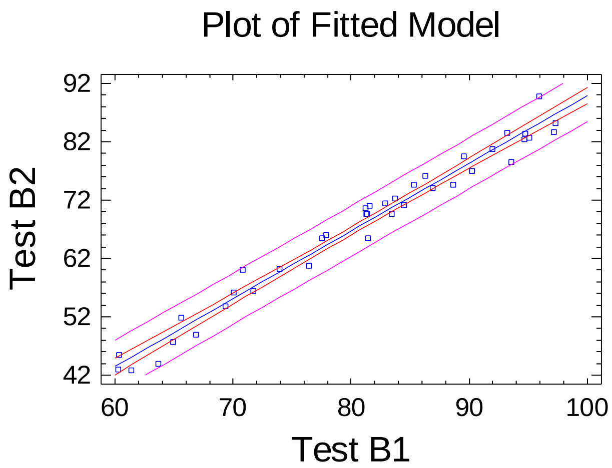
  Estimate σ_ε

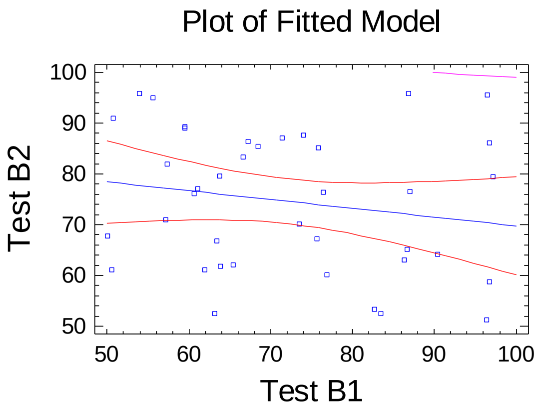
Student	Test 1	Test 2
1	50	32
2	51	33
3	52	34
4	53	35
5	54	36
6	55	37
7	56	39
8	57	40
9	58	41
10	59	42
11	60	43
12	61	44
13	62	46
14	63	47
15	64	48
16	65	49
17	66	50
18	67	51
19	68	53
20	69	54
21	70	55
22	71	56
23	72	57

Linear Regression Analysis…

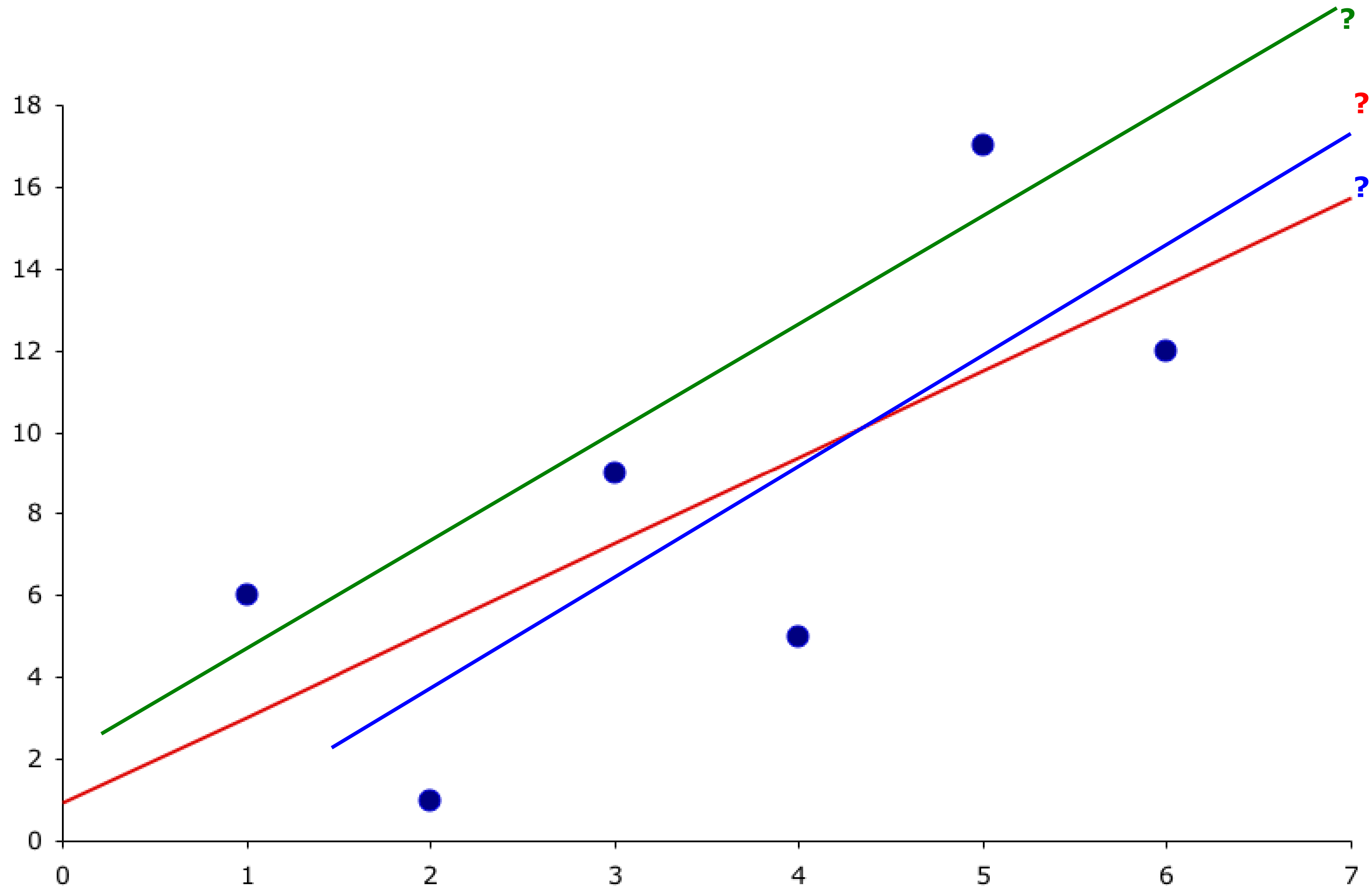
$$y = \beta_0 + \beta_1 x + \varepsilon$$







