Statistics in Evidence Based Medicine II

Lecture 2: Regression and Correlation

Rizwana Rehman, PhD
Regional Statistician
Southeast Epilepsy Center of Excellence
Durham VA Medical Center, Durham NC
Rizwana.Rehman@va.gov
(919)286-0411 ext: 5024

Audio Information: Dial 1-855-767-1051
Conference ID 61304911

Text Books

- **Main:** **Statistics at Square One (2010)**

  M J Campbell & T D V Swinscow

  http://www.phsource.us/PH/EPI/Biostats/

- **Secondary:** **Basic and Clinical Biostatistics (2004)**

  Beth Dawson, Robert G. Trapp

  http://www.accessmedicine.com/resourceTOC.aspx?resourceID=62

- For more information, program materials, and to complete evaluation for CME credit visit

  www.epilepsy.va.gov/Statistics

Audio Information: Dial 1-855-767-1051
Conference ID 61304911

Overview

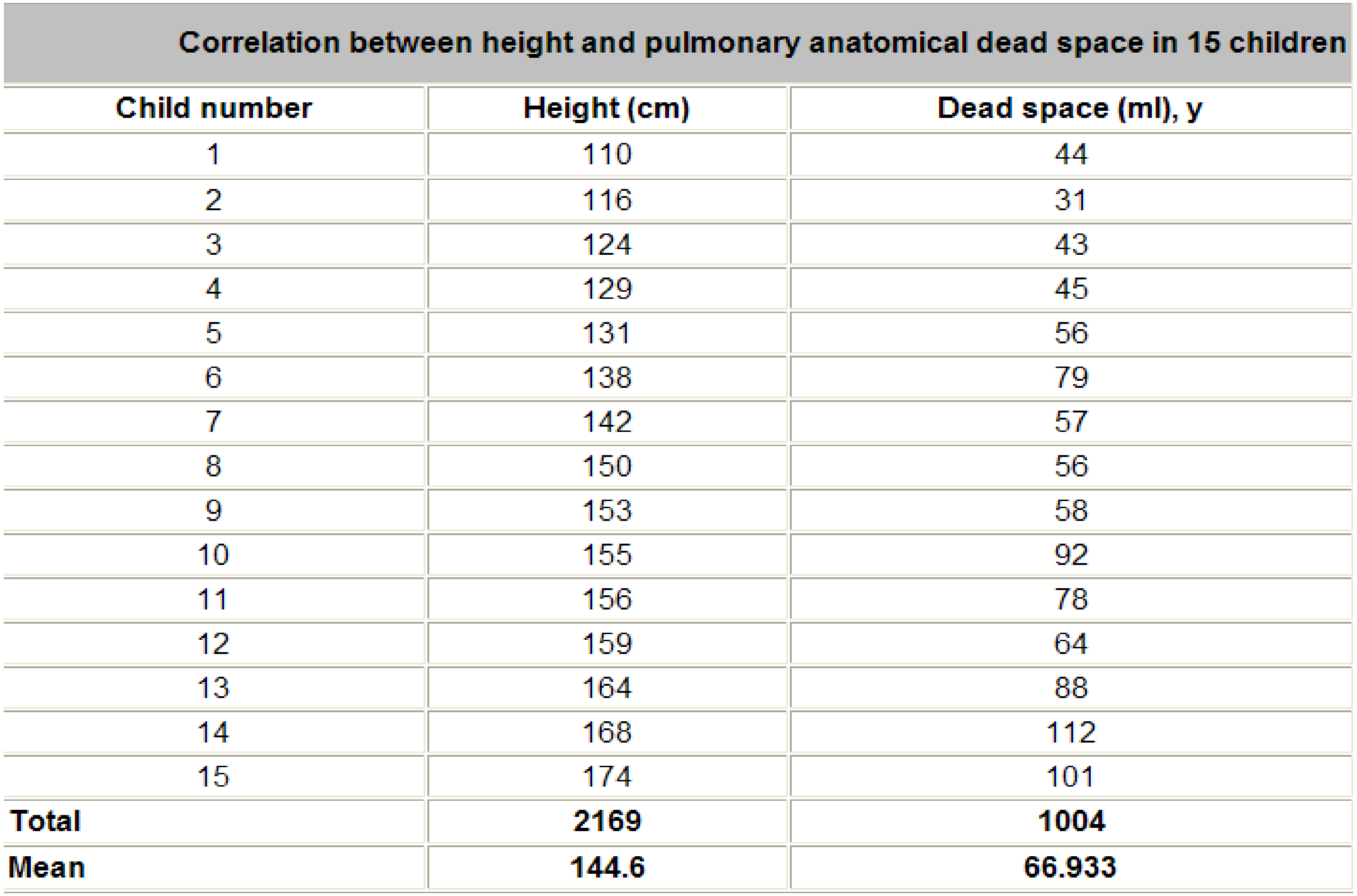
- Correlation coefficient *r*
  - Test of significance
- Regression
  - Test of significance
- Coefficient of determination r^2

Correlation

- In correlation we look for a linear association between two continuous variables *x* and *y*.

- Strength of association is summarized by the correlation coefficient *r*.

