



## Linear regression models

[Notes on linear regression analysis \(pdf\)](#)

[Introduction to linear regression analysis](#)

[Mathematics of simple regression](#)

[Regression examples](#)

- [Baseball batting averages](#)
- [Beer sales vs. price, part 1: descriptive analysis](#)
- [Beer sales vs. price, part 2: fitting a simple model](#)
- [Beer sales vs. price, part 3: transformations of variables](#)
- [Beer sales vs. price, part 4: additional predictors](#)
- [NC natural gas consumption vs. temperature](#)

[What to look for in regression output](#)

[What's a good value for R-squared?](#)

[What's the bottom line? How to compare models](#)

[Testing the assumptions of linear regression](#)

[Additional notes on regression analysis](#)

[Stepwise and all-possible-regressions](#)

[Excel file with simple regression formulas](#)

[Excel file with regression formulas in matrix form](#)

If you are a PC Excel user, you *must* check this out:

[RegressIt: free Excel add-in for linear regression and multivariate data analysis](#)

## Introduction to linear regression analysis

[History of regression](#)

[Justification for regression assumptions](#)

[Correlation and simple regression formulas](#)

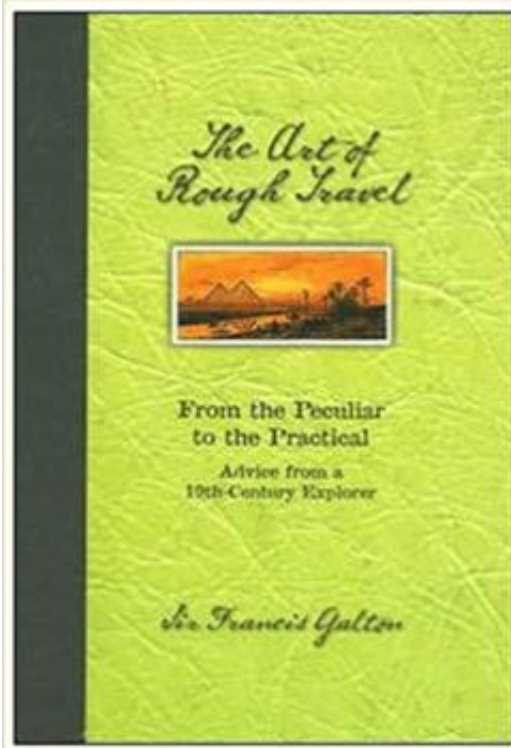
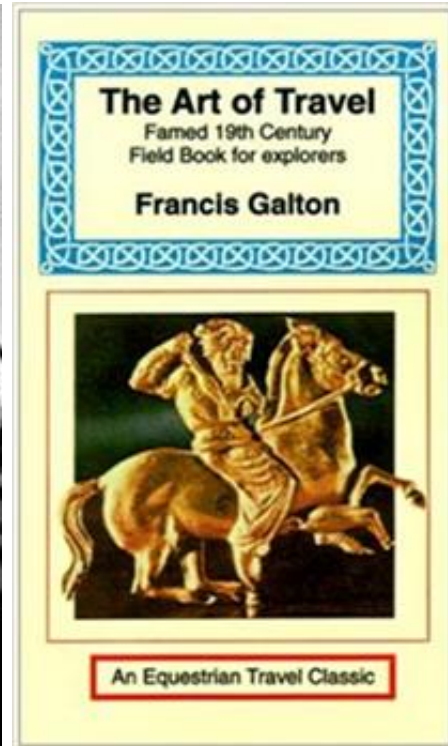
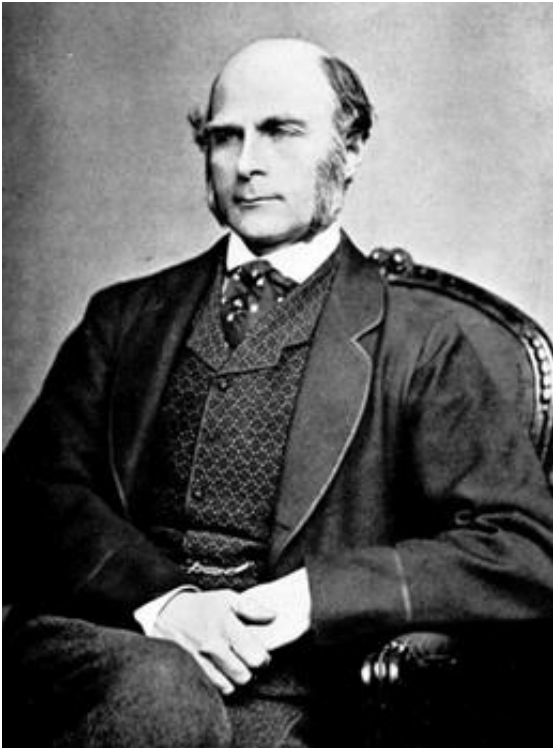
Linear regression analysis is the most widely used of all statistical techniques: it is the study of *linear*, *additive* relationships between variables. Let  $Y$  denote the “dependent” variable whose values you wish to predict, and let  $X_1, \dots, X_k$  denote the “independent” variables from which you wish to predict it, with the value of variable  $X_i$  in period  $t$  (or in row  $t$  of the data set) denoted by  $X_{it}$ . Then the equation for computing the predicted value of  $Y_t$  is:

