

Vancomycin Pharmacokinetic Dosing
Analysis of Patient Data & Optimization of Kinetic Parameters
6/07

Methods:

All patients' data from pharmacokinetic consults was placed in an Excel spreadsheet (age, height, weight, sex, serum creatinine at time of levels, dose and frequency at time of levels, and measured trough). Patients were excluded if they had acute renal failure, unstable renal function, were receiving hemodialysis, or did not have steady state serum levels. Summary data is presented in the table below.

Number of Patients = 107		
	Average (SD)	Range
Age (years)	58.5 (17)	22-90
Male	47%	
Serum Creatinine (mg/dl)	0.98 (0.51)	0.4-4
Creatinine Clearance Calculated at Time of Levels (ml/min)	80 (33)	14.9-173
Height (Inches)	67.8 (4.2)	60-81
Ideal or Actual Body Weight (kg) (which ever was less)	64.9 (11)	43-98.3
Actual Body Weight (kg)	90.9 (25.9)	43-171.8
% of Patients Obese (20% or more above Ideal Weight)	64%	
% Above Ideal Weight (TBW-IBW)*100/IBW	42% (42%)	0-184%
Volume of Distribution = 0.65 * Total Body Weight (liters)	59 (16.8)	27.9-111.7
Dose Administered At Time of Level (mg)	1214 (352)	500-2000
Dosage Interval (hours)	18.5 (8)	8-48
Trough Measured At Time of Level (mcg/ml)	9.6 (3.9)	1.5-22.4
Predicted Trough Using Matzke Elimination Rate Constant (0.00083*creatinine clearance + 0.0044)	11 (3.2)	3.6-19.5
Predicted Trough Using Updated Equation (0.00107*creatinine clearance*1.73/Surface Area) +0.005216005	9.3 (3.2)	1.8-19.9

Ideal body weight and body surface area were calculated for each patient using the Devine and DuBois & DuBois formulas respectively. The program currently uses Cockcroft-Gault equation with ideal body weight and actual serum creatinine to calculate creatinine clearance. The current volume of distribution is 0.65 l/kg of total body weight.

One compartment, open model, pharmacokinetic dosing equations were incorporated into the spreadsheet and were used to calculate the predicted trough for each patient using their individual data. A non linear fitting routine to minimize the sum of the square of the errors, sum of (actual level-predicted level)², was used to optimized pharmacokinetic parameters. The parameters that were optimized are bolded in the equations below. Actual serum creatinine values were used and were not adjusted upward when they were less than 1 mg/dl. The following equations were optimized during the fittings.

$$\text{Creatinine clearance (ml/min or ml/min/1.73 Meters}^2) = (140 - \text{age}) * [\text{IBW} + ((\text{TBW} - \text{IBW}) * \text{Fat Factor})] * (\mathbf{1.73/\text{SA}_{\text{patient}}})^{1 \text{ or } 0} / (72 * \text{Serum creatinine}_{\text{mg/dl}})$$

When the exponent is set to 0 the result for the quantity within the parentheses is equal to 1.

$$\text{Vd (liters)} = \mathbf{\text{Vd}_{\text{l/kg}}} * \text{total body weight}$$

$$\text{K (hours}^{-1}\text{)} = \mathbf{\text{slope}} * \mathbf{\text{Clcr}_{\text{ml/min or ml/min/1.73 M}^2}} + \mathbf{\text{intercept}}$$

$$\text{Predicted level (mcg/ml)} = \text{Dose}_{\text{mg}} (1 - e^{-k*t'}) * e^{-k*(\text{Tau} - t')} / [\text{TBW} * \mathbf{\text{Vd}_{\text{l/kg}}} * \mathbf{k} * t' * (1 - e^{-k*\text{Tau}})]$$

k = elimination rate constant (hour⁻¹)

t' = infusion period (hours)

Tau = Dosage Interval (hours)

IBW = Ideal Body Weight (kg)

TBW = Total Body Weight (kg)

SA_{patient} = Surface Area (M²)

When the fat factor was optimized its limits were set to a range of ≥ 0 and ≤ 1 .

When Vd (l/kg) was optimized its limits were set to a range of ≥ 0.3 and ≤ 1.3 .

When surface area was included in the equation the exponent was set to 1.

When the slope of elimination rate was optimized its limits were set to a range of ≥ 0.0001 and ≤ 0.002

When the intercept to the elimination rate was optimized its limits were set to a range of ≥ 0 and ≤ 0.01

Data Fittings:

The following data fitting were performed. The fit results for the parameters are noted.

Fittings	Vd (l/kg)	Elimination Rate Constant Slope	Elimination Rate Constant Intercept	Fat Factor	Surface Area Factor	(Actual-Predicted) Average	(Actual-Predicted) Standard Deviation	Sum of (Actual-Predicted) <i>Bias</i>	Absolute of (Actual-Predicted) Average	Absolute of (Actual - Predicted) Standard Deviation	Sum of Absolute (Actual - Predicted) <i>Precision</i>	Sum of Square of Errors (Actual - Predicted) ²
Current Model	0.65	0.00083	0.0044	Not Used	No Used	-1.49	3.86	-159	3.5	2.2	374	1821
1	0.65 Fixed	Fit Value 0.000896	Fit Value 0.00663531	Not Used	Not Used	0.26	3.7	27.7	3.1	2.1	327	1461
2	0.65 Fixed	Fit Value 0.000896	Fit Value 0.00663531	Fit Value 0	Not Used							
3	Fit Value 1.3	Fit Value 0.000565	Fit Value 0.003907852	Fit Value 0	Not Used	0.18	3.6	19.4	3	2.1	319	1410
4	0.65 Fixed	Fit Value 0.001053	Fit Value 0.005126106	Fit Value 0.039709	1.73/SA	0.24	3.4	25.6	2.74	2.1	294	1257
5	0.65 Fixed	Fit Value 0.00107	Fit Value 0.005216005	Fit Value 0	1.73/SA	0.23	3.4	24.4	2.72	2.1	291	1261
6	Fit Value 1.3	Fit Value 0.000674	Fit Value 0.002908673	Fit Value 0	1.73/SA	0.14	3.2	15.3	2.68	2	287	1171
7	0.65 Fixed	Fit Value 0.000772	Fit Value 0.008617933	0.4 Fixed	Not Used	0.52	4.2	55.7	3.37	2.54	360	1902
8	0.65 Fixed	Fit Value 0.000926	Fit Value 0.005627497	0.4 Fixed	1.73/SA	0.33	3.7	34.9	2.92	2.3	312	1436

Results:

- Data fittings that did not include the patient’s body surface area all gave similar results and were an improvement over the current model with the exception of the fitting including a fixed fat factor of 0.4 which made predictions worse. Allowing the fat factor and/or Vd (l/kg) to be included in the fitting did not improve the results.
- When the patient’s surface area was included in the equation the data fittings improved and bias related to weight, height, and the patient’s body surface area were minimized. Including the fat factor with the surface area factor did not improve the fitting and setting the fat factor to 0.4 made predictions worse.