Which line has the best "fit" to the data?



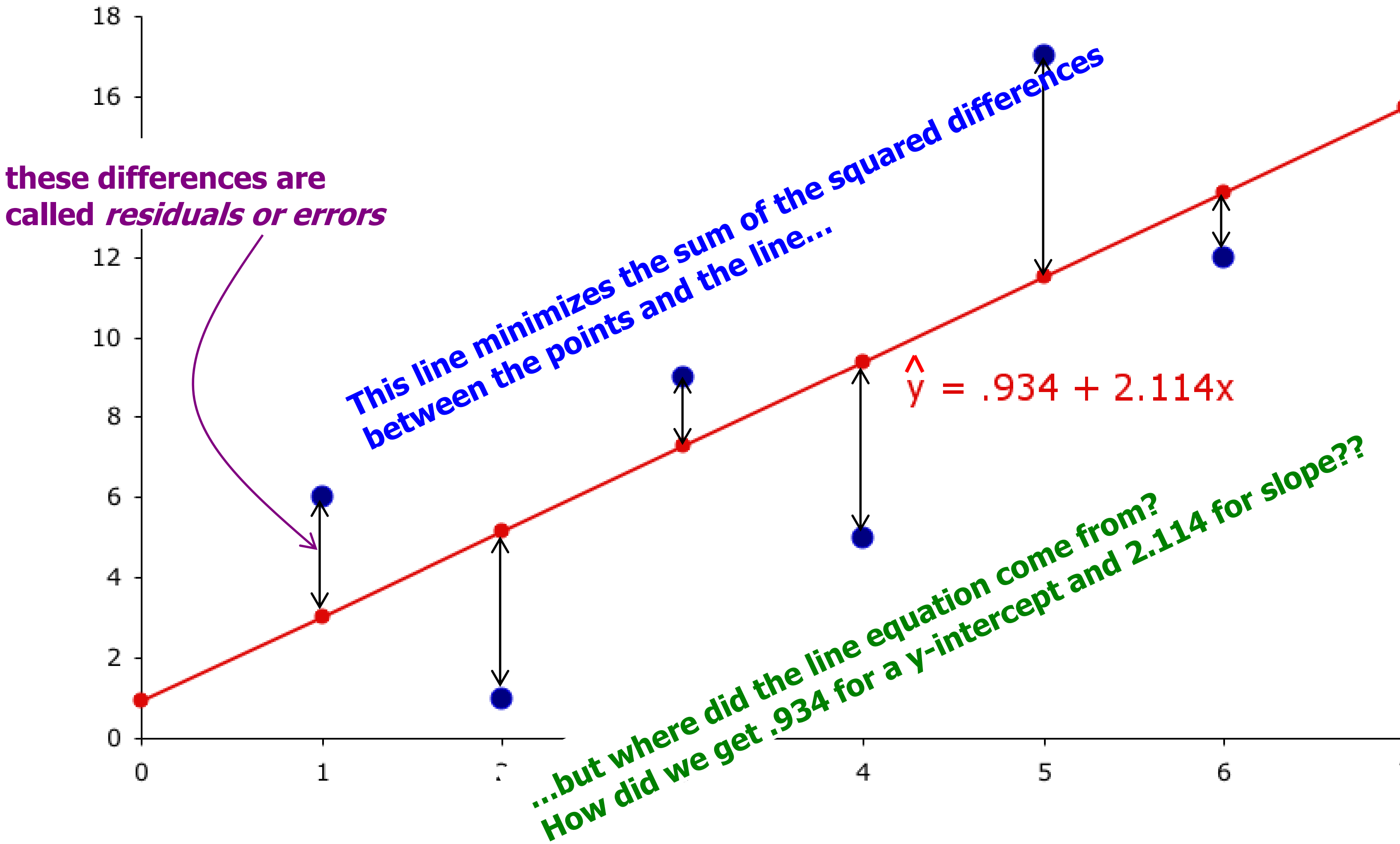
Estimating the Coefficients…

- In much the same way we base estimates o μ on $\overline{x}$, we estimate β_0 with b_0 and β_1 with b_1, the y-intercept and slope (respectively) of the ***least squares*** or ***regression line*** given by:

$$\boxed{\hat{y} = b_0 + b_1 x} \qquad\qquad y = \beta_0 + \beta_1 x$$

- (This is an application of the least squares method and it produces a straight line that ***minimizes*** the sum of the squared differences between the points and the line)

Least Squares Line…



Least Squares Line…
[sure glad we have computers now!]

- The coefficients **b_1** and **b_0** for the least squares line…