$$\hat{Y}_t = b_0 + b_1 X_{1t} + b_2 X_{2t} + \dots + b_k X_{kt}$$

This formula has the property that the prediction for  $Y$  is a straight-line function of each of the  $X$  variables, holding the others fixed, and the contributions of different  $X$  variables to the predictions are additive. The slopes of their individual straight-line relationships with  $Y$  are the constants  $b_1, b_2, \dots, b_k$ , the so-called *coefficients* of the variables. That is,  $b_i$  is the change in the predicted value of  $Y$  per unit of change in  $X_i$ , other things being equal. The additional constant  $b_0$ , the so-called *intercept*, is the prediction that the model would make if all the  $X$ 's were zero (if that is possible). The coefficients and intercept are estimated by *least squares*, i.e., setting them equal to the unique values that minimize the sum of squared errors within the sample of data to which the model is fitted. And the model's prediction errors are typically assumed to be *independently and identically normally distributed*.

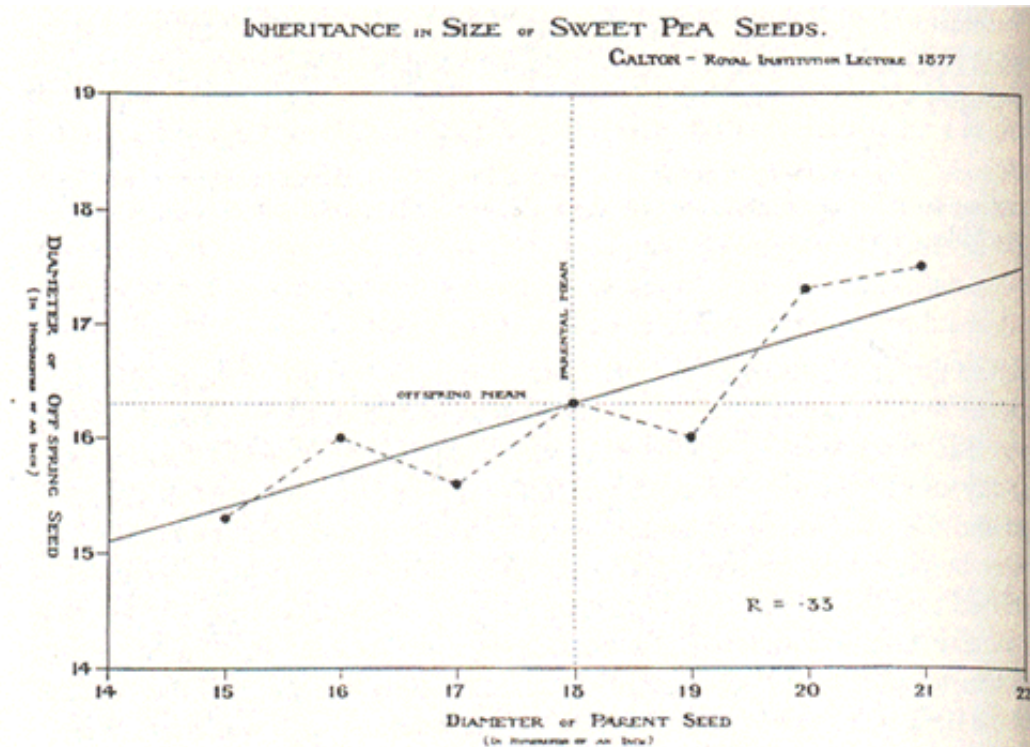
The first thing you ought to know about linear regression is how the strange term *regression* came to be applied to models like this. They were first studied in depth by a 19th-Century scientist, **Sir Francis Galton**. Galton was a self-taught naturalist, anthropologist, astronomer, and statistician--and a real-life

