Biostatistics

Chapter 7B Correlation and Regression

Jing Li

jing.li@sjtu.edu.cn

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

Dept of Bioinformatics & Biostatistics, SJTU

Review Questions (5 min)

- Describe the assumption of simple linear correction ?
- Write down the teachers' names in the last two classes.

Pearson's Correlation Coefficient

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

Linear Regression

- Variables: $y = \beta_0 + \beta_1 x + \varepsilon$

  X = Independent Variable (we provide this)

  Y = Dependent Variable (we observe this)

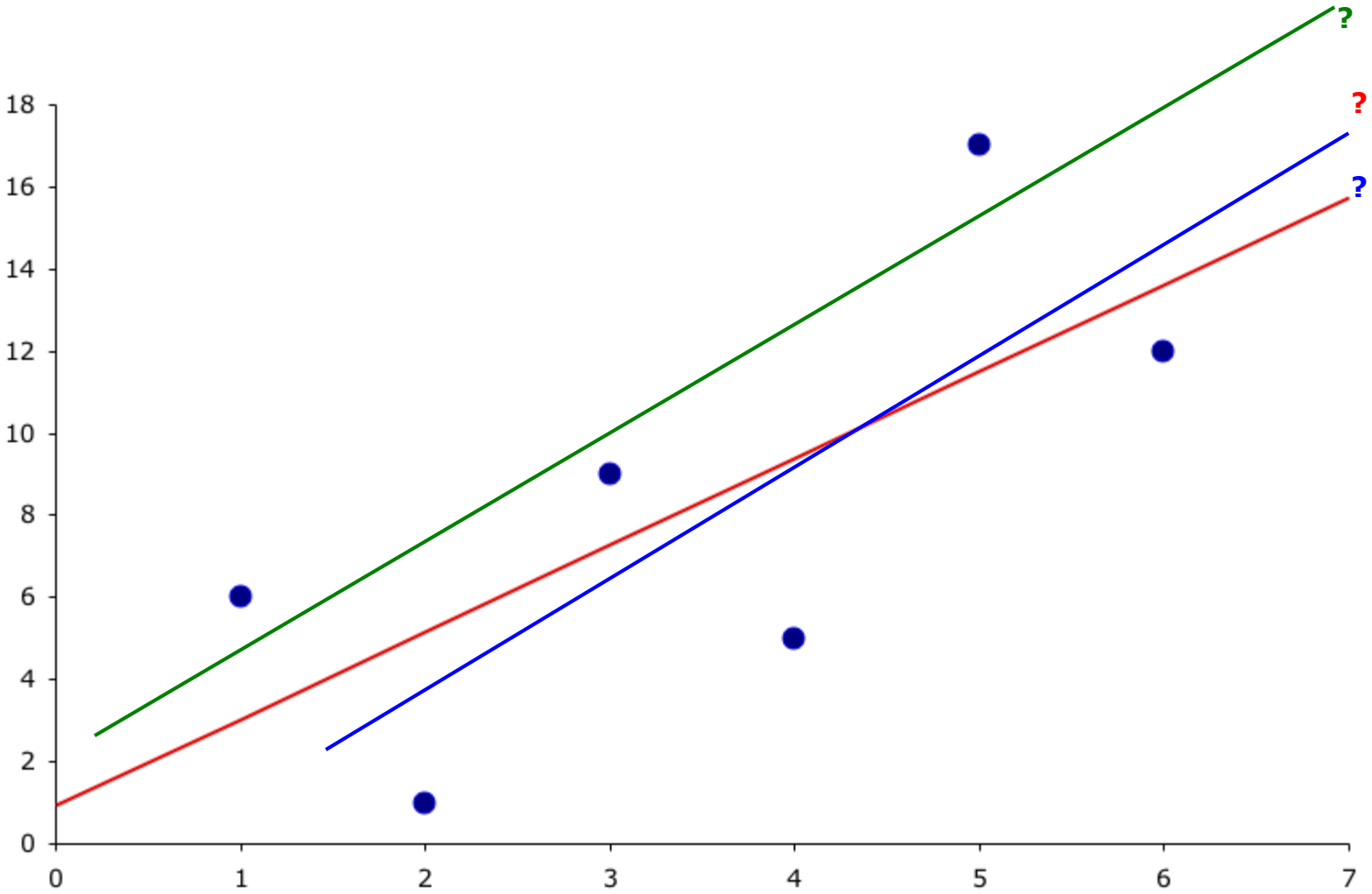
- Parameters:

  β_0 = Y-Intercept

  β_1 = Slope

  $\varepsilon \sim$ Normal Random Variable ($\mu_\varepsilon = 0, \sigma_\varepsilon$ = ???) [Noise]

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



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:

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

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

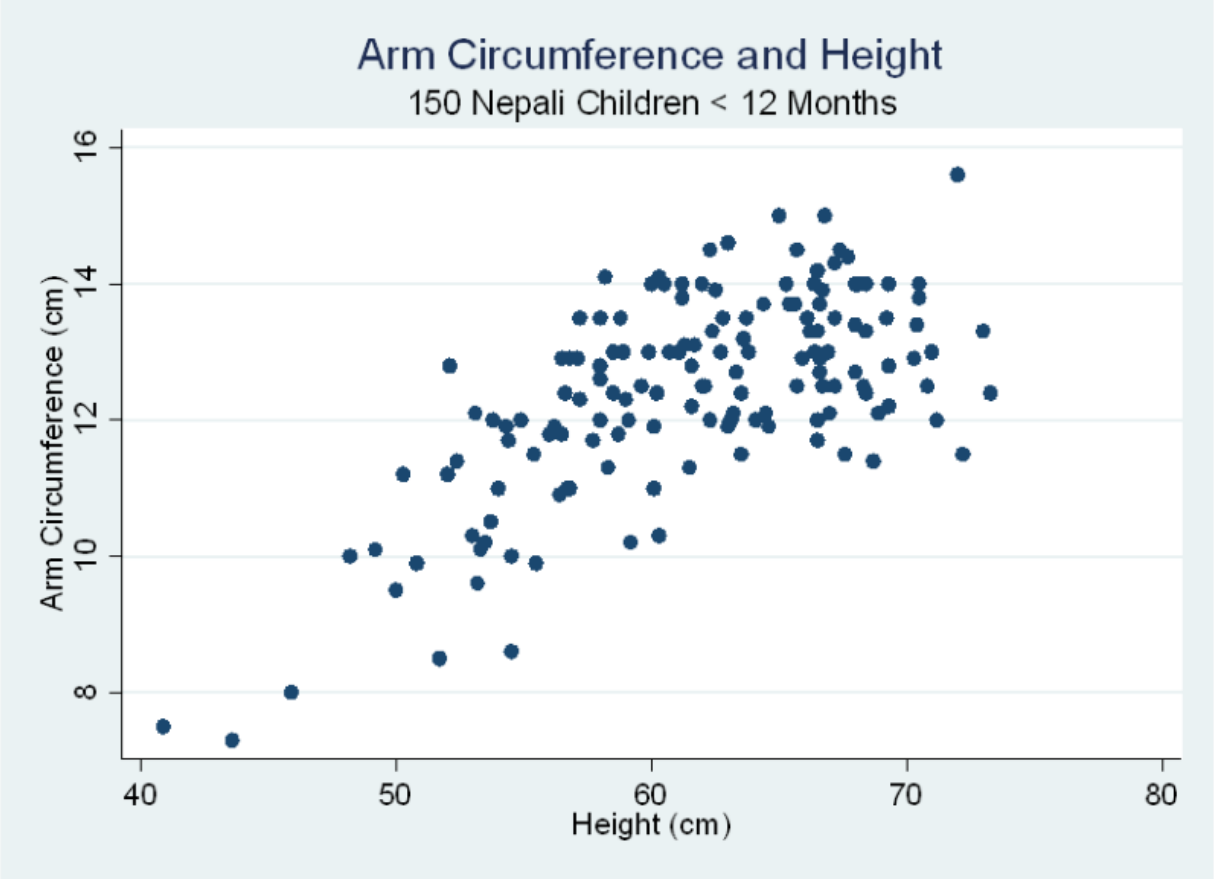
$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

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

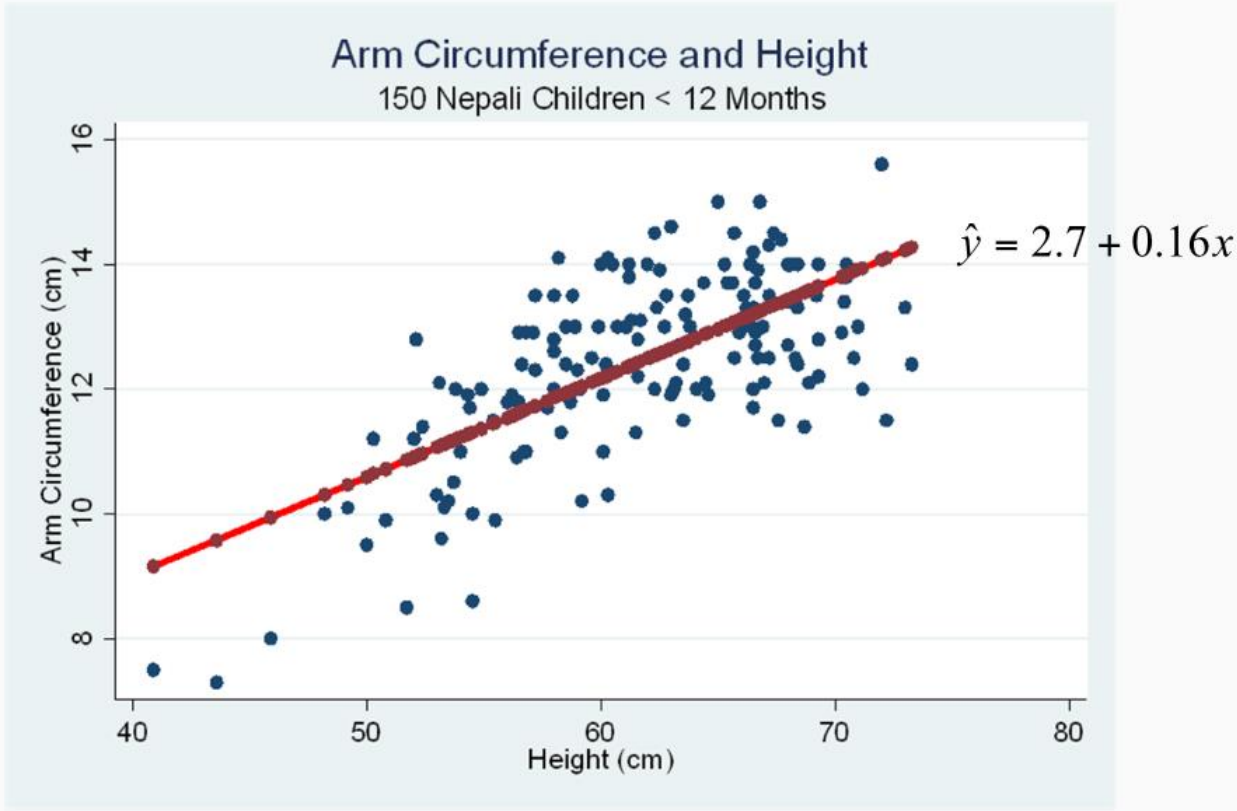
$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