The selected equation to update the dosing program is bolded below:

$$K \text{ (hours-1)} = \mathbf{0.00107} * \text{Clcr}_{\text{ml/min}} * \mathbf{(1.73/SA_{\text{patient}})} + \mathbf{0.005216005}$$

The programs equation for creatinine clearance will remain unchanged as the surface area factor may be incorporated into the elimination rate constant equation above and will give the same results without impacting the dosage calculations in the aminoglycoside program.

The following equations will remain unchanged.

$$\text{Clcr (ml/min)} = [(140-\text{age}) * \text{IBW} / (72 * \text{Serum creatinine})] , * 0.85 \text{ if female}$$

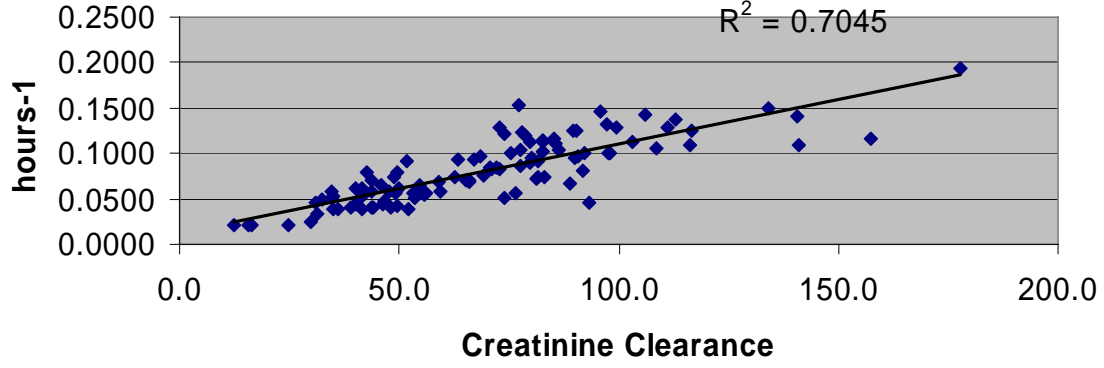
$$\text{Vd (liters)} = 0.65 \text{ l/kg} * \text{total body weight}$$

The results of data fitting 6 will not be used to optimized the program as incorporating a V_d of 1.3 l/kg would cause the loading dose calculated to double from 20-25 mg/kg to 40-50 mg/kg, which is unreasonable. The vast majority of studies have found a V_d of around 0.65 l/kg. The V_d value of 1.3 l/kg could potentially cause the user to select a large maintenance dose to be given at long dosing intervals.