Example of Correlation

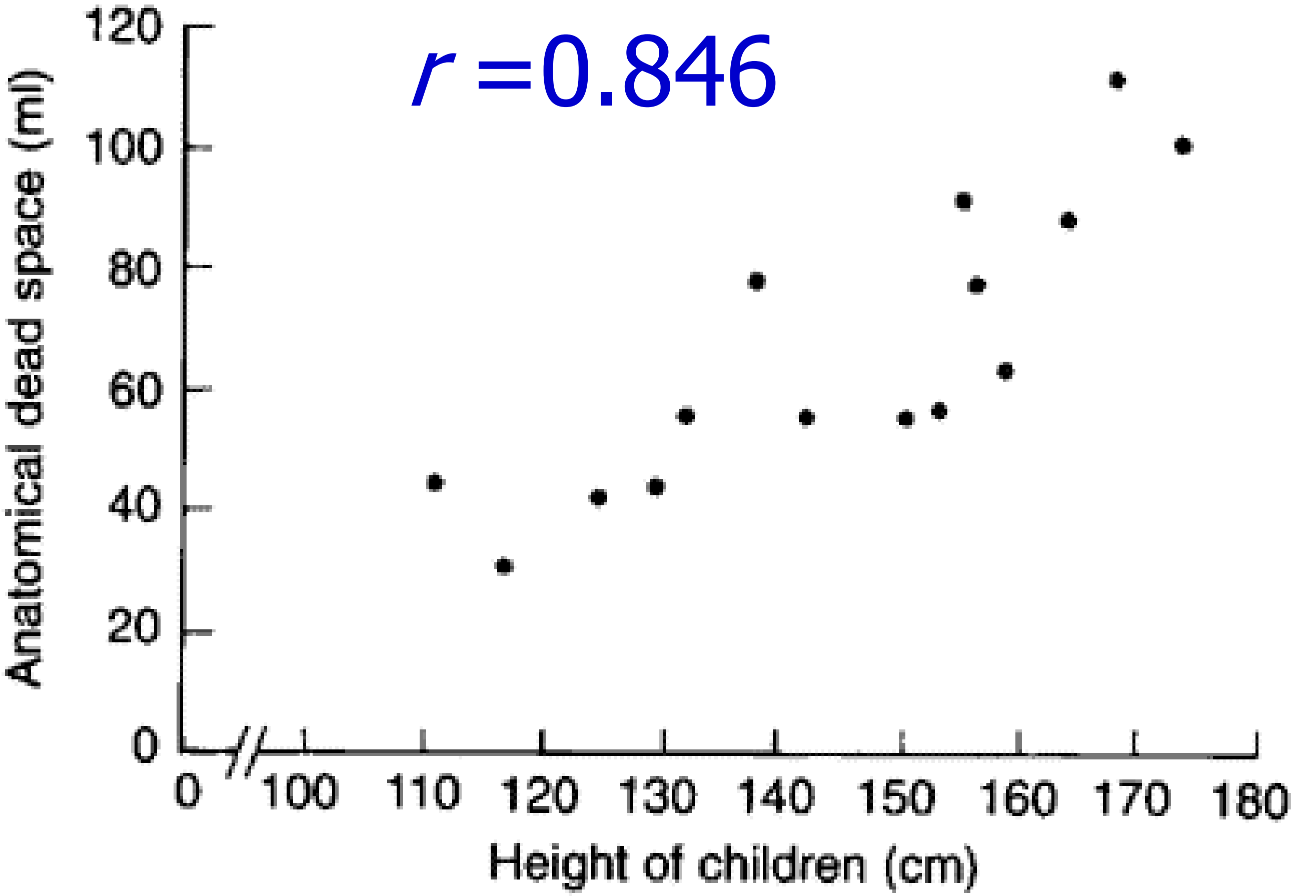


Correlation between height and pulmonary anatomical dead space in 15 children

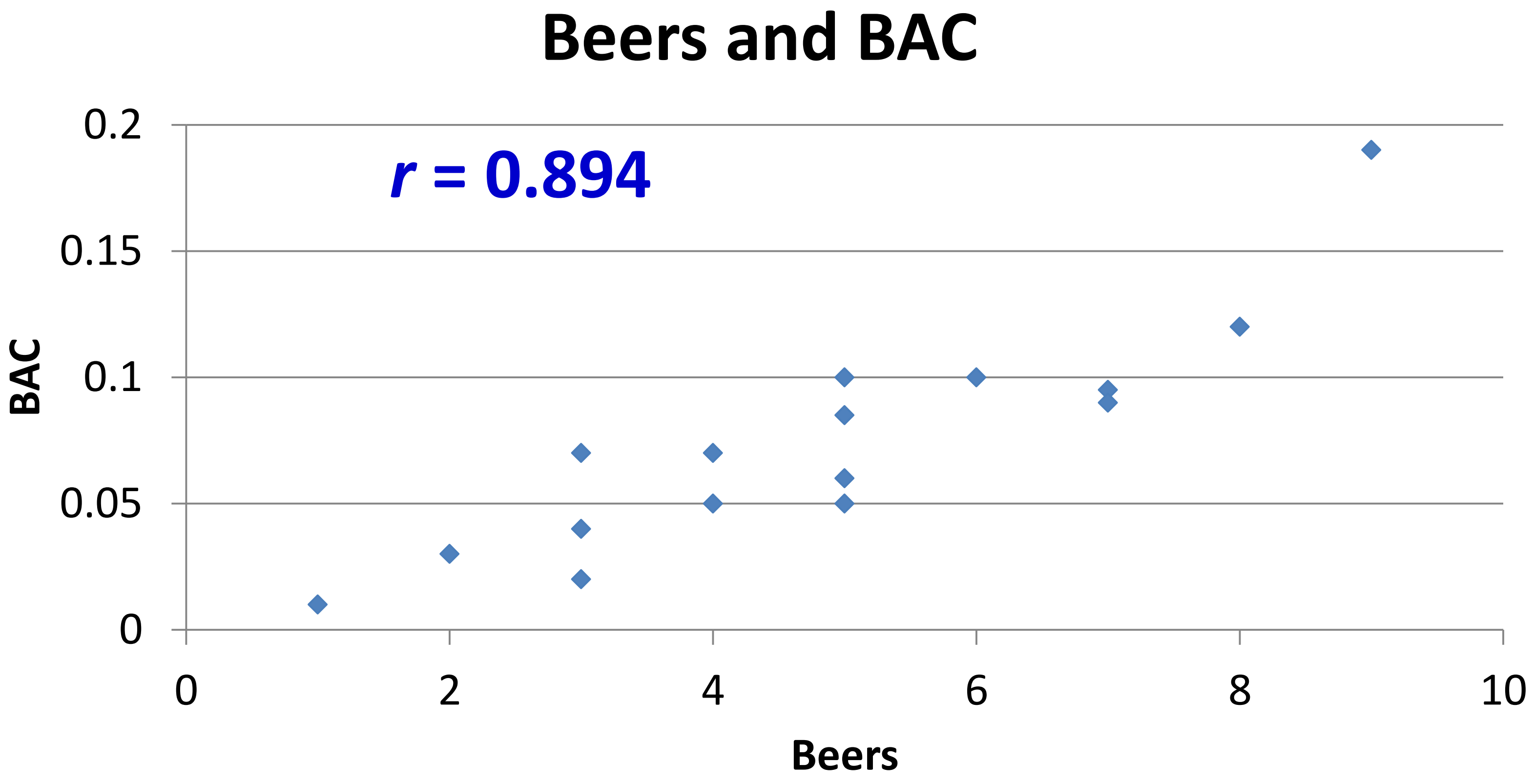
Child number	Height (cm)	Dead space (ml), y
1	110	44
2	116	31
3	124	43
4	129	45
5	131	56
6	138	79
7	142	57
8	150	56
9	153	58
10	155	92
11	156	78
12	159	64
13	164	88
14	168	112
15	174	101
Total	**2169**	**1004**
Mean	**144.6**	**66.933**