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

Visualizing Arm Circumference and Height Relationship

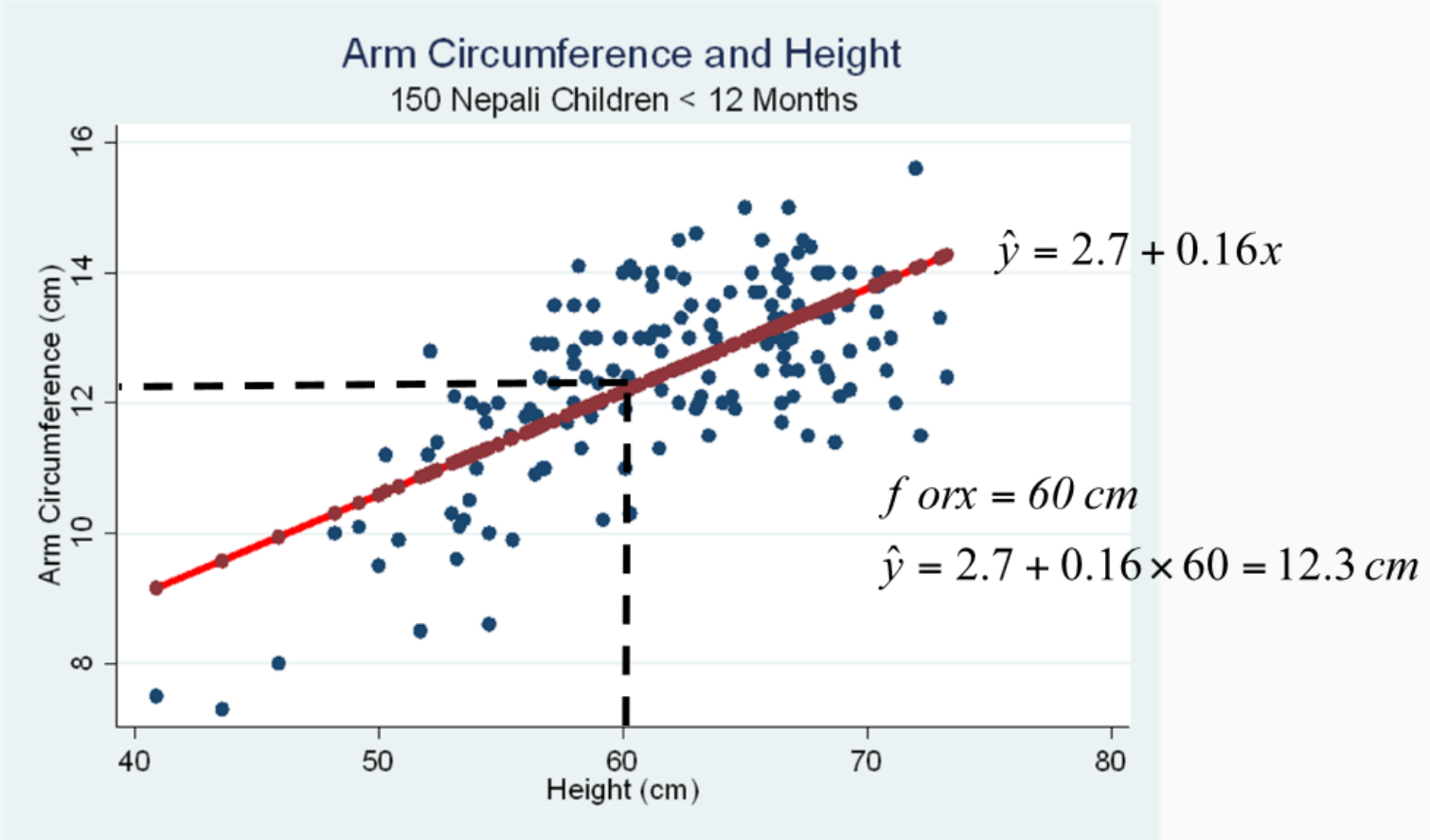


Scatterplot with regression line



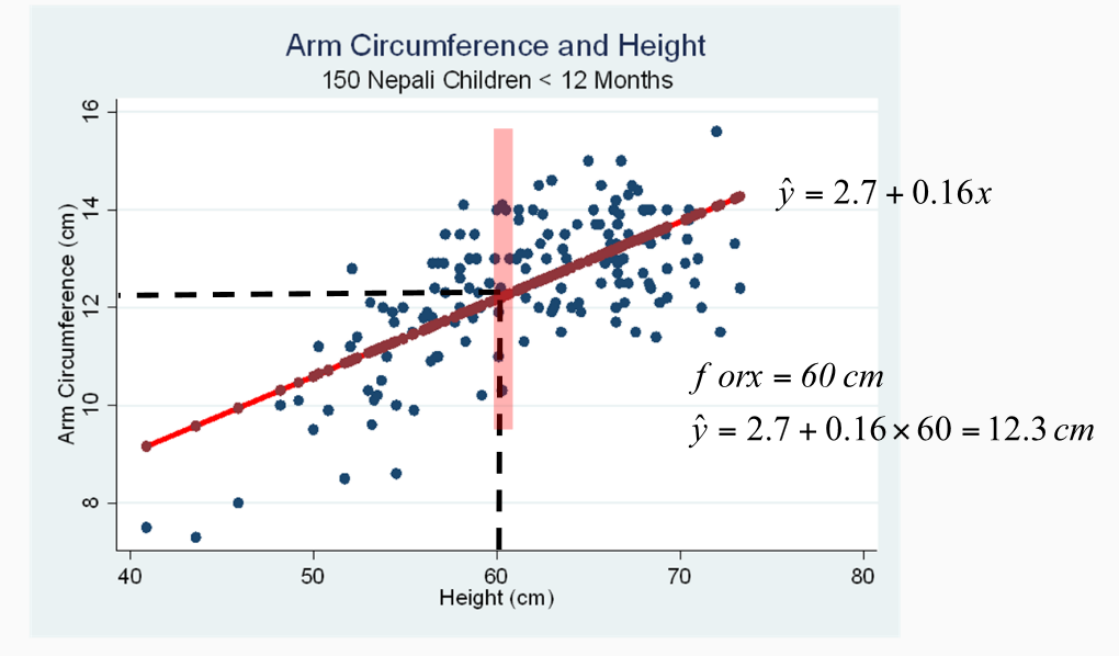
Example: Arm Circumference and Height

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



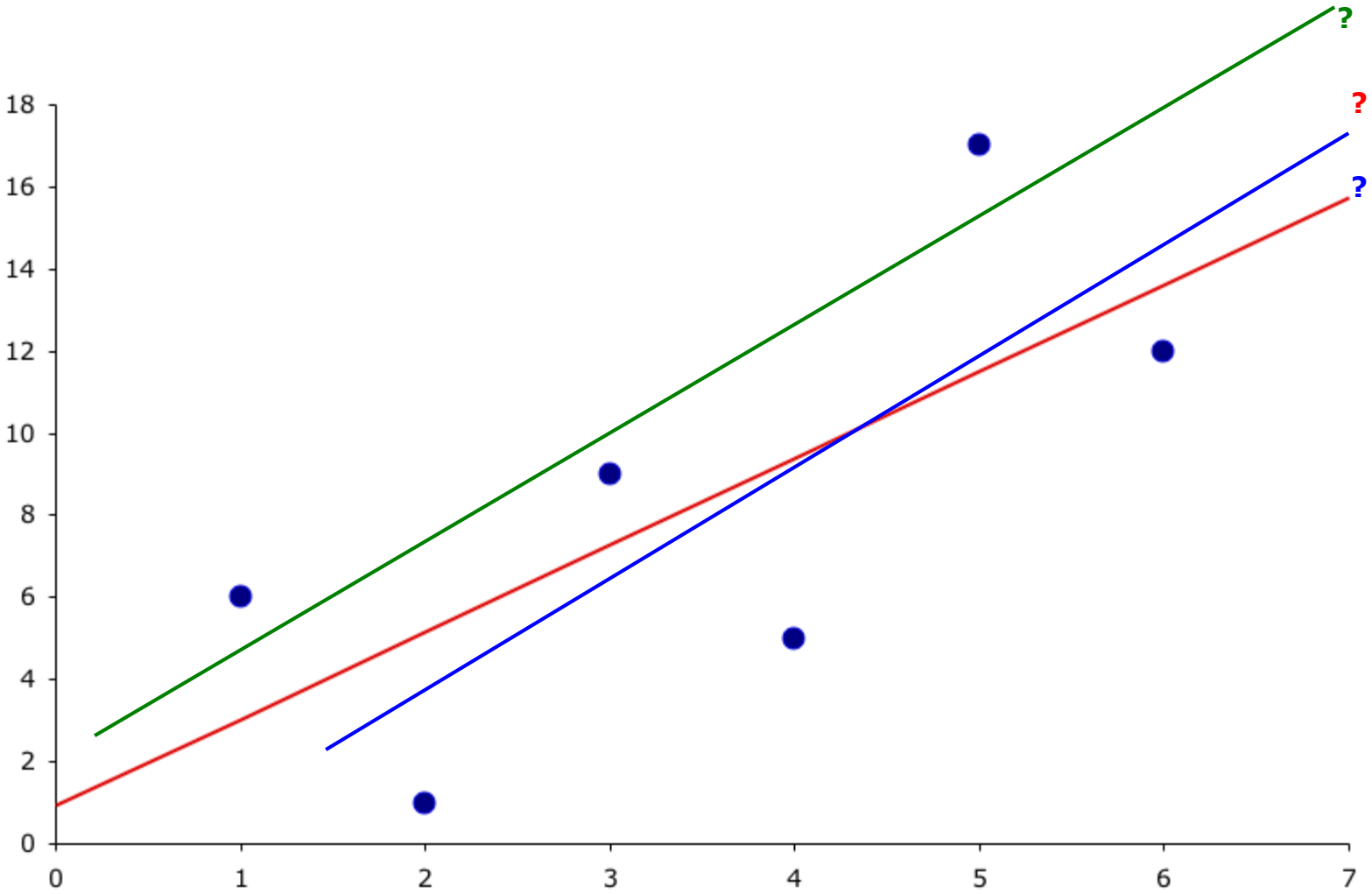
Example: Arm Circumference and Height

- **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

The best "fit" is good enough?