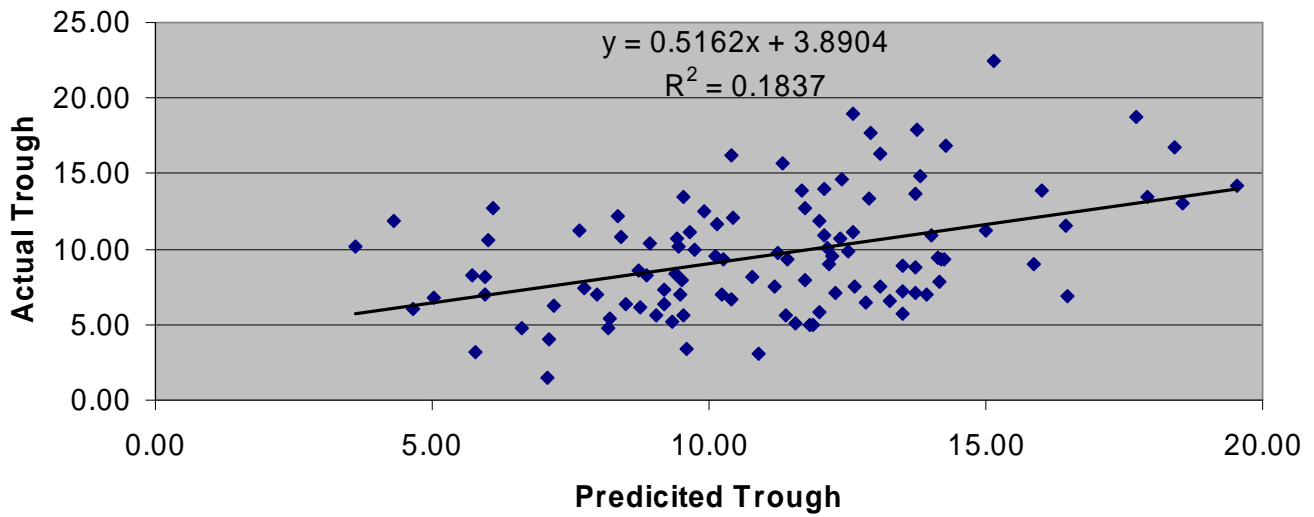
K Actual versus Clcr with fitting

$$y = 0.001x + 0.0128$$

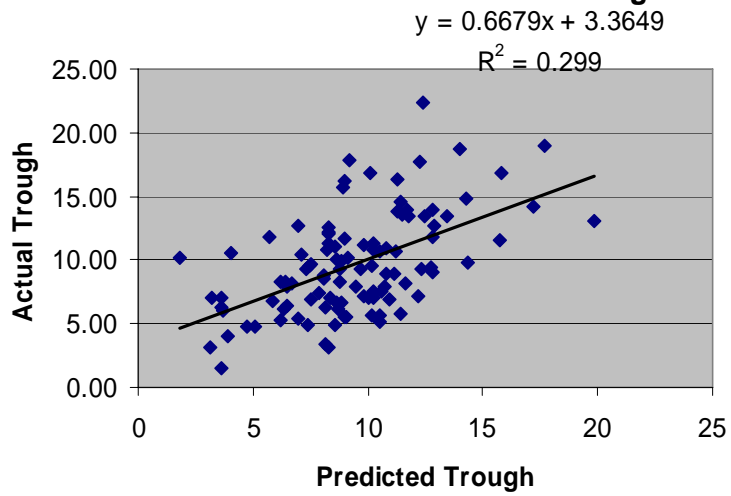
$$R^2 = 0.7045$$

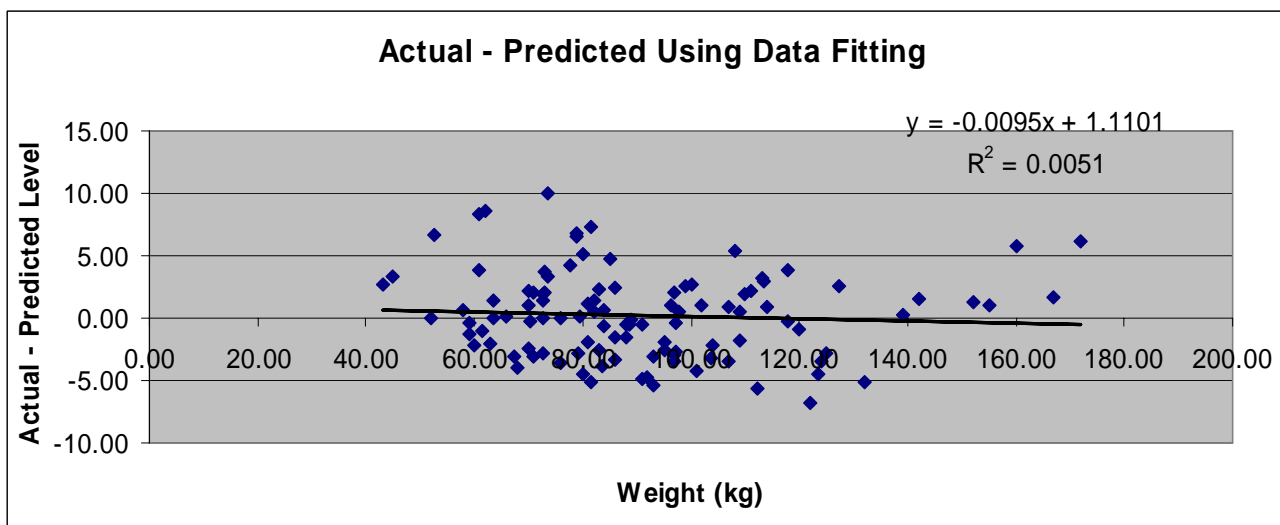
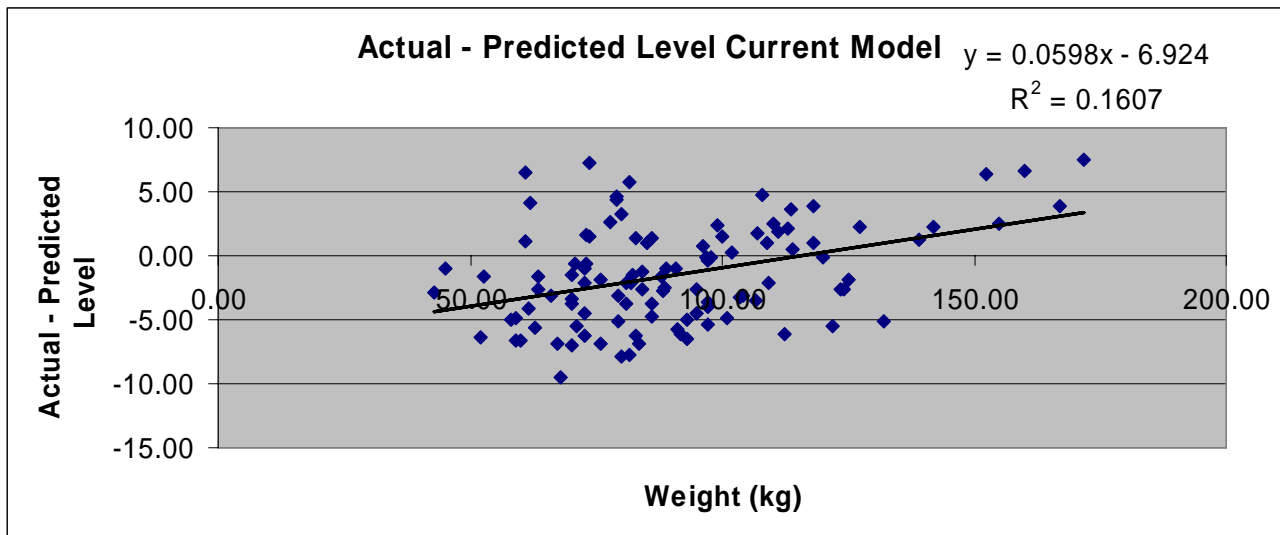