Statistics at Square One

Scatter Diagram for Correlation

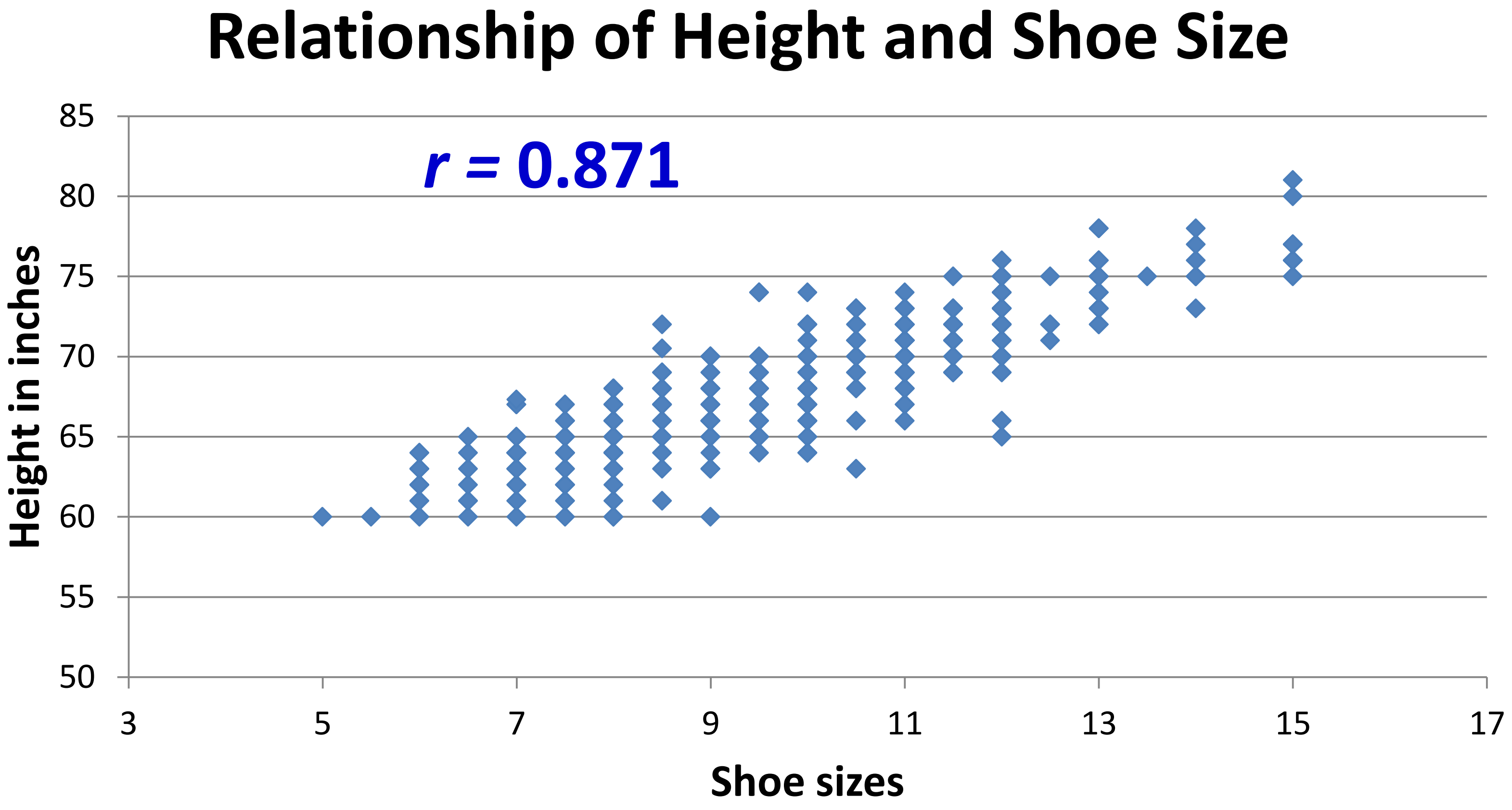


Example of Correlation



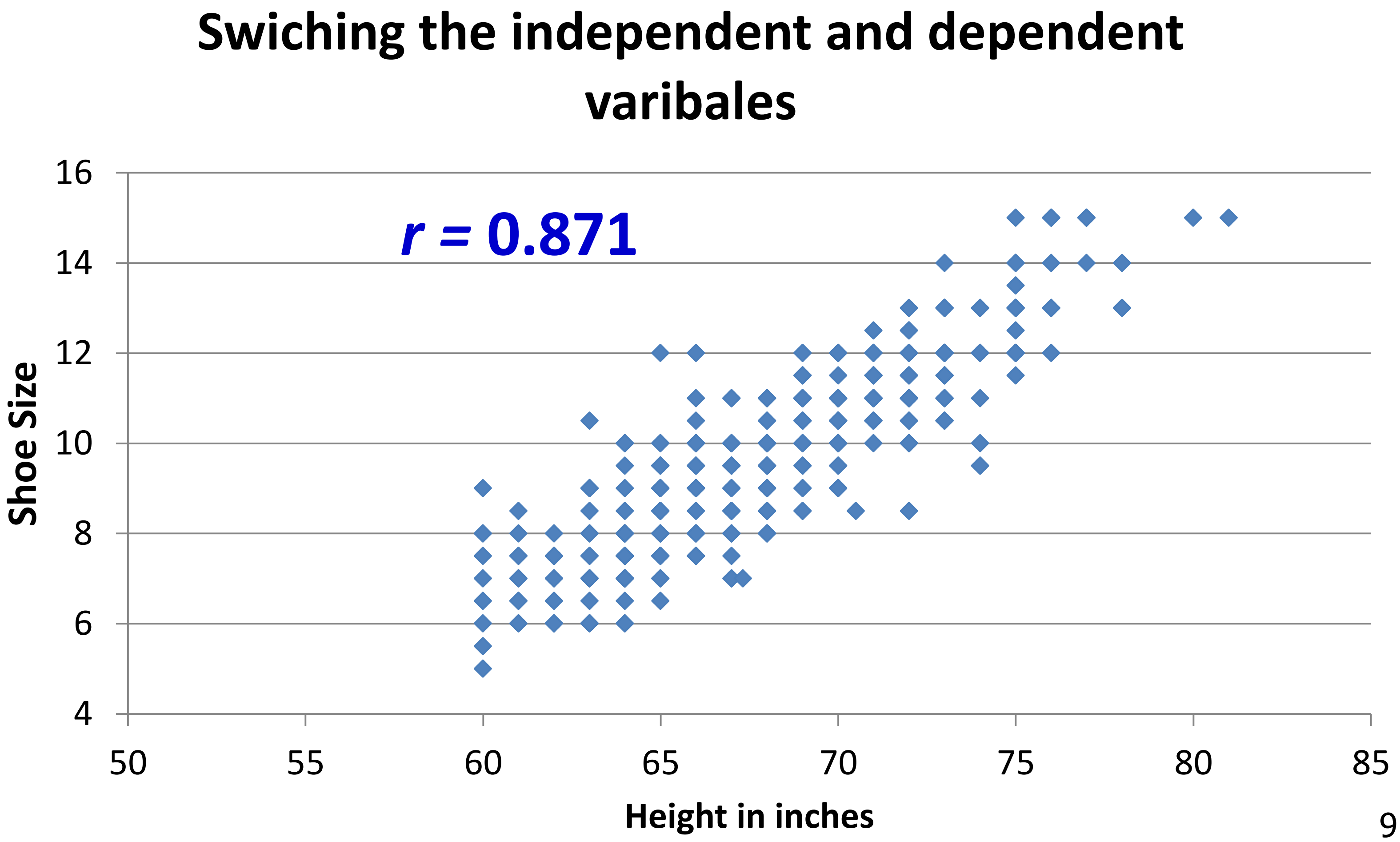
Data provided by Dr. Roger Woodard Department of Statistics NCSU