Assessing the Model…

- The least squares method will always produce a straight line, even if there is no relationship between the variables, or if the relationship is something other than linear.

- Hence, in addition to determining the coefficients of the least squares line, we need to assess it to see how well it "fits" the data. We'll see these evaluation methods now. They're based on the what is called sum of squares for errors (SSE).

Sum of Squares for Error (SSE – another thing to calculate)…

- The sum of squares for error is calculated as:

$$SSE = (n-1)\left(s_y^2 - \frac{s_{xy}^2}{s_x^2}\right)$$

and is used in the calculation of the ***standard error of estimate***:

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$$

- If s_ε is zero, all the points fall on the regression line.

Standard Error…

- If s_ε is small, the fit is excellent and the linear model should be used for forecasting. If s_ε is large, the model is poor…

But what's the cutoff of s_ε for a good model ?

Standard Error…

- Judge the value of s_ε by comparing it to the sample mean of the dependent variable ($\bar{y}$).

For example,

- s_ε = .3265 and
- $\bar{y}$ = 14.841

- so (relatively speaking) it appears to be "small", hence our linear regression model of car price as a function of odometer reading is "good".

Testing the Slope…

- If no linear relationship exists between the two variables, we would expect the regression line to be ***horizontal***, that is, to have a ***slope of zero***.

- We want to see if there is a linear relationship, i.e. we want to see if the slope (β_1) is something other than zero. Our research hypothesis becomes:

  H_1: $\neq 0$ β_1

- Thus the null hypothesis becomes:

  H_0: $= 0$ β_1

Testing the Slope…

- We can implement this test statistic to try our hypotheses:
- H0: $\beta_1 = 0$

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

- where s_{b_1} is the standard deviation of b_1, defined as:

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1)s_x^2}}$$

- If the error variable is normally distributed, the test statistic has a Student **t**-distribution with n–2 degrees of freedom. The rejection region depends on whether or not we're doing a one- or two- tail test (two-tail test is most typical).

Relationship with correlation

$$\hat{r} = \hat{\beta} \frac{SD_x}{SD_y}$$

In correlation, the two variables are treated as equals. In regression, one variable is considered independent (=predictor) variable (*X*) and the other the dependent (=outcome) variable *Y*.

Power of the model : Coefficient of Determination…

- Tests thus far have shown if a linear relationship ***exists***; it is also useful to measure the ***strength of the relationship***. This is done by calculating the ***coefficient of determination*** – R^2.

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}, \text{ or } R^2 = 1 - \frac{SSE}{\sum (y_i - \bar{y})^2}$$

- The coefficient of determination is the square of the coefficient of correlation (**r**), hence $R^2 = (r)^2$

R^2 has a value of .6483. This means 64.83% of the variation in y is explained by your regression model. The remaining 35.17% is unexplained, i.e. due to error.

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

Regression Diagnostics…

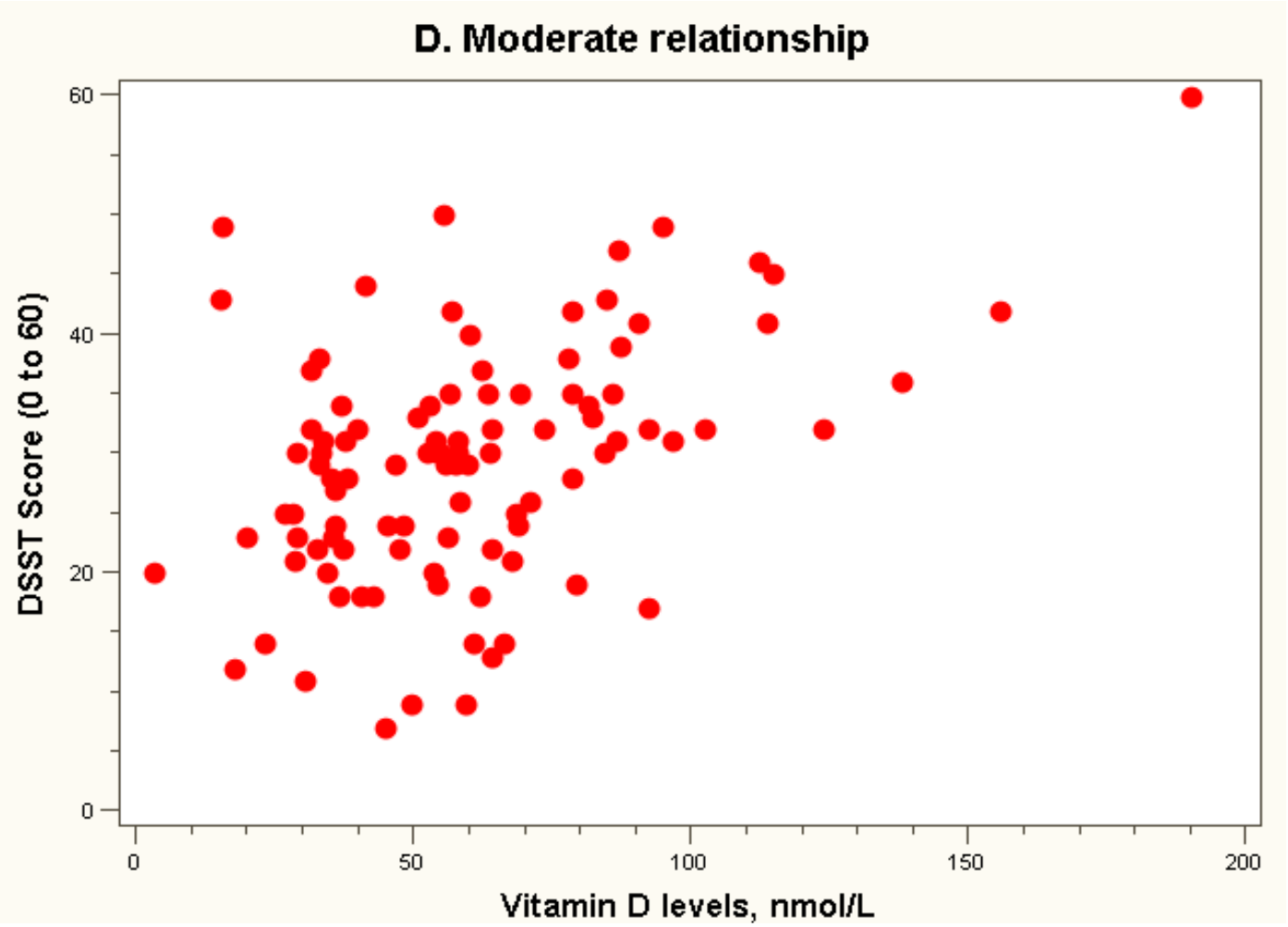
How can we **diagnose** violations of these conditions?

➔ **Residual Analysis**, that is, examine the ***differences*** between the actual data points and those predicted by the linear equation…

There are three conditions about error that are required in order to perform a regression analysis. These are:

- **The error variable must be normally distributed,**
- **The error variable must have a constant variance**
- **The errors must be independent of each other.**

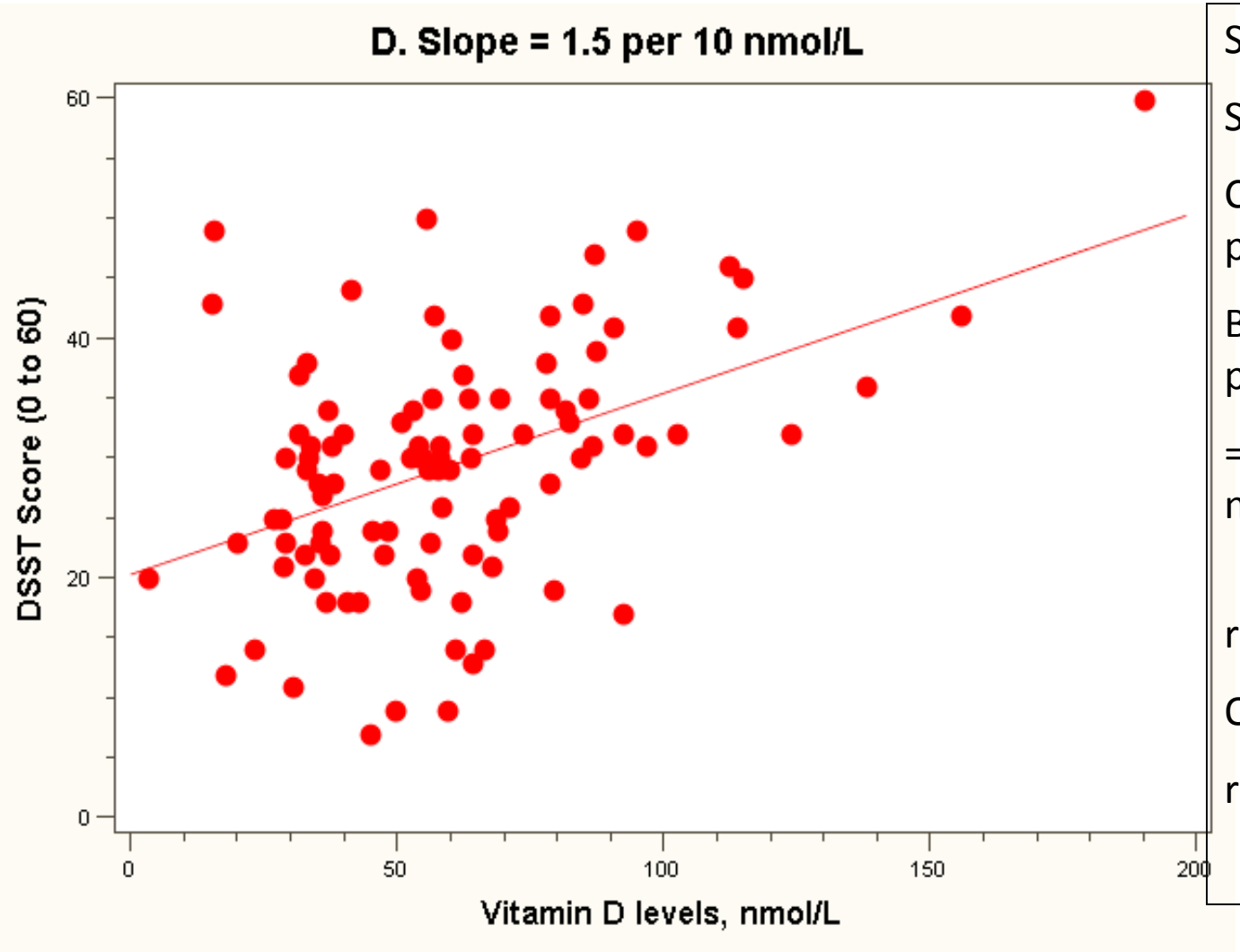
Example:
relationship between cognitive function and vitamin D



Cross-sectional study of 100 middle-aged and older European men.