Actual Vrs Predicted Levels Current Model



Actual vrs Predicted Levels With Fitting

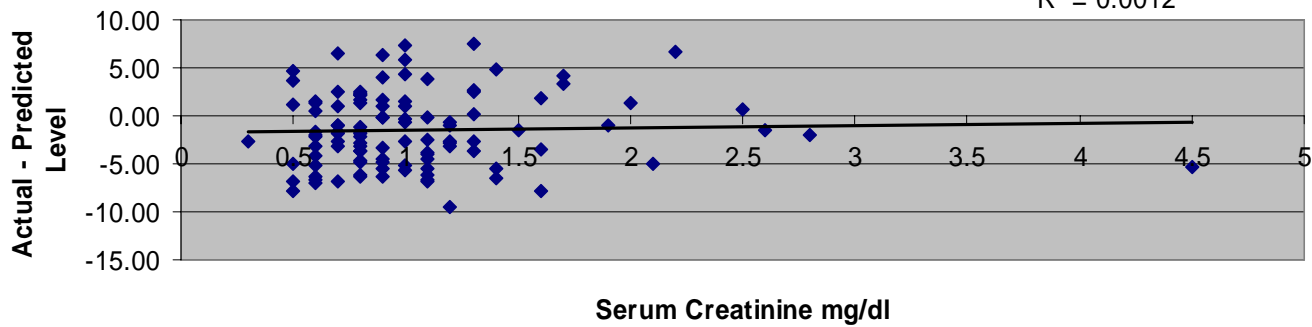




Actual Level - Predicted Level Current Model

$$y = 0.2333x - 1.7321$$

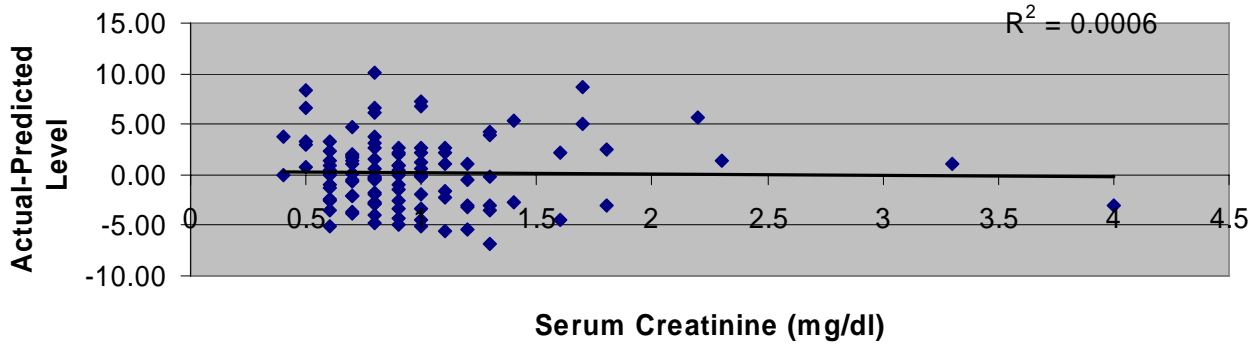
$$R^2 = 0.0012$$

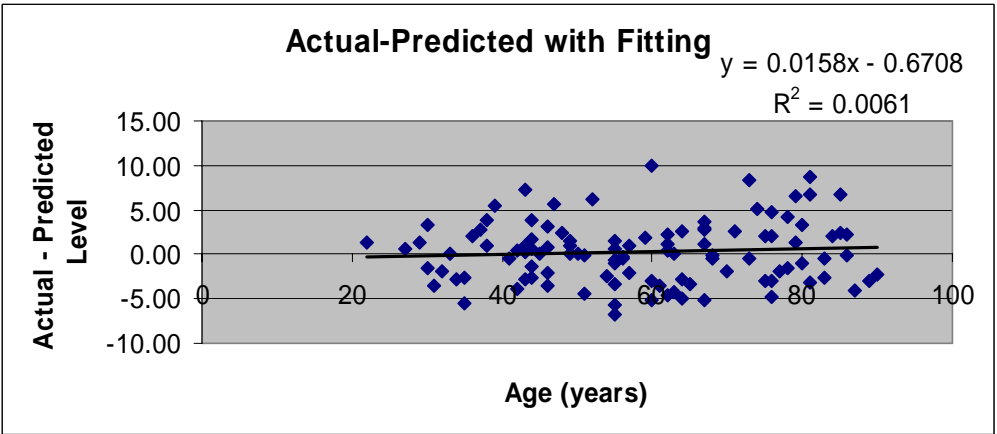
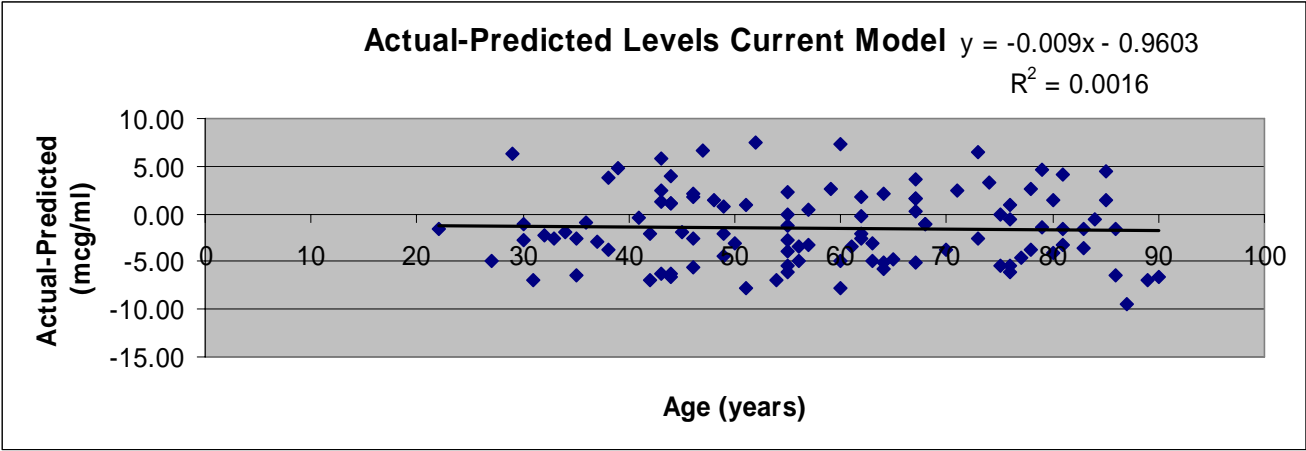


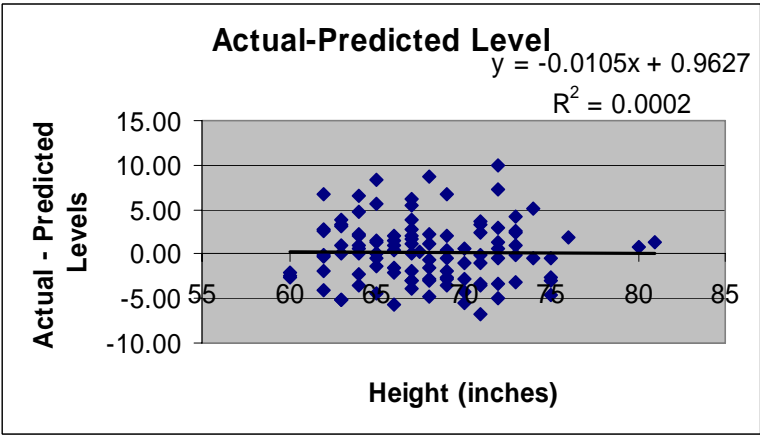
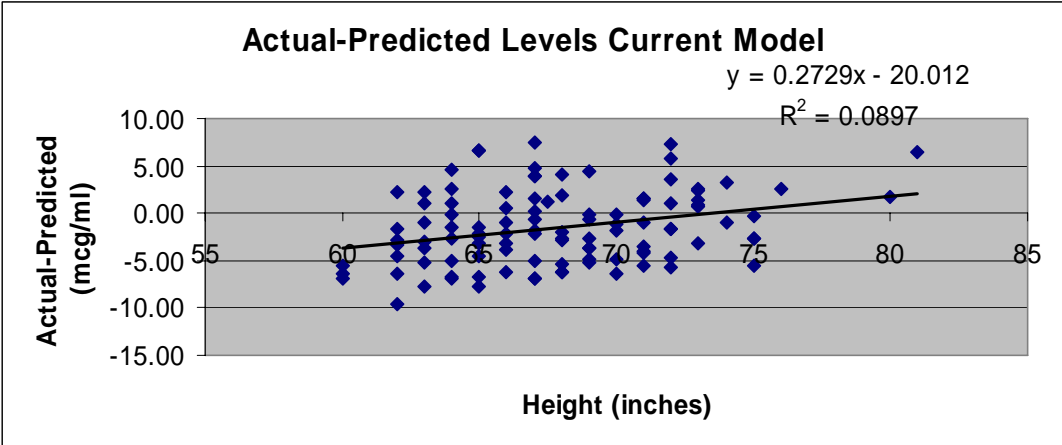
Actual-Predicted Level With Data Fitting