Example of Correlation

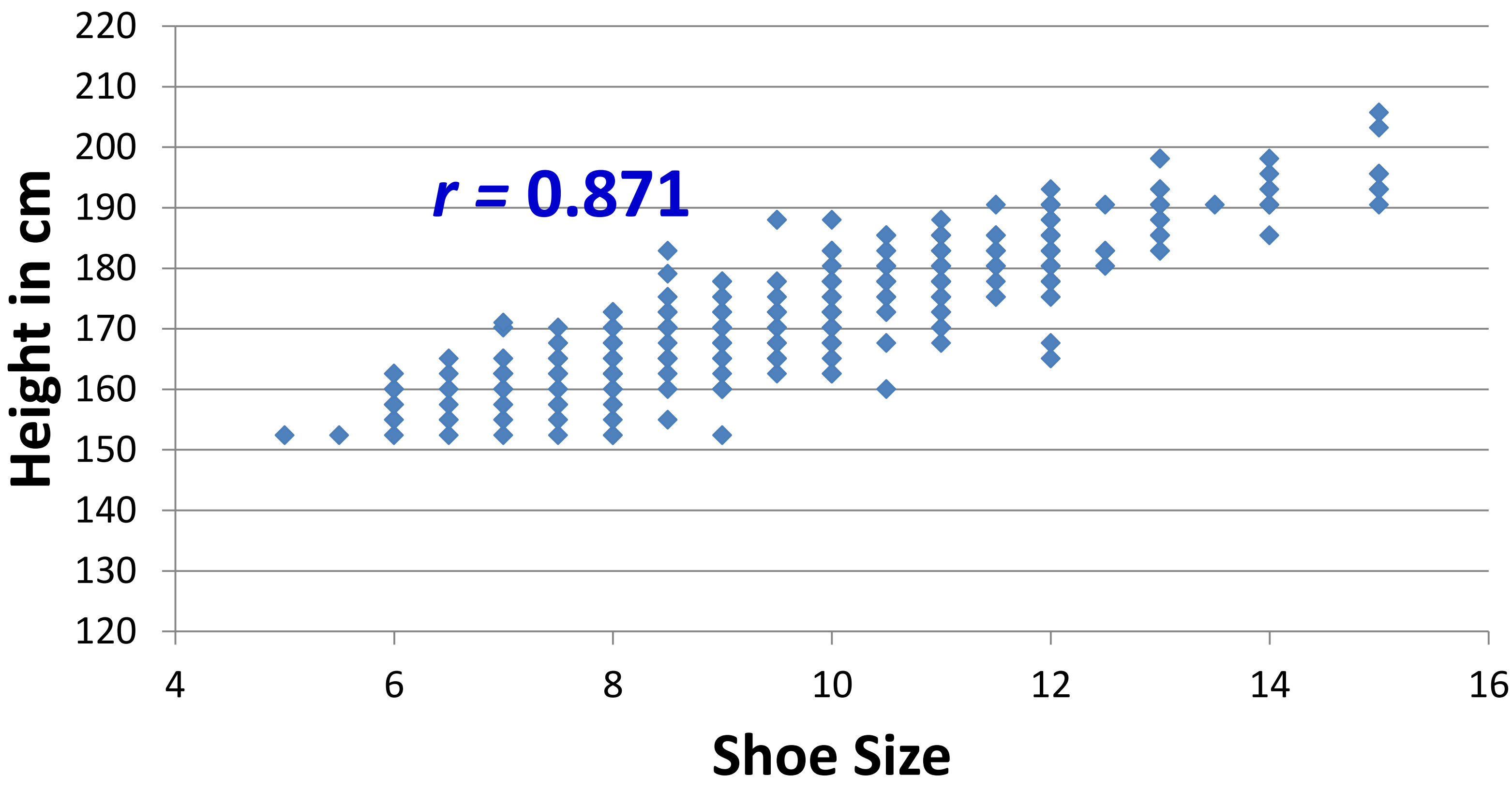


Using the height and shoe size to introduce correlation and regression

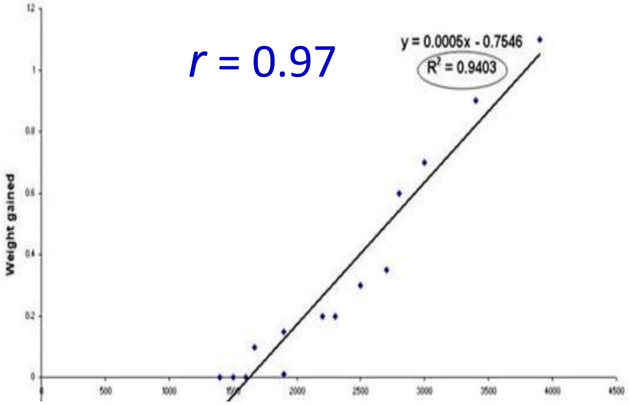
For Correlation Choice of X and Y does not Matter