Cognitive function is measured by the Digit Symbol Substitution Test (DSST).

Lee DM, Tajar A, Ulubaev A, et al. Association between 25-hydroxyvitamin D levels and cognitive performance in middle-aged and older European men. J Neurol Neurosurg Psychiatry. 2009 Jul;80(7):722-9.



SDx = 33 nmol/L

SDy= 10 points

Cov(X,Y) = 163 points*nmol/L

Beta = $163/33^2 = 0.15$ points per nmol/L

= 1.5 points per 10 nmol/L

r = 163/(10*33) = 0.49

Or

r = 0.15 * (33/10) = 0.49

Significance testing…

Slope

Distribution of slope ~ $T_{n-2}(\beta, s.e.(\hat{\beta}))$

H0: $\beta 1 = 0$ (no linear relationship)

H1: $\beta 1 \neq 0$ (linear relationship does exist)

$$T_{n-2} = \frac{\hat{\beta} - 0}{s.e.(\hat{\beta})}$$

Formula for the standard error of beta (you will not have to calculate by hand!):

$$s_{\hat{\beta}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{s_{y/x}^2}{SS_x}}$$

$$\text{where } SS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\text{and } \hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$$

Example: dataset 4

- Standard error (beta) = 0.03
- T_{98} = 0.15/0.03 = 5, p<.0001

- 95% Confidence interval = 0.09 to 0.21

Residual Analysis: check assumptions

$$e_i = Y_i - \hat{Y}_i$$

- The residual for observation i, e_i, is the difference between its observed and predicted value
- Check the assumptions of regression by examining the residuals
  - **Examine for linearity assumption**
  - **Examine for constant variance for all levels of X (homoscedasticity)**
  - **Evaluate normal distribution assumption**
  - **Evaluate independence assumption**
- Graphical Analysis of Residuals
  - **Can plot residuals vs. X**

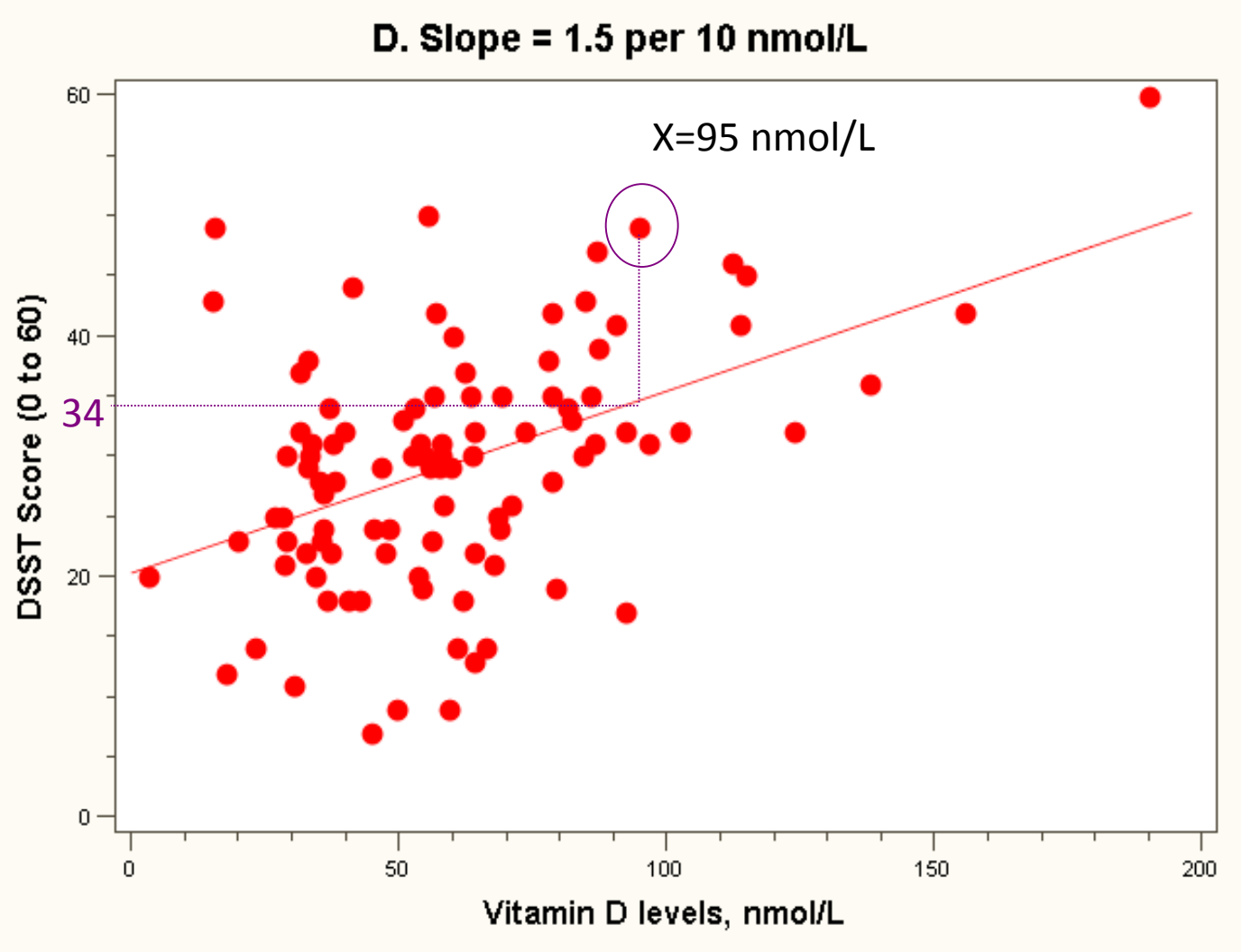
Predicted values…

$$\hat{y}_i = 20 + 1.5x_i$$

For Vitamin D = 95 nmol/L (or 9.5 in 10 nmol/L):