$$y = -0.1639x + 0.4105$$

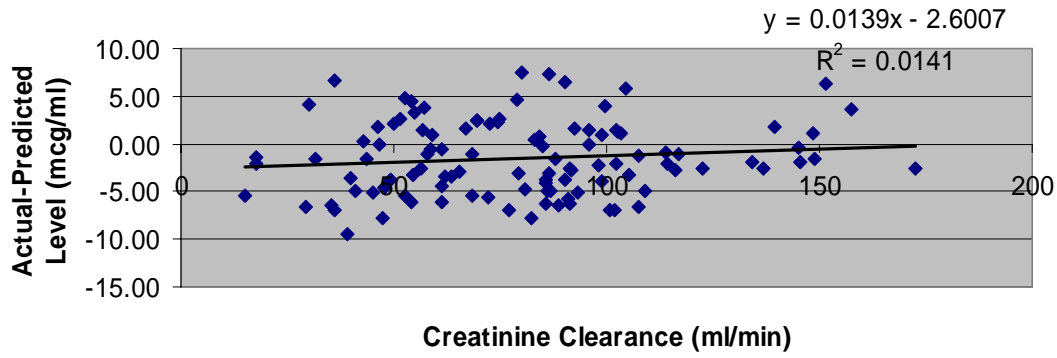
$$R^2 = 0.0006$$



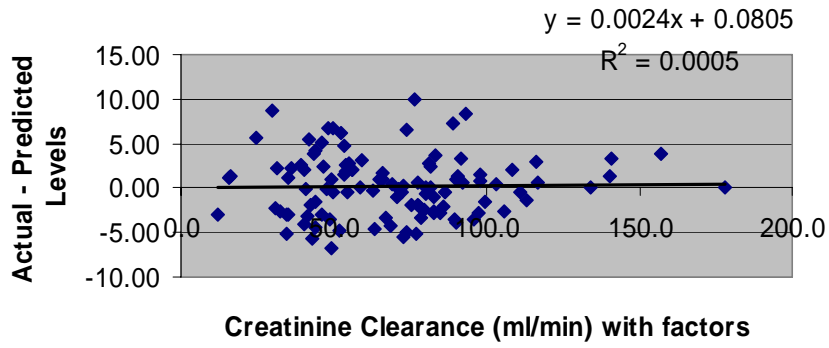




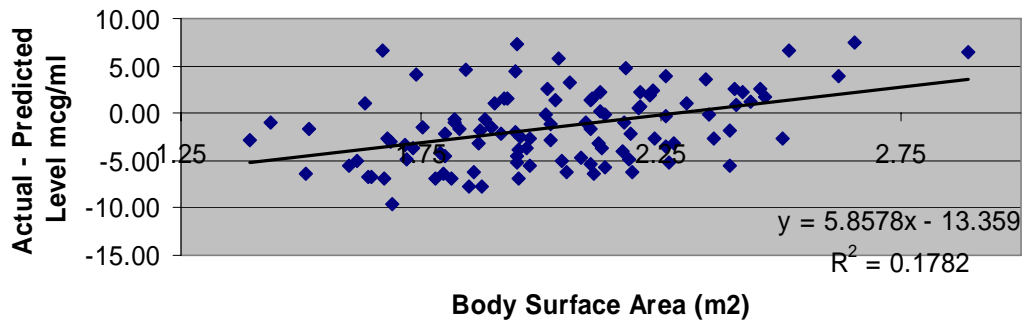
Actual-Predicted Levels Current Model



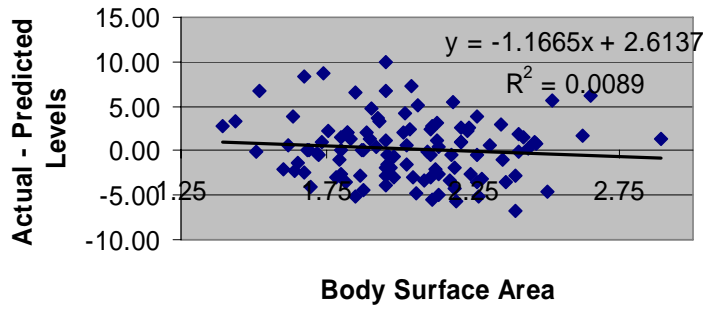
Actual-Predicted Using Fitting



Actual - Predicted Level Current Model



Actual-Predicted with fitting



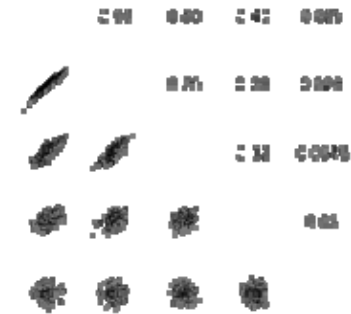
Correlation

From Wikipedia, the free encyclopedia

(Redirected from [Correlation coefficient](#))

Jump to: [navigation](#), [search](#)

This article is about the correlation coefficient between two variables. The term correlation can also mean the [cross-correlation](#) of two [functions](#) or [electron correlation](#) in molecular systems.



Positive linear correlations between 1000 pairs of numbers. The data are graphed on the lower left and their correlation coefficients listed on the upper right. Each square in the upper right corresponds to its mirror-image square in the lower left, the "mirror" being the diagonal of the whole array. Each set of points correlates maximally with itself, as shown on the diagonal (all correlations = +1).

In [probability theory](#) and [statistics](#), **correlation**, also called **correlation coefficient**, indicates the strength and direction of a linear relationship between two [random variables](#). In general statistical usage, *correlation* or correlation refers to the departure of two variables from independence, although [correlation does not imply causation](#). In this broad sense there are several coefficients, measuring the degree of correlation, adapted to the nature of data.

A number of different coefficients are used for different situations. The best known is the [Pearson product-moment correlation coefficient](#), which is obtained by dividing the [covariance](#) of the two variables by the product of their [standard deviations](#). Despite its name, it was first introduced by [Francis Galton](#).