$$\hat{y} = b_0 + b_1 x$$

- …are calculated as:

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$$\boxed{b_1 = \frac{s_{xy}}{s_x^2}}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\boxed{b_0 = \bar{y} - b_1 \bar{x}}$$

$$SSE = \text{å}\,(Y - \hat{Y})^2 = \text{å}\,(Y - b_0 - b_1 X)^2$$

Least Squares Line… See if you can estimate Y-intercept and slope from this data

Recall…

Data → Statistics → Information

Data Points:

x	y
1	6
2	1
3	9
4	5
5	17
6	12

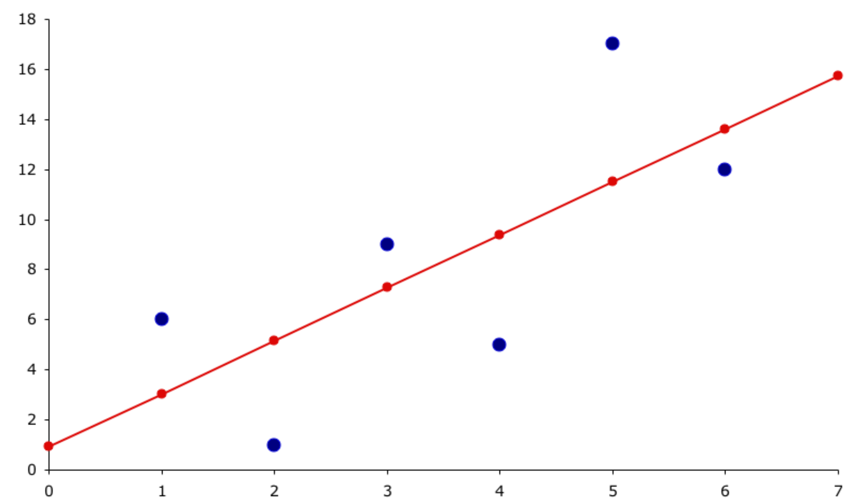
$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$\boxed{b_1 = \frac{s_{xy}}{s_x^2}}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\boxed{b_0 = \bar{y} - b_1\bar{x}}$$