$$\hat{y}_i = 20 + 1.5(9.5) = 34$$

Residual = observed - predicted

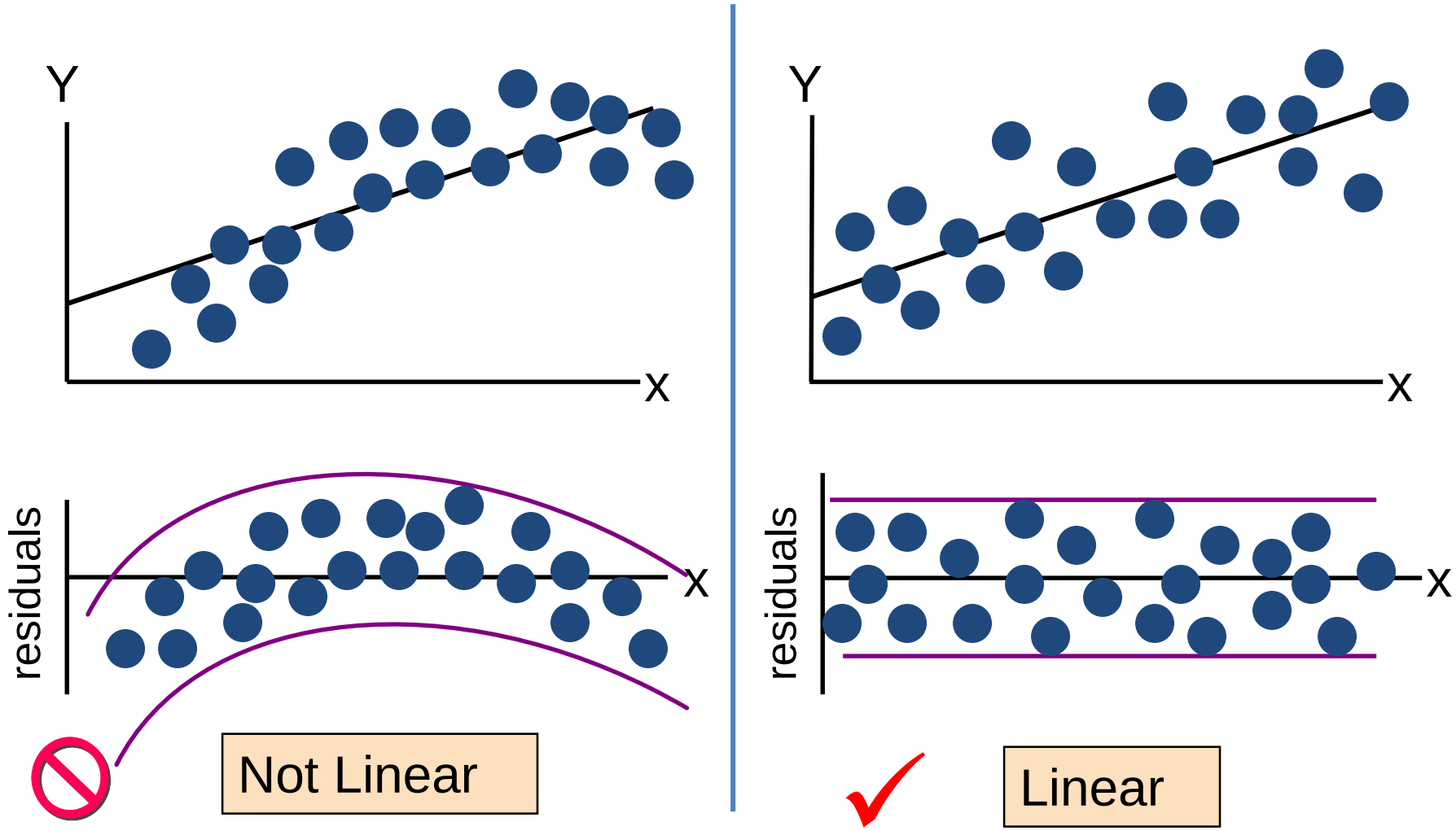


$$y_i = 48$$

$$\hat{y}_i = 34$$

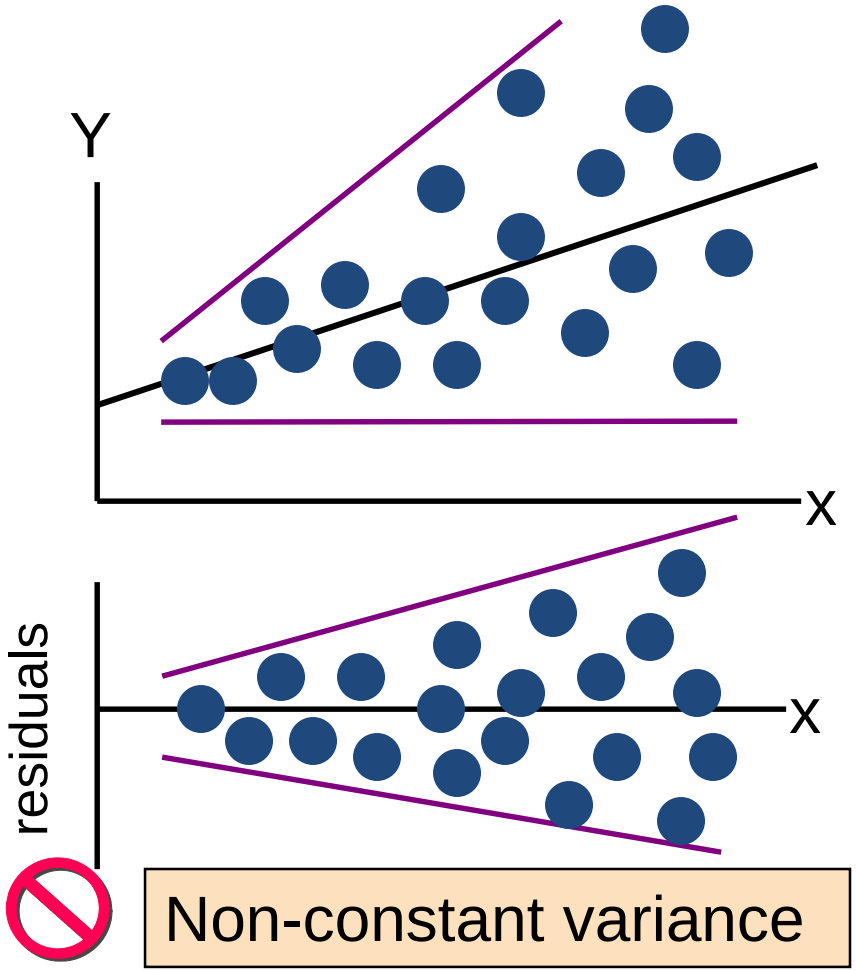
$$y_i - \hat{y}_i = 14$$

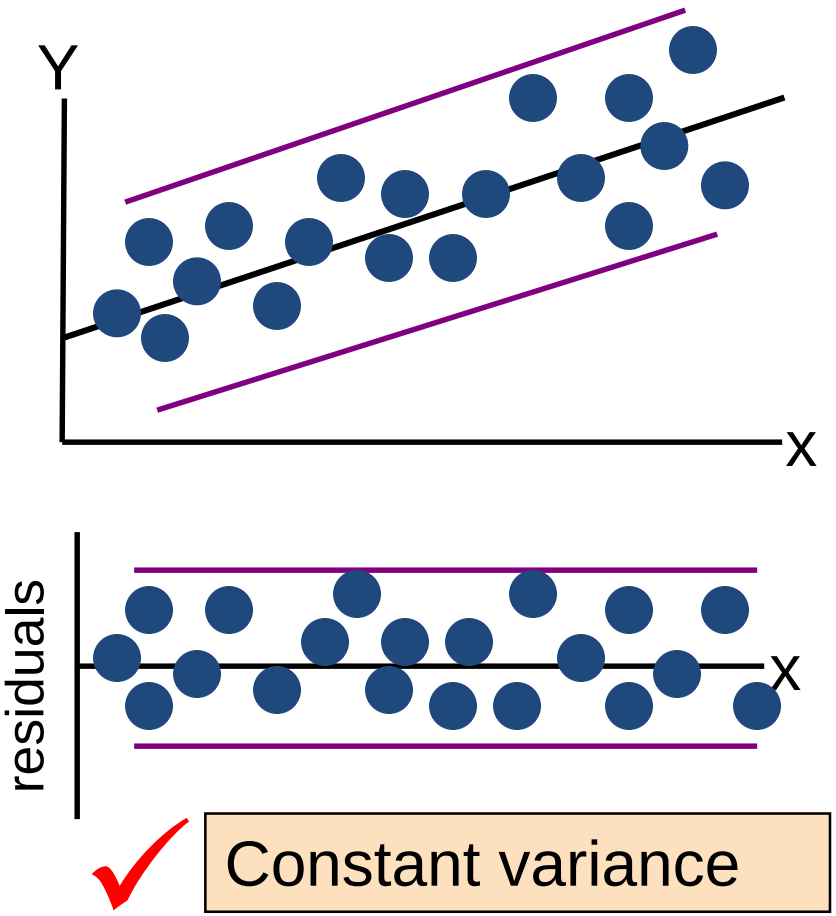
(1) Residual Analysis for Linearity



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

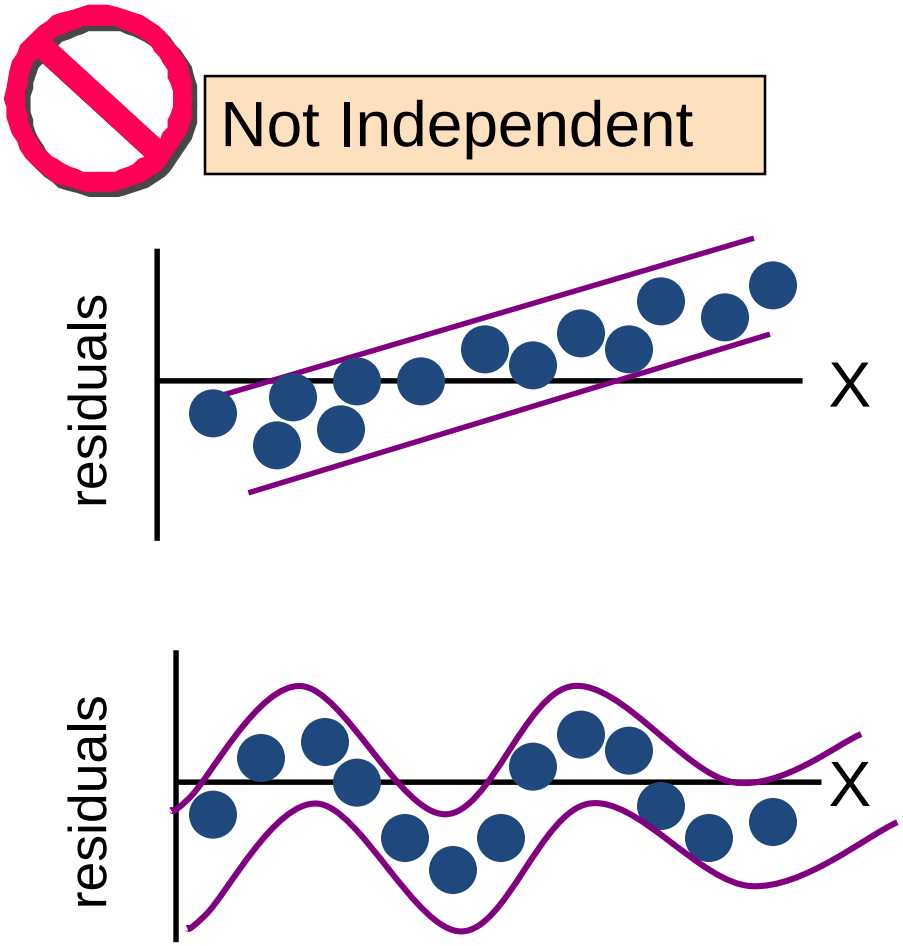
(2) Residual Analysis for Homoscedasticity