Correlation is Independent of Units

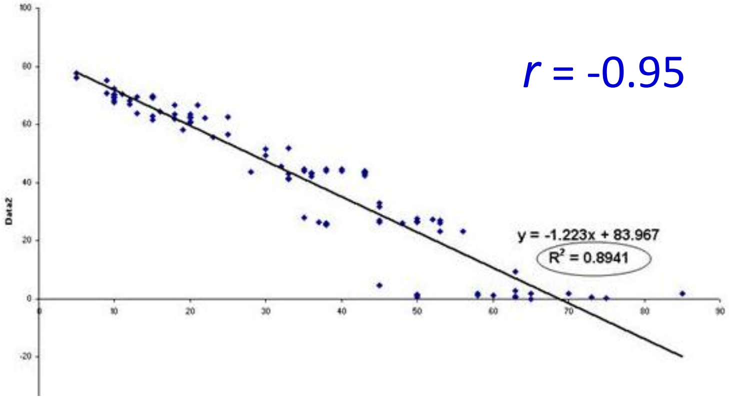


More Examples of Correlation



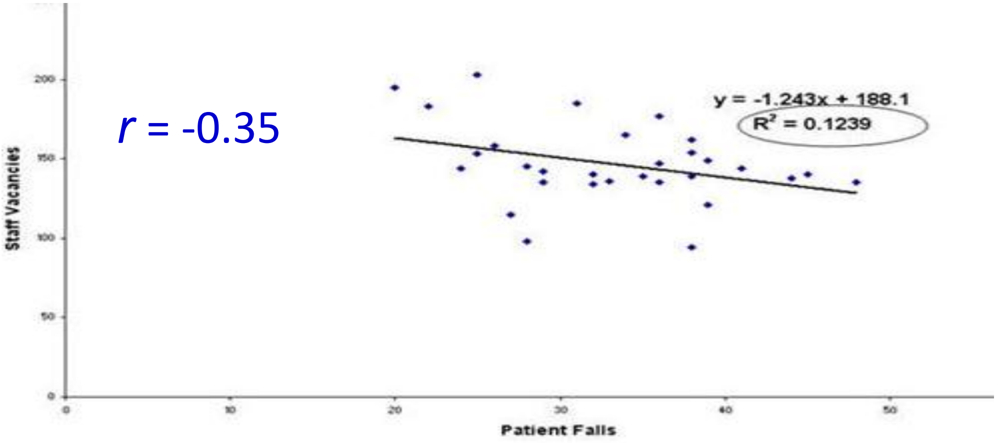
r = 0.97

Strong positive correlation



r = -0.95

Strong negative correlation



r = -0.35

Very week correlation

Correlation Coefficient *r*

- Also known as Pearson product moment correlation coefficient.
- Always ranges between -1 & 1.
- The value of *r* is independent of particular unit used.
- Correlation does not care about independent and dependent variables.

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2(y-\bar{y})^2}}$$

Rule of Thumb for Interpretation of *r*

- No relationship if *r* between 0 & ± 0.24.
- Weak relationship if *r* between 0.25 & 0.39 or -0.25 & -0.39
- Moderate if *r* between 0.40 &0.59 or -0.40 & -0.59.
- Strong if *r* between 0.60 &0.79 or -0.60 & -0.79.
- Excellent relationship if *r* greater than 0.80 or less than -0.80

Limitations of *r*

- Sensitive to outliers
- Sensitive to skewed data

Remedies

- Transform the data
- Use Spearman correlation coefficient
  - The assumption of normality is not required.
  - Can be used for outliers or ordered categorical such as pain scores

When We Should Not Use *r*

- There is a strong association but
  - Relationship is not linear.
  - Outliers are present in the data set that heavily influence the value of *r*.
- One of the variables is determined in advance.
- When the variables are measured over more than one distinct group exercise caution!

Test of Significance

- Could the observed correlation between two variables have arisen by chance alone?

H_0: ρ = 0

H_A: ρ ≠ 0

$$t = \frac{r}{SE(r)}, \quad t \text{ has } n\text{-}2 \text{ d.f}$$

- For n>10, can use Fisher’s z transformation

Assumptions for Significance

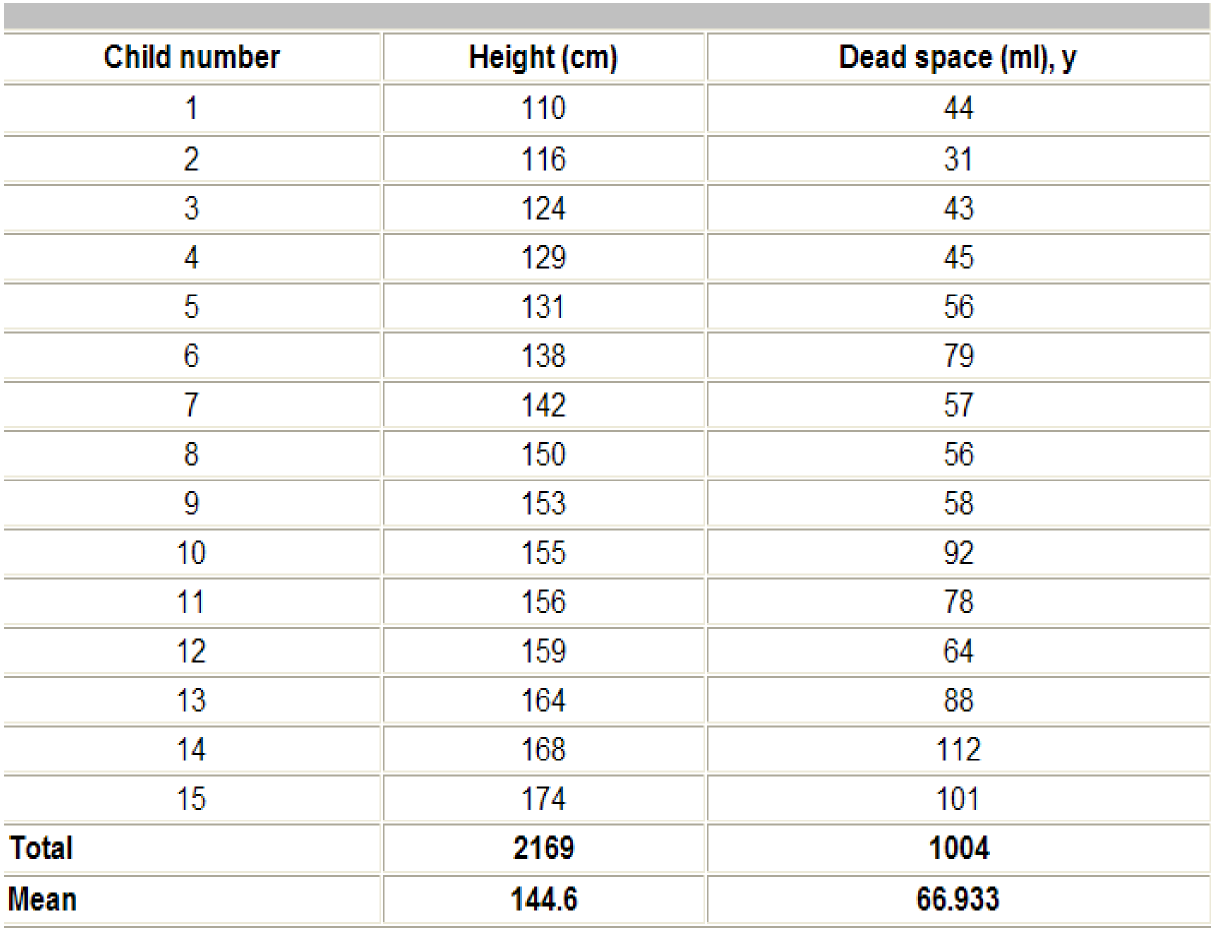
- Both variables are random samples.
- There is a linear relationship between variables.
- At least one has a normal distribution.
- The null hypothesis is that there is no relationship between variables.