$\hat{y} = .934 + 2.114x$

Least Squares Line… See if you can estimate Y-intercept and slope from this data

	X	Y	X - Xbar	Y - Ybar	(X-Xbar)*(Y-Ybar)	$(X - Xbar)^2$
	1	6	-2.500	-2.333	5.833	6.250
	2	1	-1.500	-7.333	11.000	2.250
	3	9	-0.500	0.667	-0.333	0.250
	4	5	0.500	-3.333	-1.667	0.250
	5	17	1.500	8.667	13.000	2.250
	6	12	2.500	3.667	9.167	6.250
Sum =	**21**	**50**	**0.000**	**0.000**	**37.000**	**17.500**

Xbar =	**3.500**	
Ybar =	**8.333**	
s_{xy} =	**7.400**	37.00/(6-1)
s_x^2 =	**3.500**	17.5/(6-1)
b_1 =	**2.114**	7.4/3.5
b_0 =	**0.933**	8.33 - 2.114*3.50

Example: Arm Circumference and Height

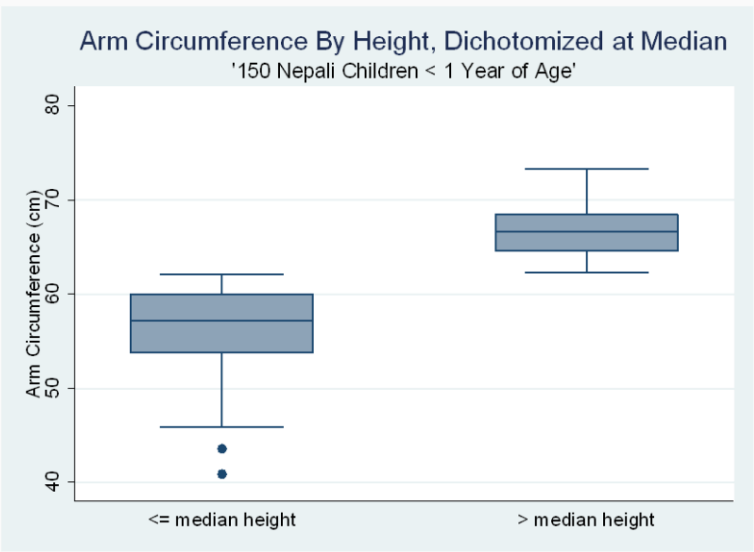
- Data on anthropomorphic measures from a random sample of 150 Nepali children [0, 12) months old

- Question: what is the relationship between average arm circumference and height

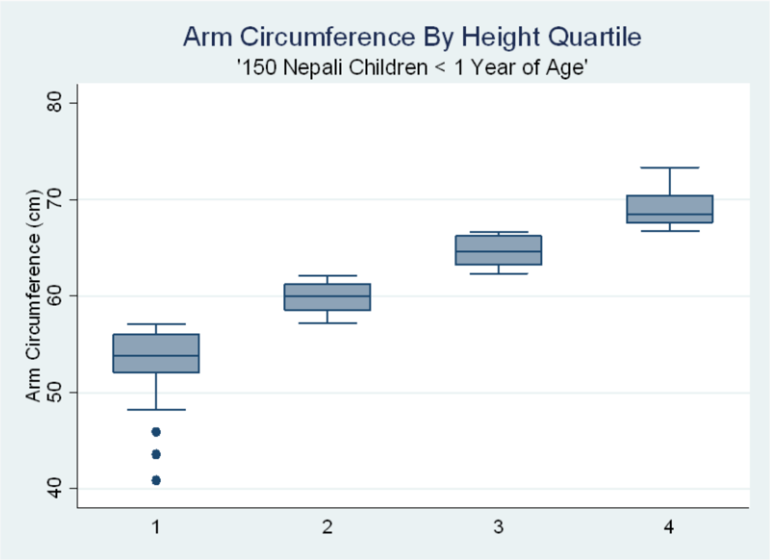
- Data:
  - Arm circumference: mean 12.4 cm, SD 1.5 cm, range 7.3 cm - 15.6 cm
  - Height: mean 61.6 cm, SD 6.3 cm, range 40.9 cm - 73.3 cm