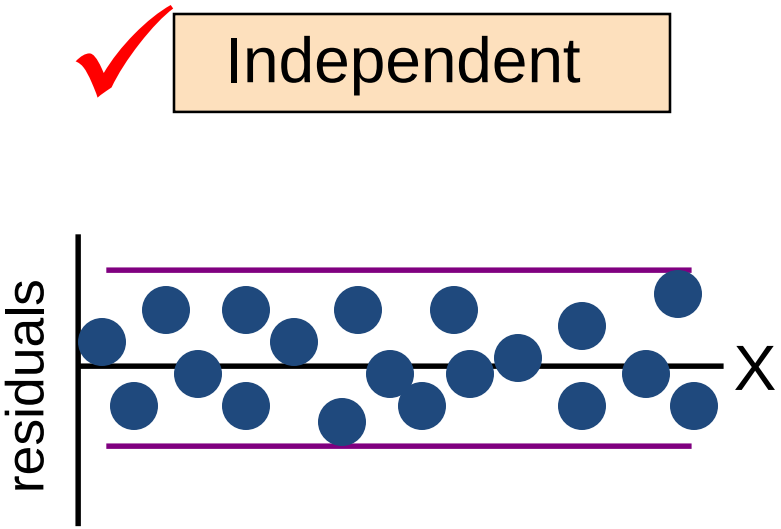




- Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

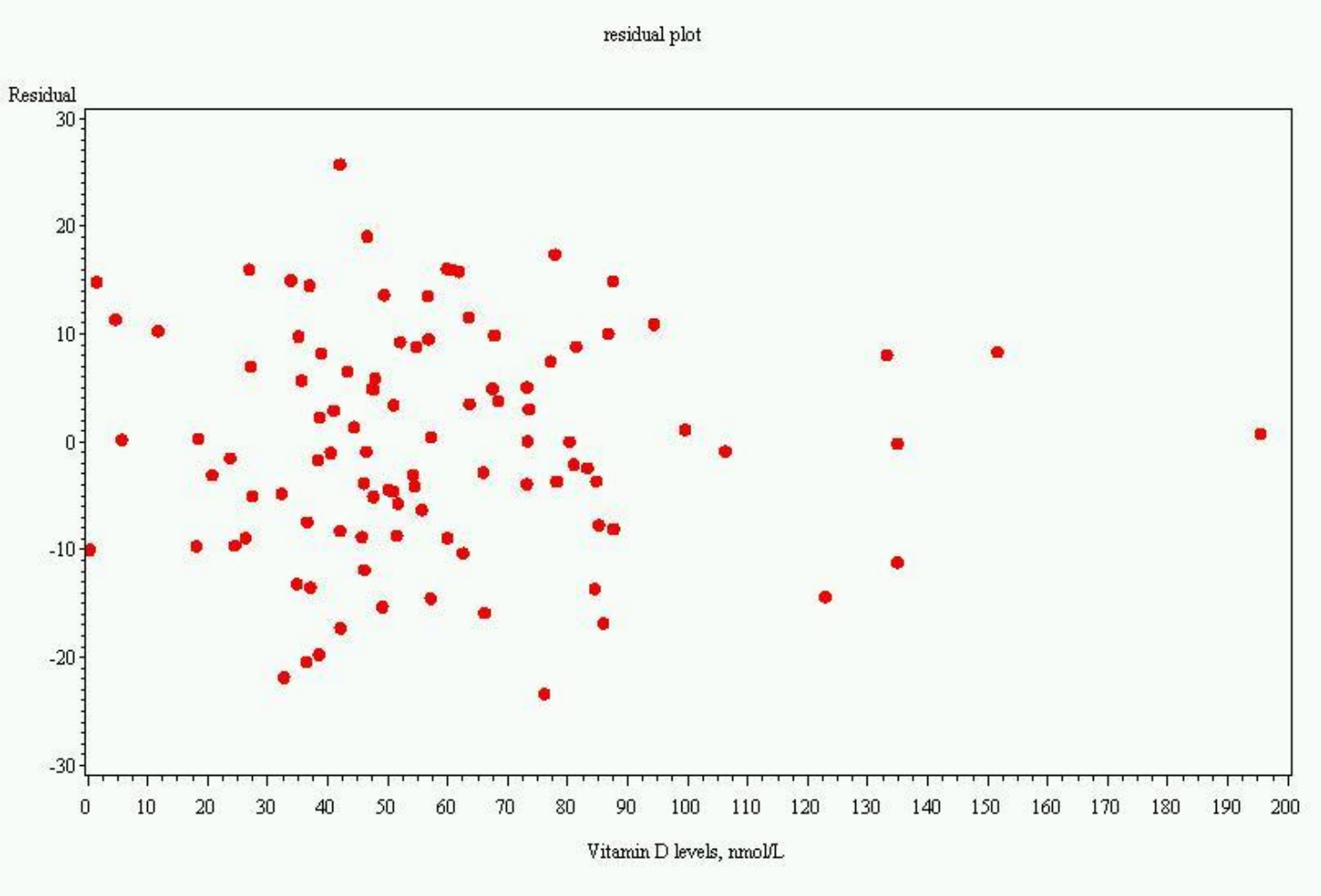
(3) Residual Analysis for Independence





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

Residual plot



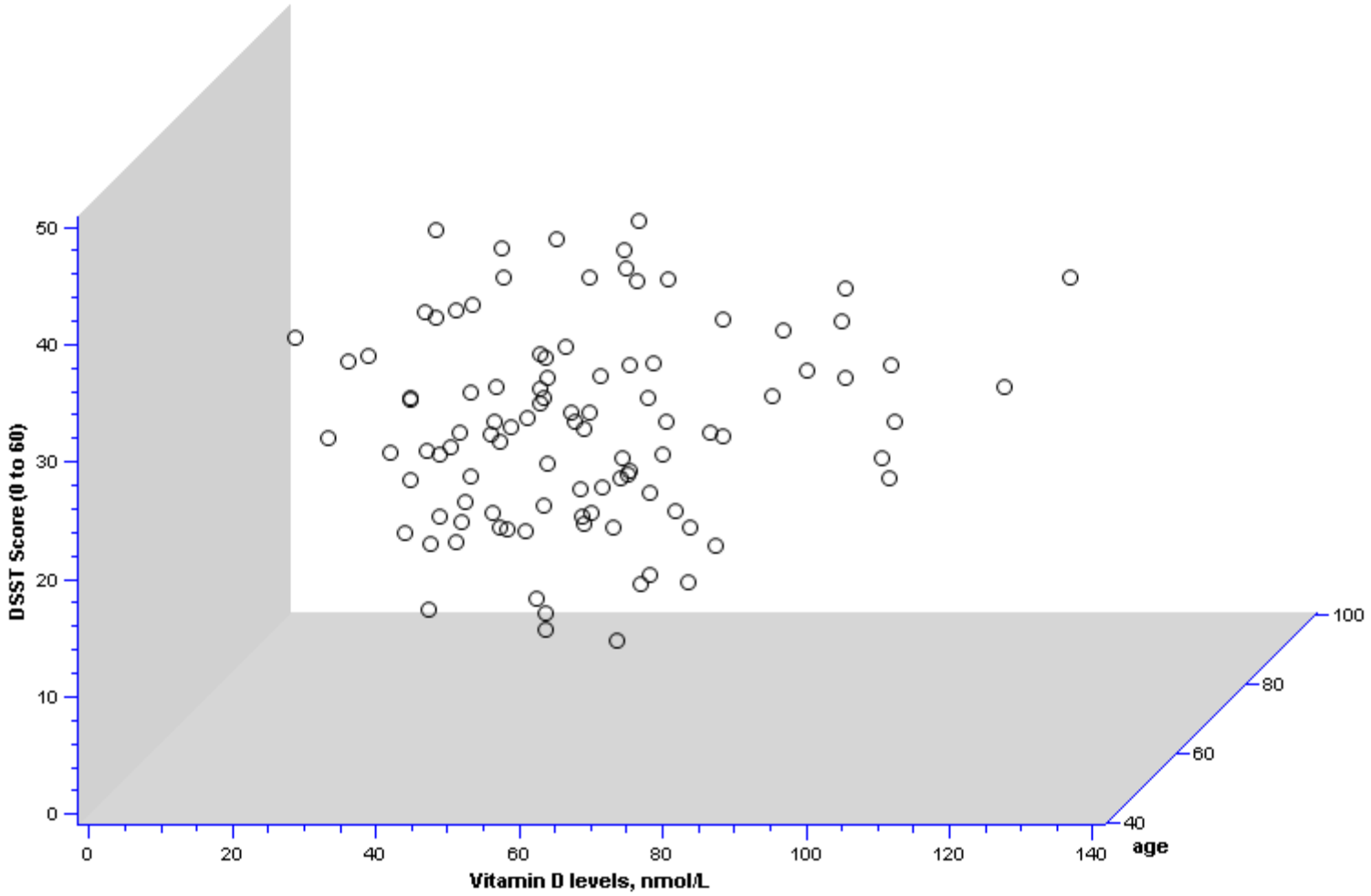
Confounder

- Confounding variables (third variables) are variables that the researcher failed to control, or eliminate, damaging the internal validity of an experiment.

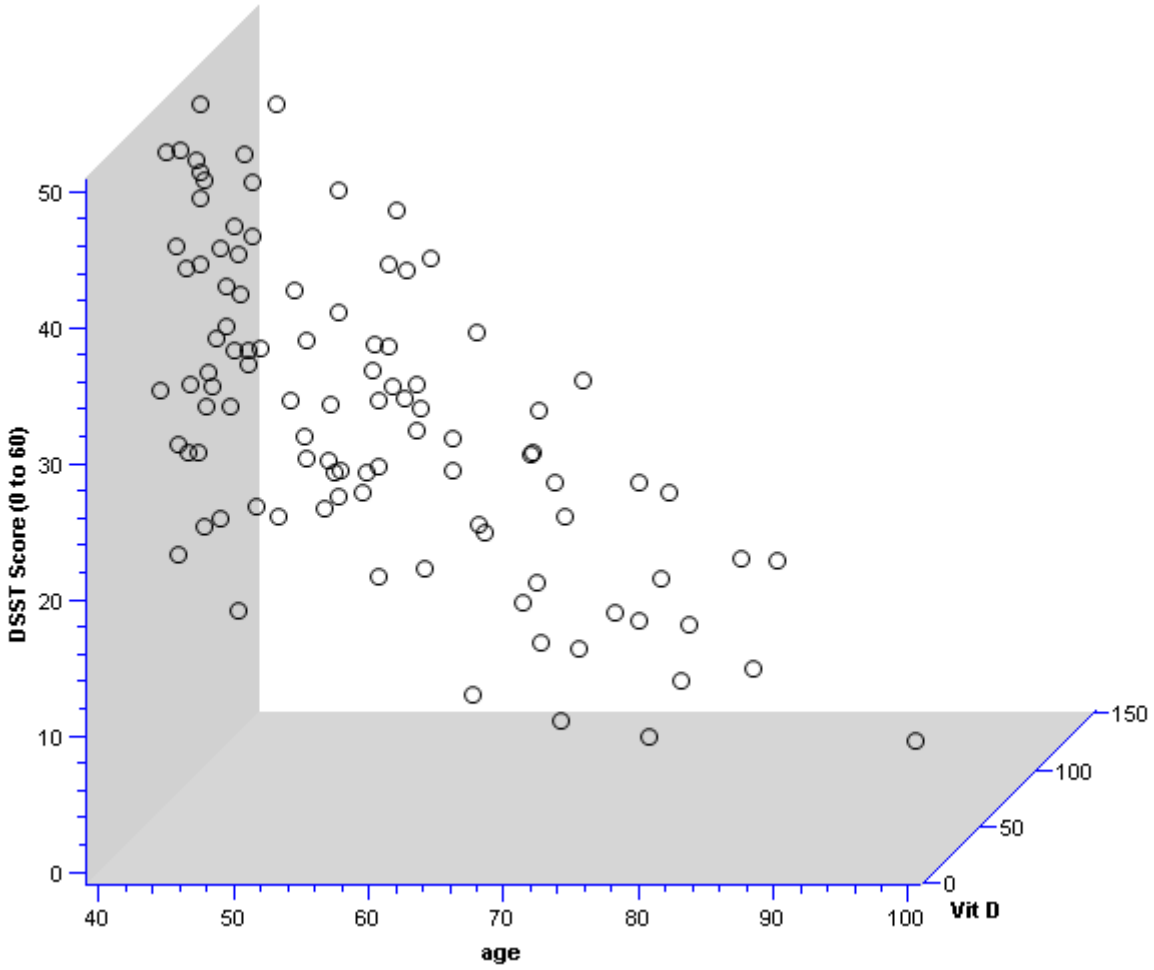
Multiple linear regression…

- What if age is a confounder here?
  - **Older men have lower vitamin D**
  - **Older men have poorer cognition**
- "Adjust" for age by putting age in the model:
  - **DSST score = intercept + slope$_1$*x*vitamin D + slope$_2$ *x*age**