Contents

[\[hide\]](#)

- [1 Pearson's product-moment coefficient](#)
 - [1.1 Mathematical properties](#)
 - [1.2 The sample correlation](#)
 - [1.3 Geometric Interpretation of correlation](#)
 - [1.4 Interpretation of the size of a correlation](#)
- [2 Non-parametric correlation coefficients](#)
- [3 Other measures of dependence among random variables](#)
- [4 Copulas and correlation](#)
- [5 Correlation matrices](#)

- [6 Common misconceptions about correlation](#)
 - [6.1 Correlation and causality](#)
 - [6.2 Correlation and linearity](#)
- [7 Computing correlation accurately in a single pass](#)
- [8 Currency correlation](#)
- [9 See also](#)
- [10 Notes and references](#)
- [11 Further reading](#)
- [12 External links](#)

[edit] Pearson's product-moment coefficient

Main article: [Pearson product-moment correlation coefficient](#)

[edit] Mathematical properties

The correlation coefficient $\rho_{X,Y}$ between two [random variables](#) X and Y with [expected values](#) μ_X and μ_Y and [standard deviations](#) σ_X and σ_Y is defined as:

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E((X - \mu_X)(Y - \mu_Y))}{\sigma_X \sigma_Y},$$

where E is the [expected value](#) operator and cov means [covariance](#). Since $\mu_X = E(X)$, $\sigma_X^2 = E(X^2) - E^2(X)$ and likewise for Y , we may also write

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)} \sqrt{E(Y^2) - E^2(Y)}}.$$

The correlation is defined only if both of the standard deviations are finite and both of them are nonzero. It is a corollary of the [Cauchy-Schwarz inequality](#) that the correlation cannot exceed 1 in [absolute value](#).

The correlation is 1 in the case of an increasing linear relationship, -1 in the case of a decreasing linear relationship, and some value in between in all other cases, indicating the degree of [linear dependence](#) between the variables. The closer the coefficient is to either -1 or 1 , the stronger the correlation between the variables.

If the variables are [independent](#) then the correlation is 0, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables. Here is an example: Suppose the random variable X is uniformly distributed on the interval from -1 to 1 , and $Y = X^2$. Then Y is completely determined by X , so that X and Y are dependent, but their correlation is zero; they are [uncorrelated](#). However, in the special case when X and Y are [jointly normal](#), independence is equivalent to uncorrelatedness.

A correlation between two variables is diluted in the presence of measurement error around estimates of one or both variables, in which case [disattenuation](#) provides a more accurate coefficient .

[edit] The sample correlation

If we have a series of n measurements of X and Y written as x_i and y_i where $i = 1, 2, \dots, n$, then the [Pearson product-moment correlation coefficient](#) can be used to estimate the correlation of X and Y . The Pearson coefficient is also known as the "sample correlation coefficient". It is especially important if X and Y are both

[normally distributed](#). The Pearson correlation coefficient is then the best estimate of the correlation of X and Y . The Pearson correlation coefficient is written:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n - 1)s_x s_y},$$

where \bar{x} and \bar{y} are the sample [means](#) of X and Y , s_x and s_y are the sample [standard deviations](#) of X and Y and the sum is from $i = 1$ to n . As with the population correlation, we may rewrite this as

$$r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}.$$

Again, as is true with the population correlation, the absolute value of the sample correlation must be less than or equal to 1. Though the above formula conveniently suggests a single-pass algorithm for calculating sample correlations, it is notorious for its numerical instability (see below for something more accurate).

The square of the sample correlation coefficient, which is also known as the [coefficient of determination](#), is the fraction of the variance in y_i that is accounted for by a linear fit of x_i to y_i . This is written

$$r_{xy}^2 = 1 - \frac{s_{y|x}^2}{s_y^2},$$

where $s_{y|x}^2$ is the square of the error of a [linear regression](#) of x_i on y_i by the [equation](#) $y = a + bx$:

$$s_{y|x}^2 = \frac{1}{n - 1} \sum_{i=1}^n (y_i - a - bx_i)^2,$$

and s_y^2 is just the variance of y :

$$s_y^2 = \frac{1}{n - 1} \sum_{i=1}^n (y_i - \bar{y})^2.$$

Note that since the sample correlation coefficient is symmetric in x_i and y_i , we will get the same value for a fit of x_i to y_i :

$$r_{xy}^2 = 1 - \frac{s_{x|y}^2}{s_x^2}.$$

This equation also gives an intuitive idea of the correlation coefficient for higher [dimensions](#). Just as the above described sample correlation coefficient is the fraction of variance accounted for by the fit of a 1-dimensional [linear submanifold](#) to a set of 2-dimensional vectors (x_i, y_i) , so we can define a correlation coefficient for a fit of an m -dimensional linear submanifold to a set of n -dimensional vectors. For example, if we fit a plane $z = a + bx + cy$ to a set of data (x_i, y_i, z_i) then the correlation coefficient of z to x and y is

$$r^2 = 1 - \frac{\sigma_{z|xy}^2}{s_z^2}.$$

[\[edit\]](#) Geometric Interpretation of correlation

The correlation coefficient can also be viewed as the [cosine](#) of the [angle](#) between the two [vectors](#) of samples drawn from the two random variables.

Caution: This method only works with centered data, i.e., data which have been shifted by the sample mean so as to have an average of zero. Some practitioners prefer an uncentered (non-Pearson-compliant) correlation coefficient. See the example below for a comparison.

As an example, suppose five countries are found to have gross national products of 1, 2, 3, 5, and 8 billion dollars, respectively. Suppose these same five countries (in the same order) are found to have 11%, 12%, 13%, 15%, and 18% poverty. Then let \mathbf{x} and \mathbf{y} be ordered 5-element vectors containing the above data: $\mathbf{x} = (1, 2, 3, 5, 8)$ and $\mathbf{y} = (0.11, 0.12, 0.13, 0.15, 0.18)$.

By the usual procedure for finding the angle between two vectors (see [dot product](#)), the *uncentered* correlation coefficient is:

$$\cos \theta = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} = \frac{2.93}{\sqrt{103}\sqrt{0.0983}} = 0.920814711.$$

Note that the above data were deliberately chosen to be perfectly correlated: $y = 0.10 + 0.01x$. The Pearson correlation coefficient must therefore be exactly one. Centering the data (shifting \mathbf{x} by $E(\mathbf{x}) = 3.8$ and \mathbf{y} by $E(\mathbf{y}) = 0.138$) yields $\mathbf{x} = (-2.8, -1.8, -0.8, 1.2, 4.2)$ and $\mathbf{y} = (-0.028, -0.018, -0.008, 0.012, 0.042)$, from which

$$\cos \theta = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} = \frac{0.308}{\sqrt{30.8}\sqrt{0.00308}} = 1,$$

as expected.

[\[edit\]](#) Interpretation of the size of a correlation

Several authors have offered guidelines for the interpretation of a correlation coefficient. Cohen (1988),^[1] for example, has suggested the following interpretations for correlations in psychological research, in the table on the right.

Correlation	Negative	Positive
Small	−0.29 to −0.10	0.10 to 0.29
Medium	−0.49 to −0.30	0.30 to 0.49
Large	−1.00 to −0.50	0.50 to 1.00

As Cohen himself has observed, however, all such criteria are in some ways arbitrary and should not be observed too strictly. This is because the interpretation of a correlation coefficient depends on the context and purposes. A correlation of 0.9 may be very low if one is verifying a physical law using high-quality instruments, but may be regarded as very high in the social sciences where there may be a greater contribution from complicating factors.