Arm Circumference and Height

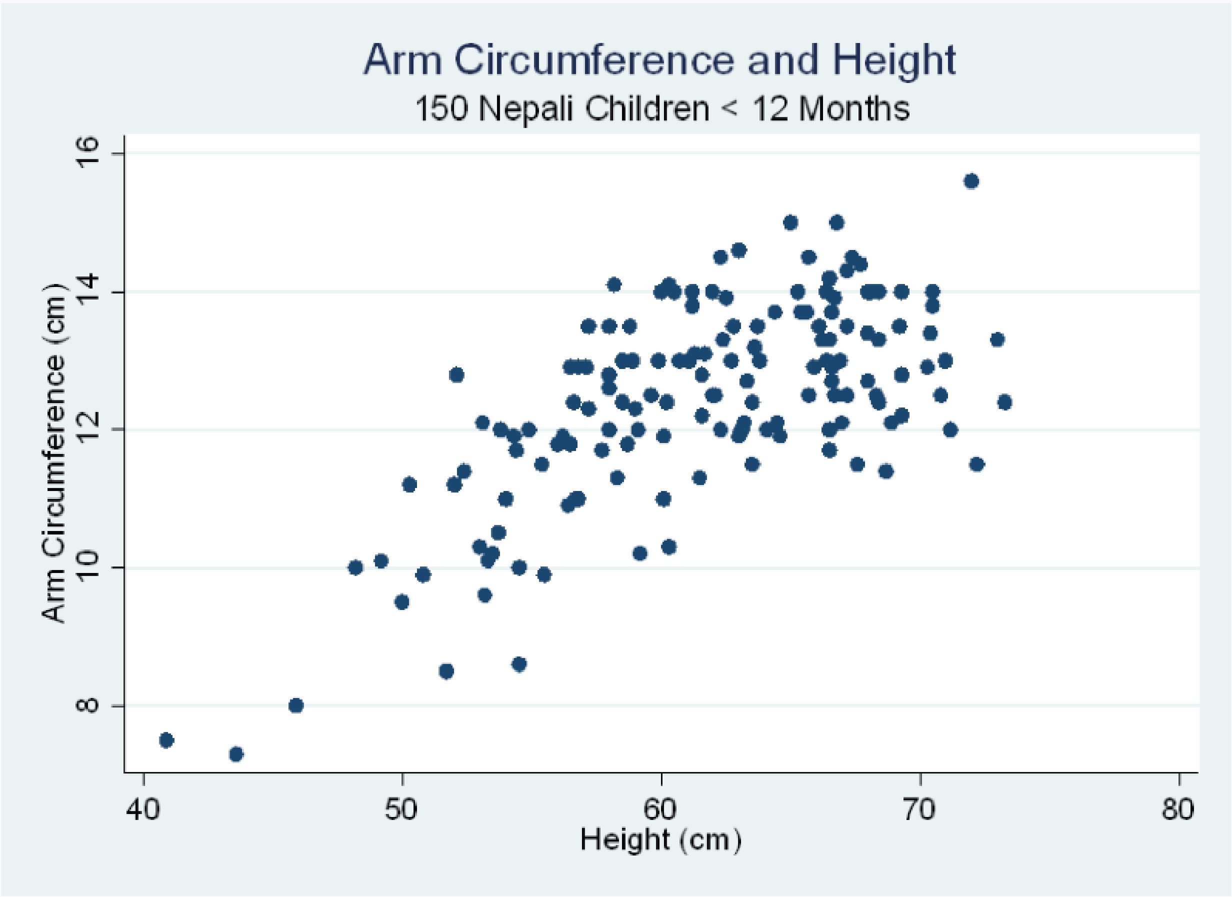
T-test



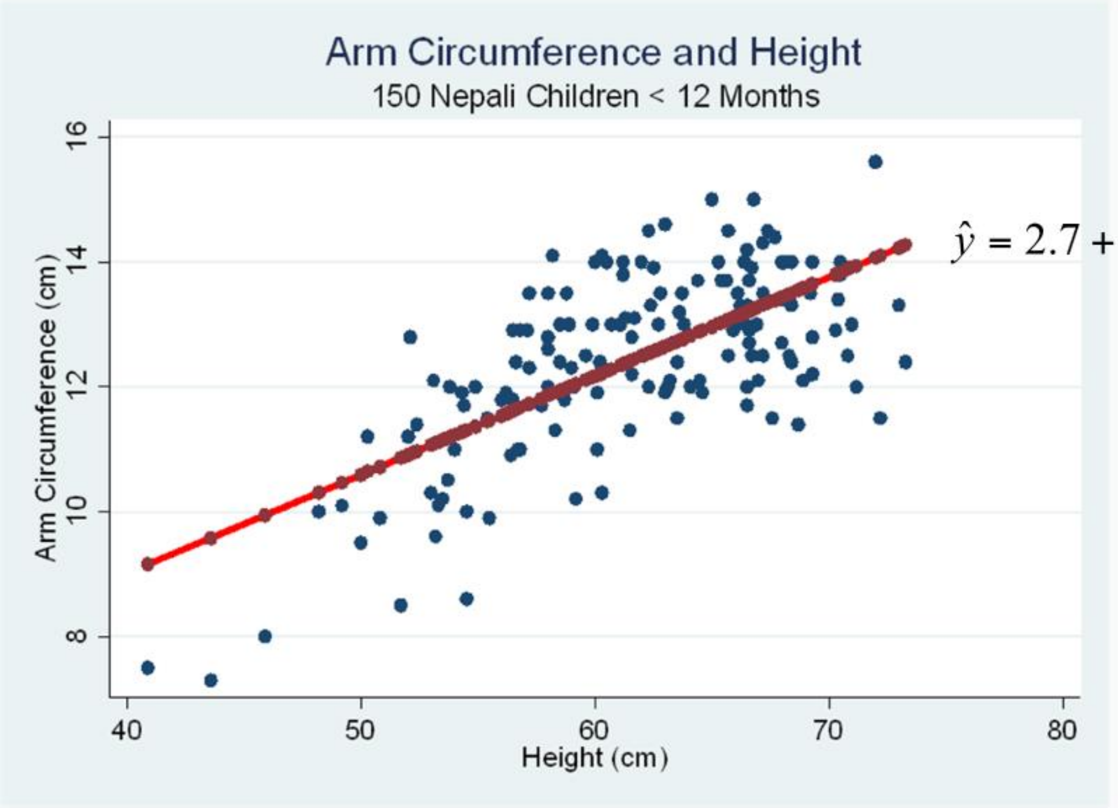
ANOVA



Visualizing Arm Circumference and Height Relationship



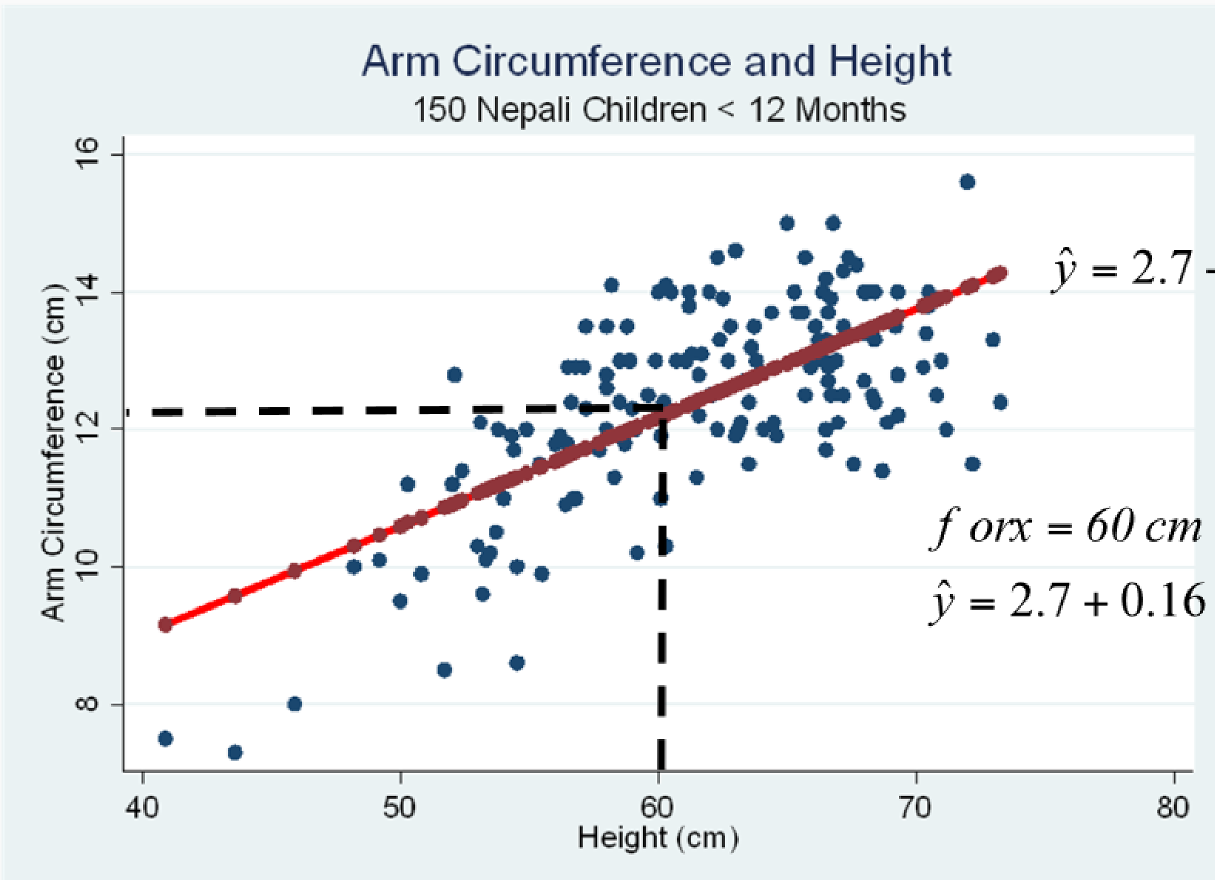
Scatterplot with regression line



$\hat{y} = 2.7 + 0.16x$

Example: Arm Circumference and Height