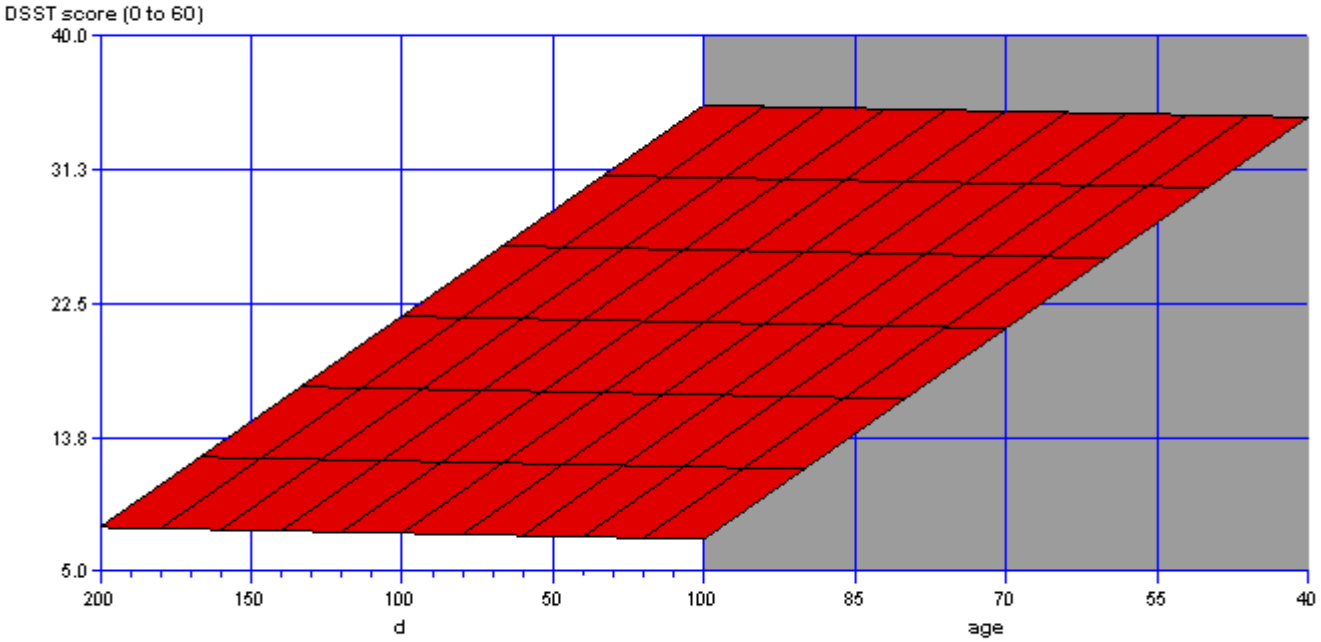
2 predictors: age and vit D…



Different 3D view…



Fit a plane rather than a line…



On the plane, the slope for vitamin D is the same at every age; thus, the slope for vitamin D represents the effect of vitamin D when age is held constant.

Equation of the "Best fit" plane…

- DSST score = 53 + 0.0039*x*vitamin D (in 10 nmol/L) - 0.46 *x*age (in years)

- P-value for vitamin D >>.05
- P-value for age <.0001

- Thus, relationship with vitamin D was due to confounding by age!

Multiple Linear Regression

- More than one predictor…

$E(y) = \alpha + \beta_1 * X + \beta_2 * W + \beta_3 * Z \ldots$

Each regression coefficient is the amount of change in the outcome variable that would be expected per one-unit change of the predictor, if all other variables in the model were held constant.

Functions of multivariate analysis:

- Control for confounders
- Test for interactions between predictors (effect modification)
- Improve predictions

Procedure for Regression Diagnostics…

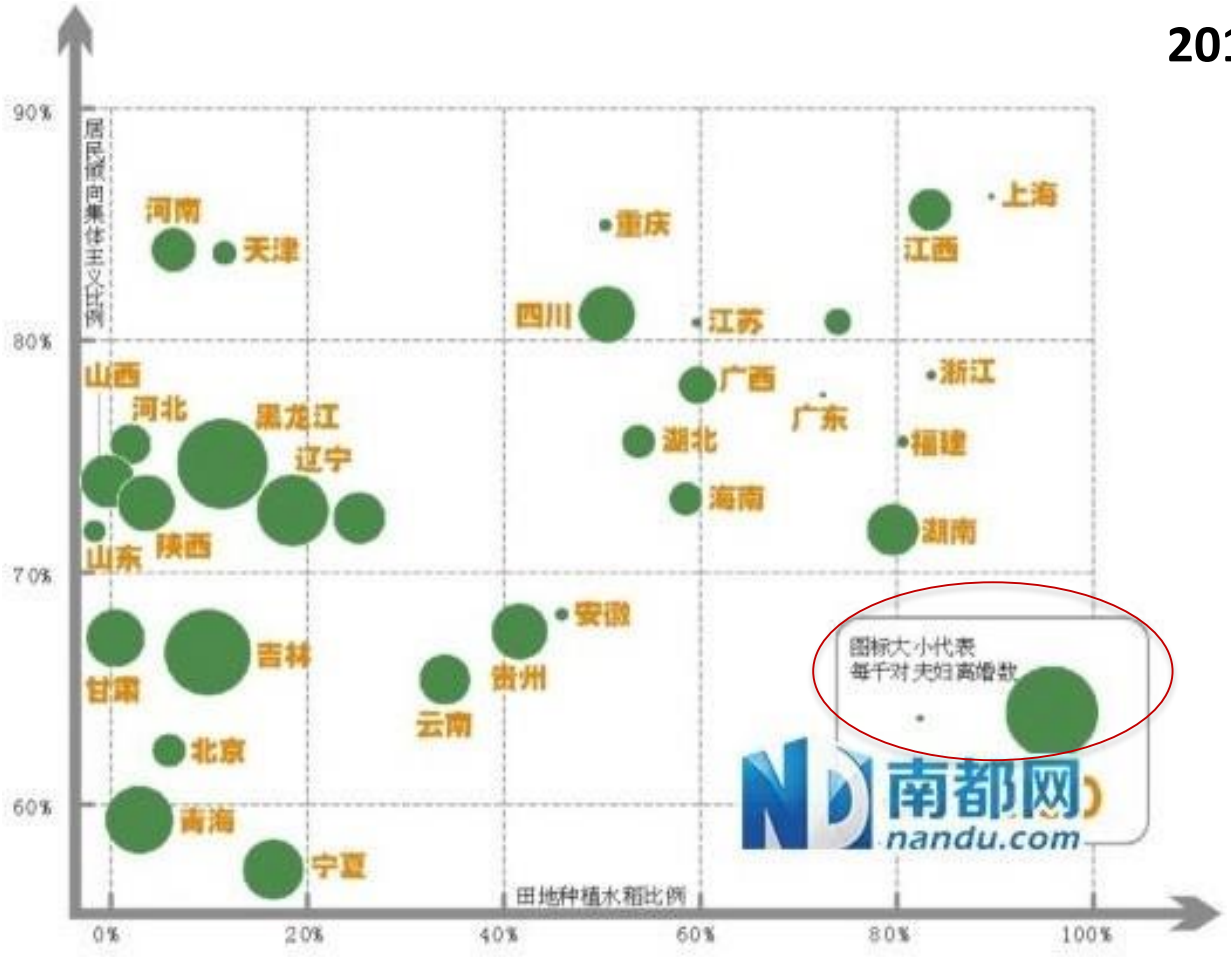
1. Develop a model that has a theoretical basis.
2. Gather data for the two variables in the model.
3. Draw the scatter diagram to determine whether a linear model appears to be appropriate.
4. Determine the regression equation.
5. Calculate the residuals and check the required conditions
6. Assess the model's fit.
7. ***If the model fits the data***, ***use the regression equation*** to predict a particular value of the dependent variable and/or estimate its mean.

Other types of multivariate regression

- Multiple linear regression is for normally distributed outcomes
- Logistic regression is for binary outcomes
- Cox proportional hazards regression is used when time-to-event is the outcome

“研究称中国南方种水稻致离婚率低于北方”

2014年05月14日 深圳新闻网



本图横轴代表各省份田地种植水稻比例，本图纵轴代表各省份居民倾向集体主义的比例

Your comments?

（1）结论可信吗？研究合理吗

（2）一个解读：“在种植水稻的南部，人们更为相互依赖，而北方小麦种植区人们则更加个人主义”。你同意吗？ 原因？

（3）如果你是论文评审，推荐这个研究发表在哪里？

Assignment: read the original paper, give your comments

- "Large-Scale Psychological Differences Within China Explained by Rice Versus Wheat Agriculture," by T. Talhelm et al. *Science*, 2014.

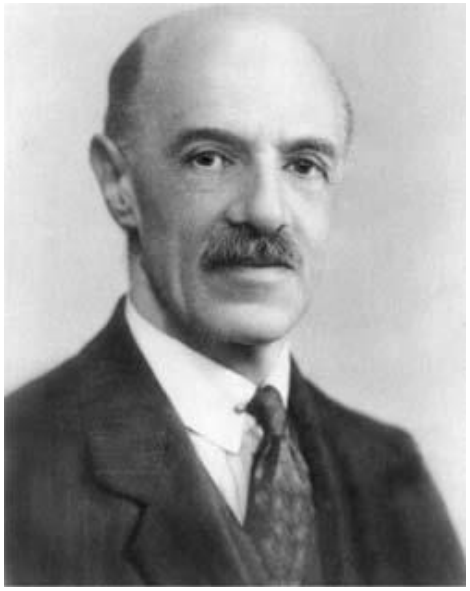
09 MAY 2014
VOL 344, ISSUE 6184

Pearson correlation assumptions

- absence of outliers
- normality of variables
- linearity
- homoscedasticity.

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

What can we do if violations of these conditions happen



Charles Spearman
(1863-1945, UK)

Karl Pearson
(1857-1936, UK)

Spearman rank correlation

a non-parametric version of the conventional Pearson correlation

Spearman rank correlation

- The Spearman rank correlation tests for association without any assumption on the association:
  - Rank the X-values, and rank the Y-values.
  - Compute ordinary sample correlation of the ranks: This is called the Spearman rank correlation.