[\[edit\]](#) Non-parametric correlation coefficients

Pearson's correlation coefficient is a [parametric statistic](#), and it may be less useful if the underlying assumption of normality is violated. [Non-parametric](#) correlation methods, such as [Chi-square](#), [Point biserial correlation](#), [Spearman's \$\rho\$](#) and [Kendall's \$\tau\$](#) may be useful when distributions are not normal; they are a little less powerful than parametric methods if the assumptions underlying the latter are met, but are less likely to give distorted results when the assumptions fail.

[\[edit\]](#) Other measures of dependence among random variables

To get a measure for more general dependencies in the data (also nonlinear) it is better to use the [correlation ratio](#) which is able to detect almost any functional dependency, or [mutual information/total correlation](#) which is capable of detecting even more general dependencies.

[\[edit\]](#) Copulas and correlation

The information given by a correlation coefficient is not enough to define the dependence structure between random variables; to fully capture it we must consider the [copula](#) between them. The correlation coefficient completely defines the dependence structure only in very particular cases, for example when the [cumulative distribution functions](#) are the [multivariate normal distributions](#). In the case of elliptic distributions it characterizes the (hyper-)ellipses of equal density, however, it does not completely characterize the dependence

structure (for example, the a multivariate t-distribution's degrees of freedom determine the level of tail dependence).

[\[edit\]](#) Correlation matrices

The correlation matrix of n random variables X_1, \dots, X_n is the $n \times n$ matrix whose i,j entry is $\text{corr}(X_i, X_j)$. If the measures of correlation used are product-moment coefficients, the correlation matrix is the same as the [covariance matrix](#) of the standardized random variables $X_i / \text{SD}(X_i)$ for $i = 1, \dots, n$. Consequently it is necessarily a [non-negative definite matrix](#).

The correlation matrix is symmetrical (the correlation between X_i and X_j is the same as the correlation between X_j and X_i).

[\[edit\]](#) Common misconceptions about correlation

[\[edit\]](#) Correlation and causality

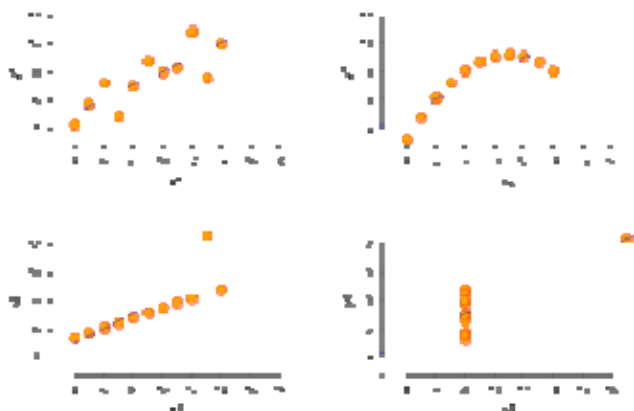
The conventional dictum that "[correlation does not imply causation](#)" means that correlation cannot be validly used to infer a causal relationship between the variables. This dictum should not be taken to mean that correlations cannot indicate causal relations. However, the causes underlying the correlation, if any, may be indirect and unknown. Consequently, establishing a correlation between two variables is a not sufficient condition to establish a causal relationship (in either direction).

Another way to say this is that [lack of correlation does not imply independence](#).

A correlation between age and height in children is fairly causally transparent, but a correlation between mood and health in people is less so. Does improved mood lead to improved health? Or does good health lead to good mood? Or does some other factor underlie both? Or is it pure coincidence? In other words, a correlation can be taken as evidence for a possible causal relationship, but cannot indicate what the causal relationship, if any, might be.

Importantly, a lack of correlation also does not imply the lack of a relationship between to variables, or even the lack of a causal relationship. For example, the variables X and $\sin X$ are uncorrelated, but have a strict functional relationship.

[\[edit\]](#) Correlation and linearity



Four sets of data with the same correlation of 0.81

While Pearson correlation indicates the strength of a linear relationship between two variables, its value alone may not be sufficient to evaluate this relationship, especially in the case where the assumption of normality is incorrect.

The image on the right shows [scatterplots](#) of [Anscombe's quartet](#), a set of four different pairs of variables created by [Francis Anscombe](#).^[2] The four y variables have the same mean (7.5), standard deviation (4.12), correlation (0.81) and regression line ($y = 3 + 0.5x$). However, as can be seen on the plots, the distribution of the variables is very different. The first one (top left) seems to be distributed normally, and corresponds to what one would expect when considering two variables correlated and following the assumption of normality. The second one (top right) is not distributed normally; while an obvious relationship between the two variables can be observed, it is not linear, and the Pearson correlation coefficient is not relevant. In the third case (bottom left), the linear relationship is perfect, except for one [outlier](#) which exerts enough influence to lower the correlation coefficient from 1 to 0.81. Finally, the fourth example (bottom right) shows another example when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear.

These examples indicate that the correlation coefficient, as a summary statistic, can not replace the individual examination of the data.

[\[edit\]](#) Computing correlation accurately in a single pass

The following algorithm (in [pseudocode](#)) will estimate correlation with good numerical stability

```
sum_sq_x = 0
sum_sq_y = 0
sum_coproduct = 0
mean_x = x[1]
mean_y = y[1]
for i in 2 to N:
    sweep = (i - 1.0) / i
    delta_x = x[i] - mean_x
    delta_y = y[i] - mean_y
    sum_sq_x += delta_x * delta_x * sweep
    sum_sq_y += delta_y * delta_y * sweep
    sum_coproduct += delta_x * delta_y * sweep
    mean_x += delta_x / i
    mean_y += delta_y / i
pop_sd_x = sqrt( sum_sq_x / N )
pop_sd_y = sqrt( sum_sq_y / N )
cov_x_y = sum_coproduct / N
correlation = cov_x_y / (pop_sd_x * pop_sd_y)
```

For an enlightening experiment, check the correlation of $\{900,000,000 + i \text{ for } i=1\dots 100\}$ with $\{900,000,000 - i \text{ for } i=1\dots 100\}$, perhaps with a few values modified. Poor algorithms will fail.

[\[edit\]](#) Currency correlation

The term "currency correlation" refers to the correlation between between two [currency pairs](#), or more generally, correlations between values of [commodities](#), [stocks](#) and [bonds markets](#). It is used as a tool to predict changes in market value.