- **Estimated mean arm circumference for children 60 cm in height**



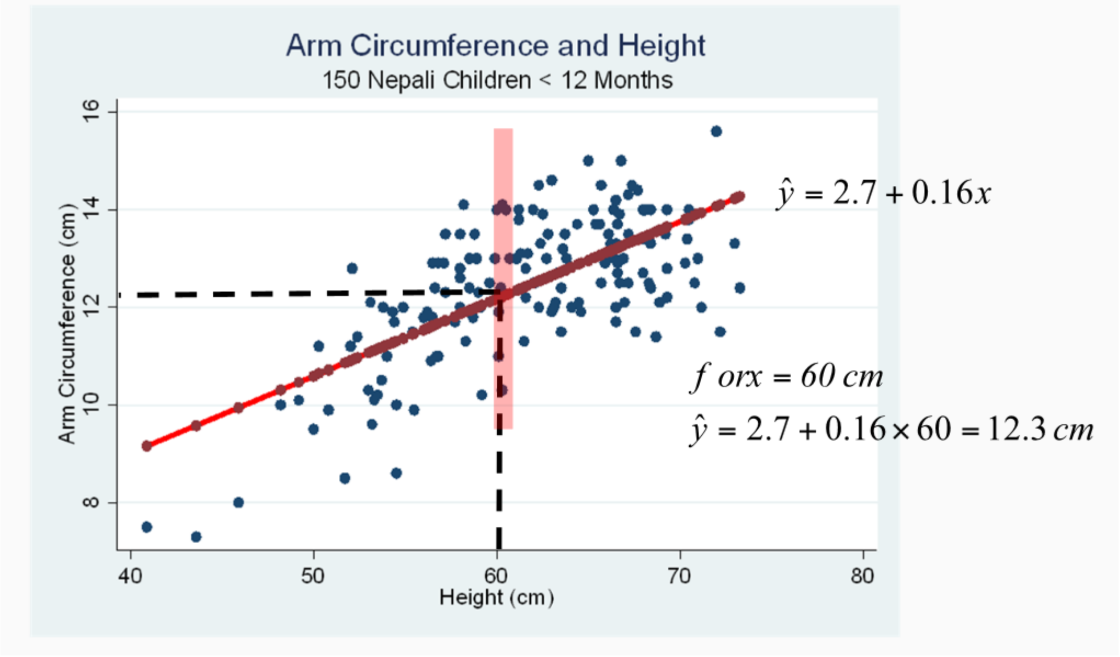
$$\hat{y} = 2.7 + 0.16x$$

$$f\ orx = 60\ cm$$

$$\hat{y} = 2.7 + 0.16 \times 60 = 12.3\ cm$$

Example: Arm Circumference and Height

- **Estimated mean arm circumference for children 60 cm in height**



Notice, most points don't fall directly on the line: we are estimating the mean arm circumference of children 60 cm tall: observed points vary about the estimated mean

Linear regression assumes that…

- The relationship between X and Y is linear
- Y is distributed normally at each value of X
- The variance of Y at every value of X is the same (homogeneity of variances)
- The observations are independent