Spearman's rank correlation

- Spearman's correlation (often denoted by the Greek letter ρ (rho) or as r_s) is then given by

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

Where d_i is the difference between the *i*th rank for *x* and the *i*th rank for *y*

Determining significance

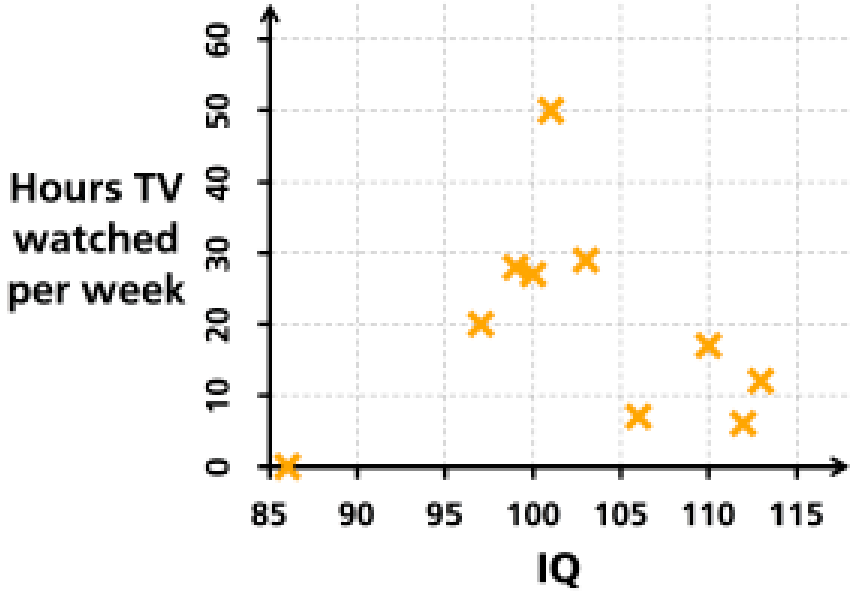
$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

which is distributed approximately as Student's t distribution with *n-2 degrees of freedom*

Example

In this example, the raw data in the table below is used to calculate the correlation between the IQ of a person with the number of hours spent in front of TV per week.

IQ, X_i	Hours of TV per week, Y_i
106	7
86	0
100	27
101	50
99	28
103	29
97	20
113	12
112	6
110	17



Example

IQ, X_i	Hours of TV per week, Y_i	rank x_i	rank y_i	d_i	d_i^2
86	0	1	1	0	0
97	20	2	6	−4	16
99	28	3	8	−5	25
100	27	4	7	−3	9
101	50	5	10	−5	25
103	29	6	9	−3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

With d_i^2 found, add them to find $\sum d_i^2 = 194$.

The value of *n* is 10.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 194}{10(10^2-1)}$$

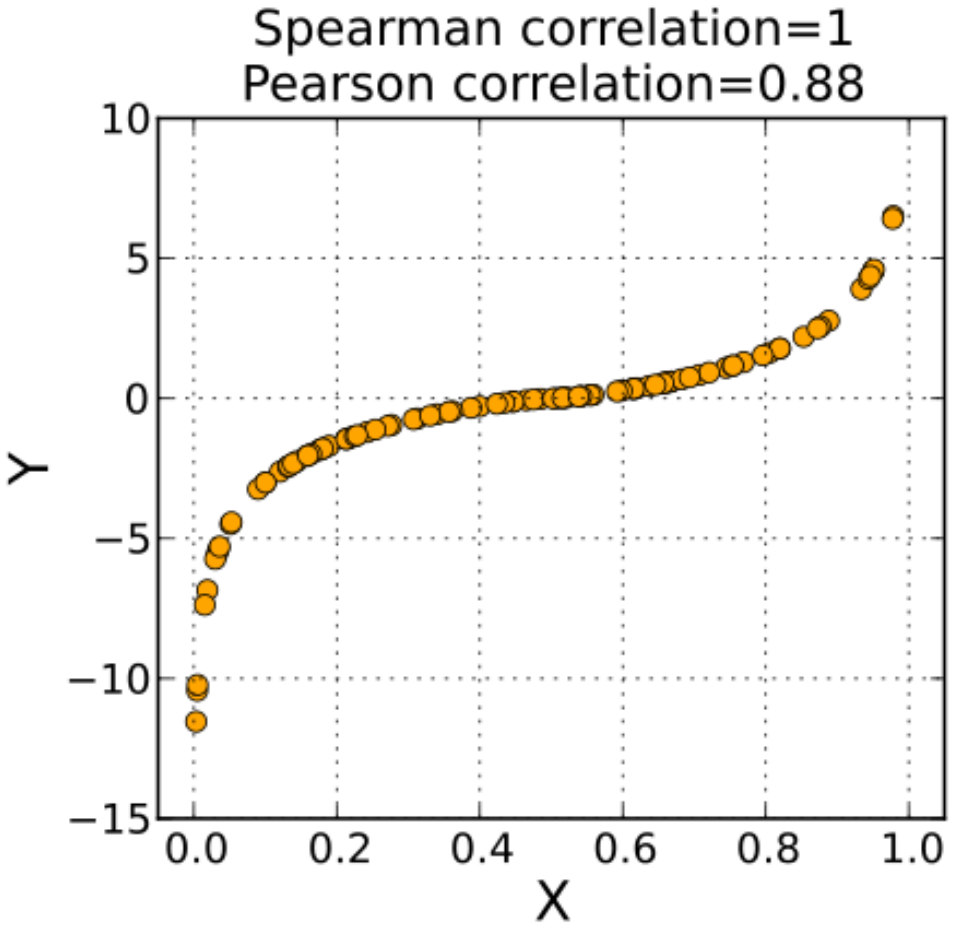
$$\rho = -29/165 = -0.175757575\ldots$$

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

P-value = 0.627188

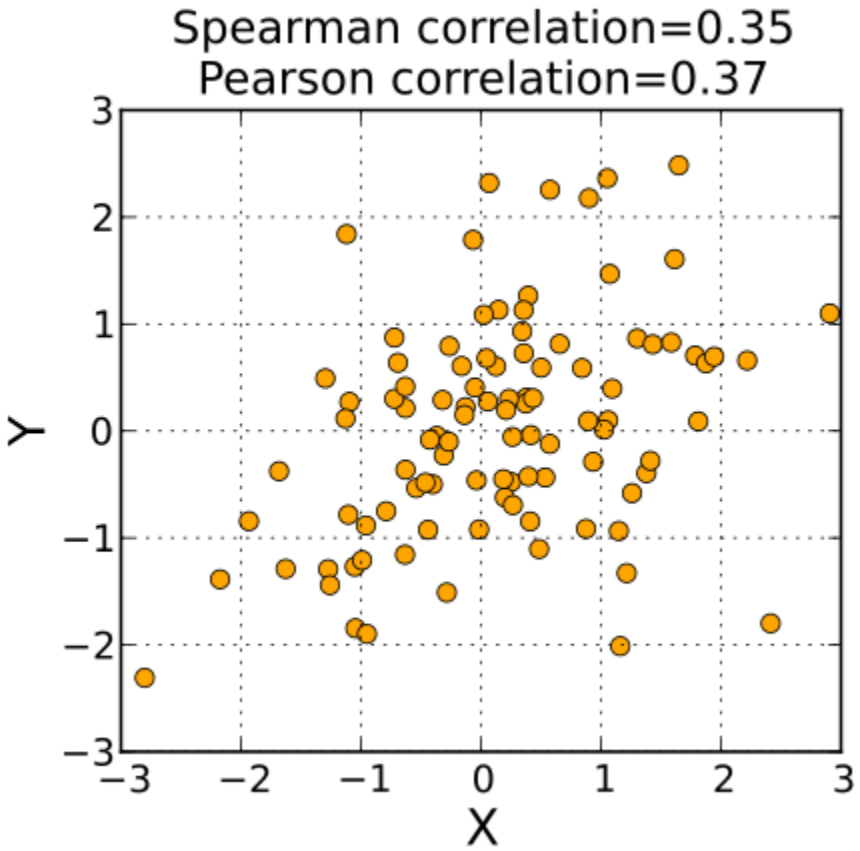
- we can not reject H0
- H0: no association between 2 variables
- H1: association between 2 variables - 2 tailed

Spearman correlation VS. Pearson correlation



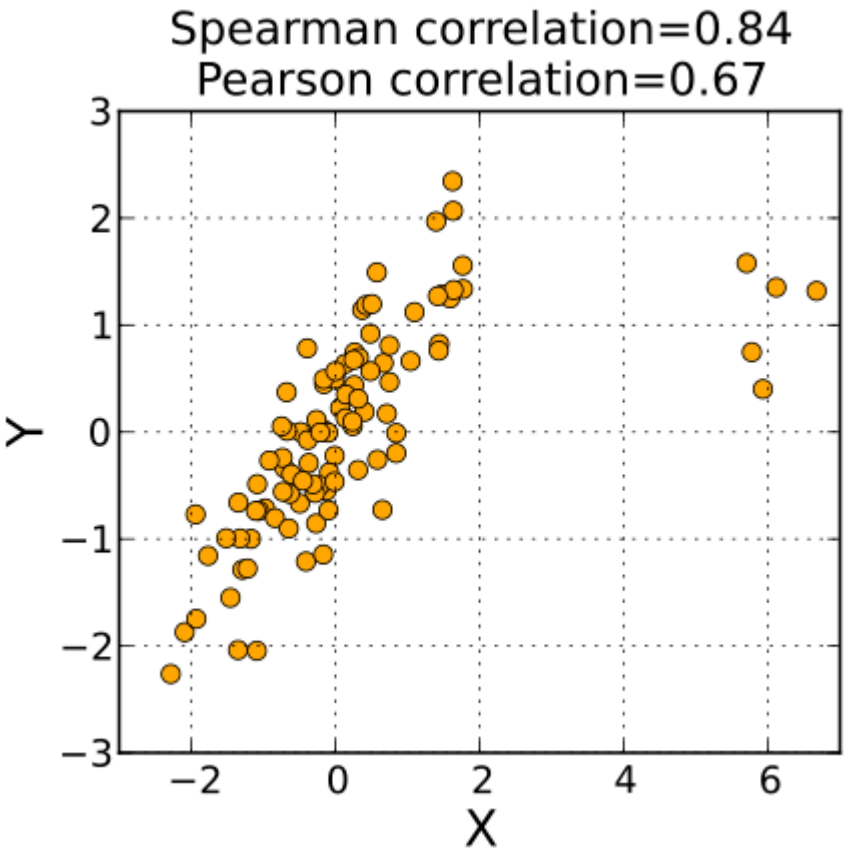
A Spearman correlation of 1 results when the two variables being compared are monotonically related, even if their relationship is not linear. This means that all data-points with greater x-values than that of a given data-point will have greater y-values as well. In contrast, this does not give a perfect Pearson correlation.

Spearman correlation VS. Pearson correlation



When the data are roughly elliptically distributed and there are no prominent outliers, the Spearman correlation and Pearson correlation give similar values.

Spearman correlation VS. Pearson correlation



The Spearman correlation is less sensitive than the Pearson correlation to strong outliers that are in the tails of both samples. That is because Spearman's rho limits the outlier to the value of its rank.

- To calculate a Spearman rank-order correlation and Pearson correlation on data

	Marks									
English	56	75	45	71	62	64	58	80	76	61
Maths	66	70	40	60	65	56	59	77	67	63