[\[edit\]](#) See also

- [Autocorrelation](#)
- [Coefficient of determination](#)

- [Fraction of variance unexplained](#)
- [Kendall's tau](#)
- [Pearson product-moment correlation coefficient](#)
- [Point-biserial correlation coefficient](#)
- [Partial correlation](#)
- [Spearman's rank correlation coefficient](#)
- [Statistical arbitrage](#)

[[edit](#)] Notes and references

1. [^] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates. [ISBN 0-8058-0283-5](#).
2. [^] Anscombe, Francis J. (1973) Graphs in statistical analysis. *American Statistician*, 27, 17–21.

[[edit](#)] Further reading

- Cohen, J., Cohen P., West, S.G., & Aiken, L.S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. (3rd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates.
- Abdi, H (2007). "Coefficients of correlation, alienation and determination.", in N.J. Salkind (ed.): [Encyclopedia of Measurement and Statistics](#). Thousand Oaks, CA: Sage.

[[edit](#)] External links

- [Understanding Correlation](#) - Introductory material by a U. of Hawaii Prof.
- [Statsoft Electronic Textbook](#)
- [Pearson's Correlation Coefficient](#) - How to calculate it quickly
- [Learning by Simulations](#) - The distribution of the correlation coefficient
- [CorrMatr.c](#) simple program for correlation matrix calculation
- [Correlation measures the strength of a *linear* relationship between two variables.](#)

Coefficient of determination

From Wikipedia, the free encyclopedia

Jump to: [navigation](#), [search](#)

In [statistics](#), the **coefficient of determination** R^2 is the proportion of variability in a data set that is accounted for by a statistical model. In this definition, the term "variability" is defined as the [sum of squares](#). There are equivalent expressions for R^2 . The version most common in [statistics](#) texts is based on an analysis of variance decomposition as follows:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}.$$

In the above definition,

$$SS_T = \sum_i (y_i - \bar{y})^2, SS_R = \sum_i (\hat{y}_i - \bar{y})^2, SS_E = \sum_i (y_i - \hat{y}_i)^2.$$

That is, SS_T is the [total sum of squares](#), SS_R is the regression sum of squares, and SS_E is the sum of squared errors. In some texts, the abbreviations SS_E and SS_R have the opposite meaning: SS_E stands for the [explained sum of squares](#) (which is another name for the regression sum of squares) and SS_R stands for the [residual sum of squares](#) (another name for the sum of squared errors).

R-square is the statistic that will give information about the goodness of fit of the model. It has a drawback: R-square increases as we increase the number of variables in the model (R-square will not decrease), so the alternative technique is to look for adjusted R-square. The explanation of this statistic is also same as R-square but it penalizes R-square by the number of variables used in the model.

Contents

[\[hide\]](#)

- [1 Explanation and interpretation of R2](#)
- [2 Inflation of R2](#)
- [3 Adjusted R2](#)
- [4 Notes on interpreting R2](#)
- [5 External links](#)
- [6 See also](#)

[\[edit\]](#) Explanation and interpretation of R^2

For expository purposes, consider a linear model of the form

$$Y_i = \beta_0 + \sum_j^p \beta_j X_{i,j} + \varepsilon_i,$$

where Y_i is the response variable, β_0, \dots, β_p are unknown coefficients; X_1, \dots, X_p are p regressors, and ε_i is a mean zero [error](#) term. The coefficient of determination R^2 is a measure of the global fit of the model. Specifically, R^2 is an element of $[0,1]$ and represents the proportion of variability in Y_i that may be attributed to some linear combination of the regressors ([explanatory variables](#)) in X .

More simply, R^2 is often interpreted as the proportion of response variation "explained" by the regressors in the model. Thus, $R^2 = 1$ indicates that the fitted model explains all variability in y , while $R^2 = 0$ indicates no 'linear' relationship between the response variable and regressors. An interior value such as $R^2 = 0.7$ may be interpreted as follows: "Approximately seventy percent of the variation in the response variable can be explained by the explanatory variable. The remaining thirty percent can be explained by unknown, [lurking variables](#) or inherent variability."

A caution that applies to R^2 , as to other statistical descriptions of [correlation](#) and association is that "[correlation does not imply causation](#)." In other words, while correlations may provide valuable clues regarding causal relationships among variables, a high correlation between two variables does not represent adequate evidence that changing one variable has resulted, or may result, from changes of other variables.

In case of a single regressor R^2 is the square of the [Pearson product-moment correlation coefficient](#) relating the regressor and the response variable.

[\[edit\]](#) Inflation of R^2

In [least squares](#) regression, R^2 is weakly increasing in the number of regressors in the model. As such, R^2 cannot be used as a meaningful comparison of models with different numbers of covariants. As a reminder of this, some authors denote R^2 by R^2_p , where p is the number of columns in X

Demonstration of this property is trivial. To begin, recall that the objective of least squares regression is (in matrix notation)

$$\min_b SS_E(b) \Rightarrow \min_b \sum_i (y_i - X_i b)^2$$

The optimal value of the objective is weakly smaller as additional columns of X are added, by the fact that relatively unconstrained minimization leads to a solution which is weakly smaller than relatively constrained minimization. Given the previous conclusion and noting that SS_T depends only on y , the non-decreasing property of R^2 follows directly from the definition above.

[\[edit\]](#) Adjusted R^2

Adjusted R^2 is a modification of R^2 that adjusts for the number of [explanatory](#) terms in a model. Unlike R^2 , the adjusted R^2 increases only if the new term improves the model more than would be expected by chance. The adjusted R^2 can be negative, and will always be less than or equal to R^2 . The adjusted R^2 is defined as

$$1 - (1 - R^2) \frac{n - 1}{n - p - 1}$$

where p is the total number of regressors in the linear model, and n is sample size.

Adjusted R^2 *does not have the same interpretation as R^2* . As such, care must be taken in interpreting and reporting this statistic. Adjusted R^2 is particularly useful in the [Feature selection](#) stage of model building.

Adjusted R^2 is not always *better* than R^2 : adjusted R^2 will be more useful only if the R^2 is calculated based on a sample, not the entire population. For example, if our [unit of analysis](#) is a [state](#), and we have data for all counties, then adjusted R^2 will not yield any more useful information than R^2 .

[\[edit\]](#) Notes on interpreting R^2

R^2 does *NOT* tell whether:

- the independent variables are a true cause of the changes in the [dependent variable](#)

- [omitted-variable bias](#) exists
- the correct regression was used; or
- the most appropriate set of independent variables has been chosen
- Co-linearity is present in the data

[[edit](#)] External links

- [Adjusted R-Square Calculator](#)
- [R-squared is an often-misused criterion for goodness-of-fit, where bigger isn't always better.](#)

[[edit](#)] See also

- [Correlation](#)
- [Fraction of variance unexplained](#)
- [Pearson product-moment correlation coefficient](#)

Retrieved from "http://en.wikipedia.org/wiki/Coefficient_of_determination"

Category: [Regression analysis](#)