Linear Regression

- For two variables *x* and *y* we assume that a change in *x* (independent) will lead directly to a change in *y* (depended).
- Often we are interested in predicting *y* from *x*.
- The equation $y = \alpha + \beta x$ is called regression equation. α is the intercept and β is the regression coefficient.

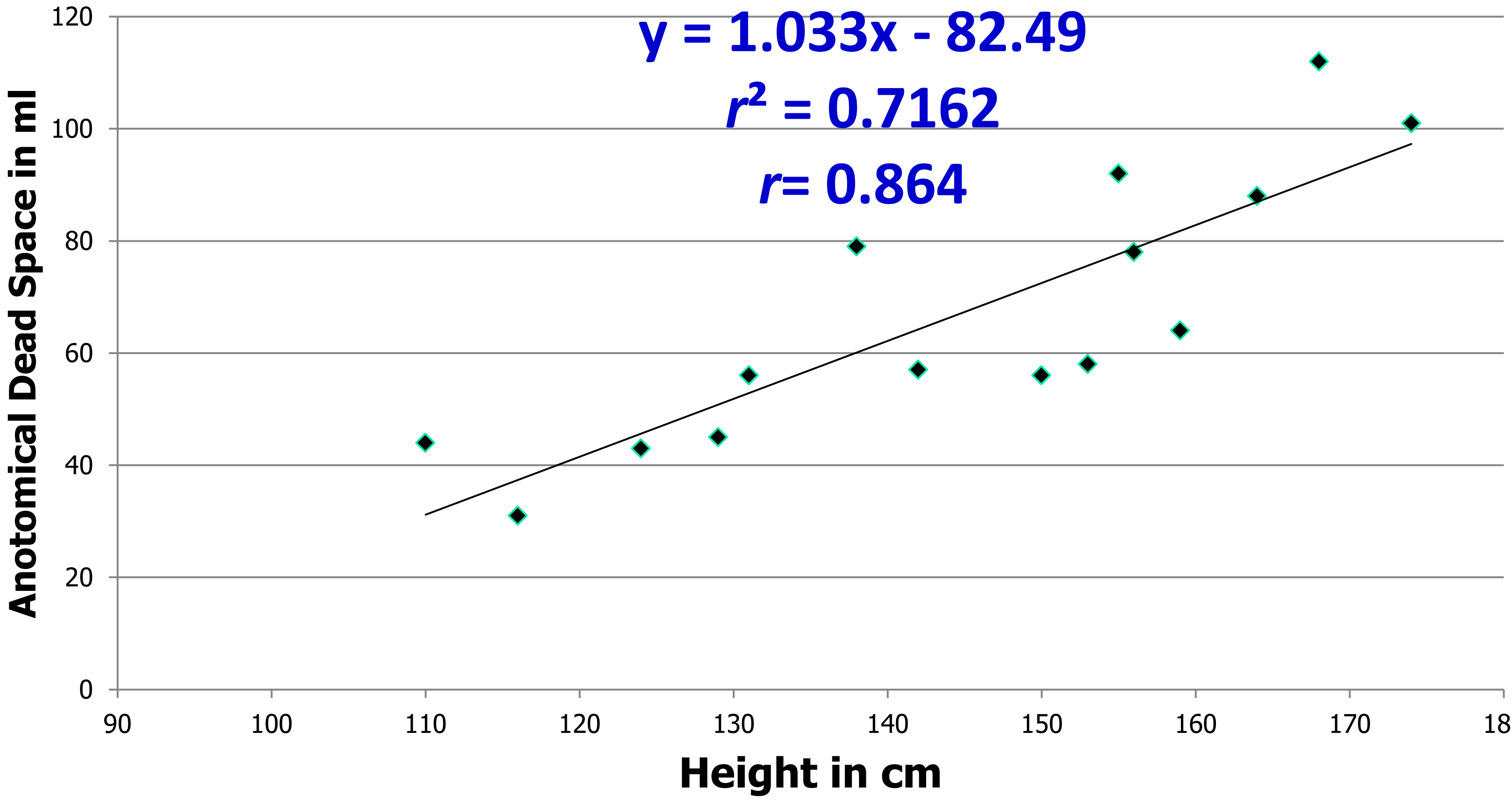
Example of Linear Regression

Can we predict pulmonary anatomical dead space from height?

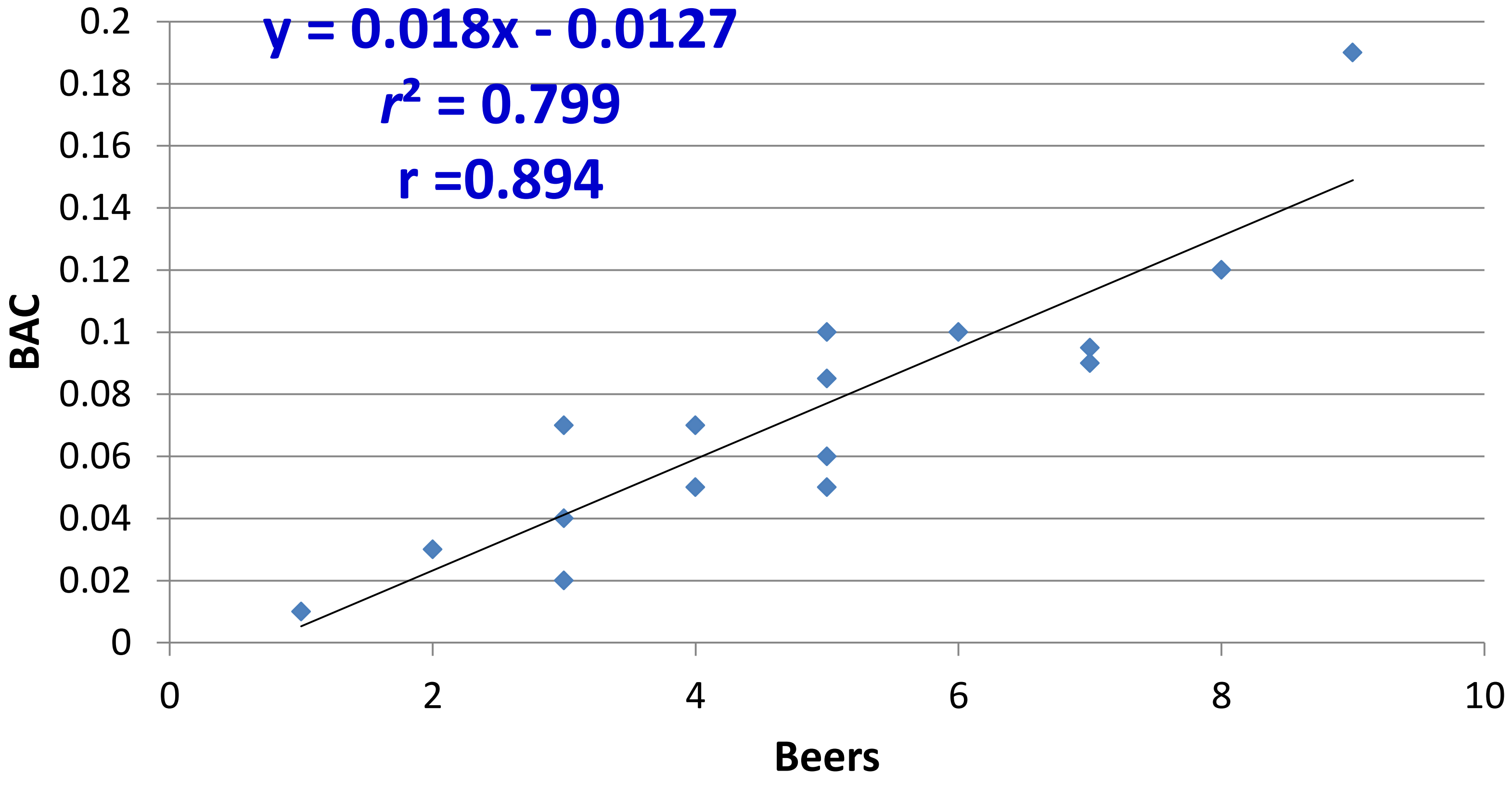


Child number	Height (cm)	Dead space (ml), y
1	110	44
2	116	31
3	124	43
4	129	45
5	131	56
6	138	79
7	142	57
8	150	56
9	153	58
10	155	92
11	156	78
12	159	64
13	164	88
14	168	112
15	174	101
Total	**2169**	**1004**
Mean	**144.6**	**66.933**

Predicting Dead Space from Height



Predicting BAC from Beers consumed

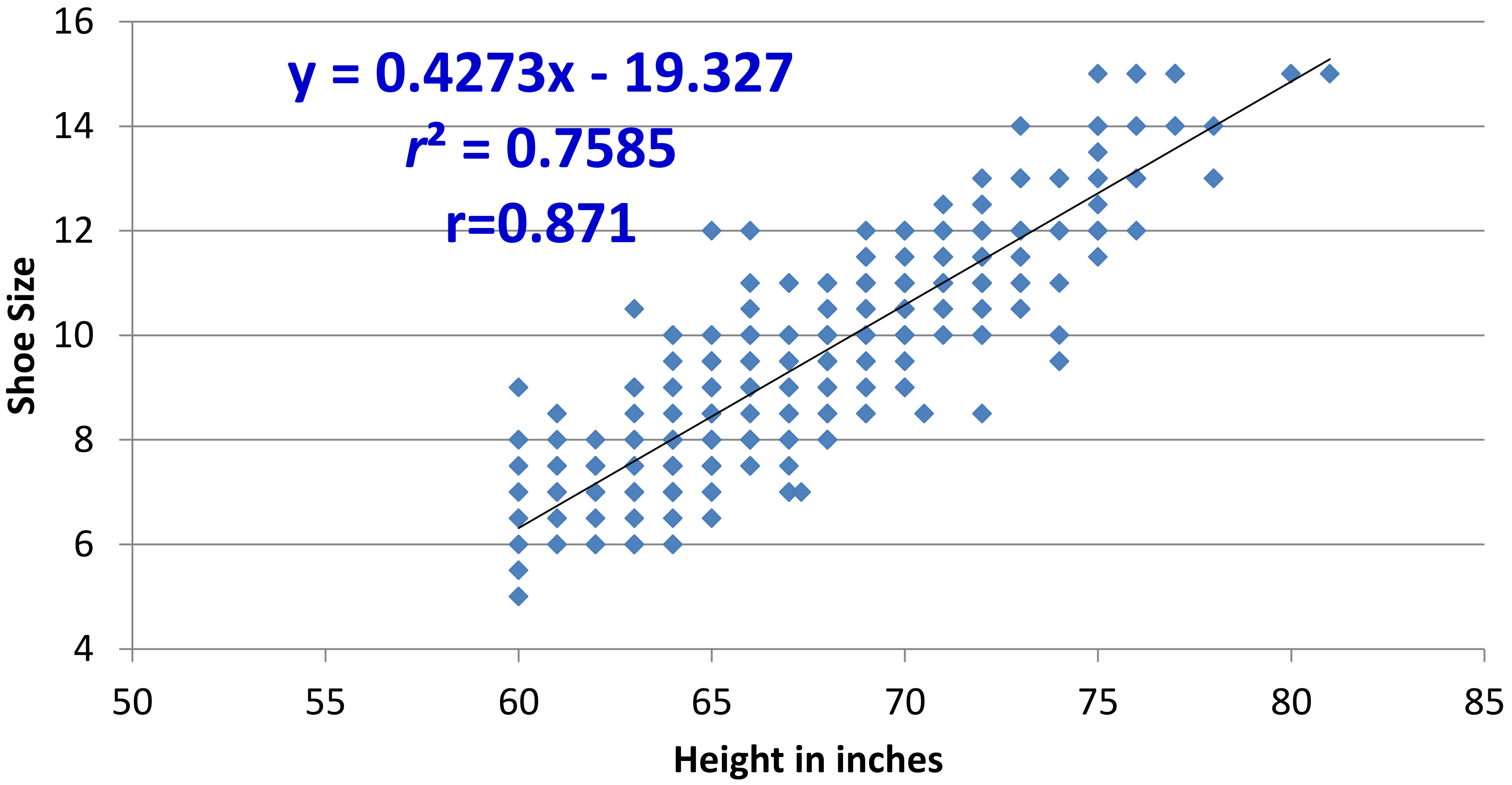


$y = 0.018x - 0.0127$

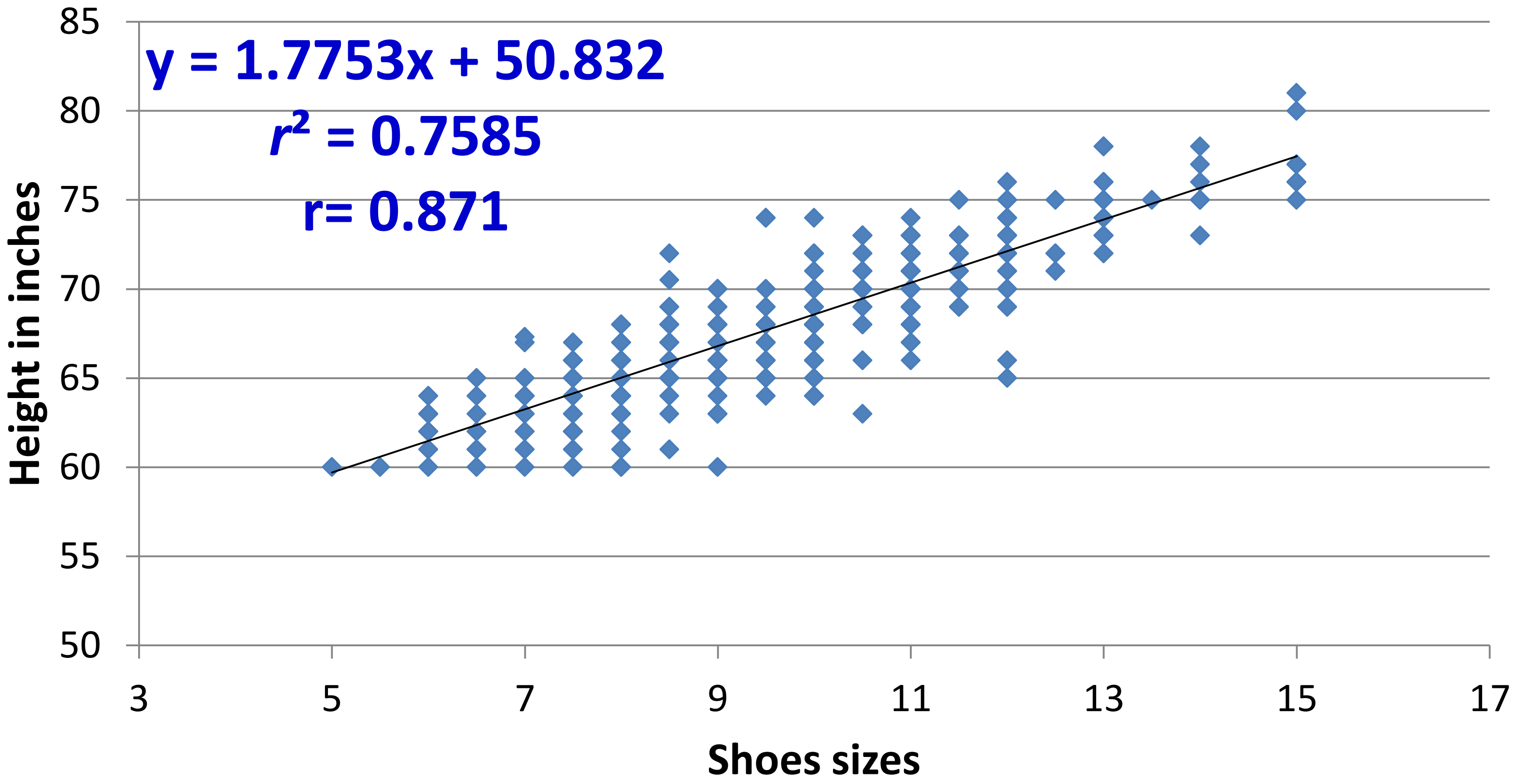
$r^2 = 0.799$

$r = 0.894$

Predicting Shoe Size from Height



Switching Independent and Depending Variables



Least Square Estimates of Population Parameters

- We require estimates of α & β from sample. We can write regression equation for *ith* observation pair as $y_i = a + bx_i$
- We want to choose *a* and *b* to minimize the sum of squares of errors $\Sigma(y_i - y)^2$

Calculating *a* and *b*

For regressionequation $y = a + bx$

$$b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

Coefficient of Determination r^2

- The square of correlation coefficient.
- Always positive and between 0 & 1.
- The coefficient of determination gives the proportion of the variance (fluctuation) of one variable (y) that is predictable from the other variable (x). It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph.
- The coefficient of determination is a measure of how well the regression line represents the data.

Test of significance

- H_0: β = 0
- *t* =*b*/SE(*b*) with n-2 d.f.
- Tests of significance for a correlation and a regression both produce the same *t* statistic and same *p* value; even though the assumptions for both are different.

Assumptions for Significance of Regression

- The relationship is approximately linear.
- The prediction error is unrelated to the predicted value.
- The residuals (errors) are normally distributed about the fitted line.
- The residuals are independent of each other.

Difference between Regression and Correlation

- Correlation does not dependent on the units of measure but the regression does.
- For regression is important which variable is X and which is Y, for correlation it is not.
- Correlation and regression are related.

$$r = b\frac{s_x}{s_y}$$

Things to Remember

- When two variables are correlated, they may not be casually related.
  - Example: Reading scores and shoe sizes in US
- If just interested in strength of relationship, use *r.*
- When there is clear causation use regression and report *r or* r^2 also.

Points when Reading a Paper

- When *r* is quoted, is the relationship likely to be linear.
- If a significant correlation is obtained and the causation inferred , could there be a third factor responsible for the association?
- If predictions are given, are they made from within the range of the observed values of the independent variables?

More Advanced Techniques

- Multiple Correlation for one continuous dependent variable and many dependent variables
  - Independent variables can be continuous or binary
- Logistic Regression for binary dependent variable
  - Categorical or continuous independent variables.

Thank you!

Questions/Comments

Rizwana.Rehman@va.gov

(919) 286-0411 ext: 5024

For more information, program materials, and to complete evaluation for CME credit visit

www.epilepsy.va.gov/Statistics