Indiana Jones character. He was famous for his explorations, and he wrote a best-selling book on how to survive in the wilderness entitled "The Art of Travel: Shifts and Contrivances Available in Wild Places," and its sequel, "The Art of *Rough* Travel: From the Practical to the Peculiar." They are still in print and still considered as useful resources. They provide many handy hints for staying alive--such as how to treat spear wounds or extract your horse from quicksand--and introduced the concept of the sleeping bag to the Western World. Click on these pictures for more details:



Galton was a pioneer in the application of statistical methods to measurements in many branches of science, and in studying data on relative sizes of parents and their offspring in various species of plants and animals, he observed the following phenomenon: a larger-than-average parent tends to produce a larger-than-average child, but the child is likely to be *less* large than the parent in terms of its relative position within its *own* generation. Thus, for example, if the parent's size is  $x$  standard deviations from the

mean within its own generation, then you should predict that the child's size will be  $rx$  ( $r$  times  $x$ ) standard deviations from the mean within the set of children of those parents, where  $r$  is a number *less than 1 in magnitude*. ( $r$  is what will be defined below as the *correlation* between the size of the parent and the size of the child.) The same is true of virtually *any* physical measurement (and in the case of humans, most measurements of cognitive and physical ability) that can be performed on parents and their offspring. Here is the first published picture of a regression line illustrating this effect, from a lecture presented by Galton in 1877:



The  $R$  symbol on this chart (whose value is 0.33) denotes the slope coefficient, not the correlation, although the two are the same if both populations have the same standard deviation, as will be shown below.

Galton termed this phenomenon a [regression towards mediocrity](#), which in modern terms is a **regression to the mean**. To a naïve observer this might suggest that later generations are going to exhibit less variability--literally more mediocrity--than earlier ones, but that is not case. It is a purely statistical phenomenon. Unless every child is *exactly* as the same size as the parent in relative terms (i.e., unless the correlation is exactly equal to 1), the predictions *must* regress to the mean regardless of biology if mean squared error is to be minimized. [\(Return to top of page.\)](#)

Regression to the mean is an inescapable fact of life. Your children can be *expected* to be less exceptional (for better or worse) than you are. Your score on a final exam in a course can be *expected* to be less good (or bad) than your score on the midterm exam, relative to the rest of the class. A baseball player's batting average in the second half of the season can be *expected* to be closer to the mean (for all players) than his batting average in the first half of the season. And so on. The key word here is "expected." This does not mean it's *certain* that regression to the mean will occur, but that's the way to bet!

We have already seen a suggestion of regression-to-the-mean in some of the time series forecasting models we have studied: plots of forecasts tend to be *smoother*--i.e., they exhibit less variability--than the plots of the original data. This is not true of random walk models, but it is generally true of moving-average models and other models that base their forecasts on more than one past observation.

The intuitive explanation for the regression effect is simple: the thing we are trying to predict usually consists of a predictable component ("signal") and a statistically independent *unpredictable* component ("noise"). The best we can hope to do is to predict (only) that part of the variability which is due to the



signal. Hence our forecasts will tend to exhibit less variability than the actual values, which implies a regression to the mean.

Another way to think of the regression effect is in terms of *selection bias*. In general a player's performance over any given period of time can be attributed to a combination of skill and luck. Suppose that we select a sample of professional athletes whose performance was much better than average (or students whose grades were much better than average) in the first half of the year. The fact that they did so well in the first half of the year makes it probable that *both* their skill and their luck were better than average during that period. In the second half of the year we may expect them to be equally skillful, but we should not expect them to be equally lucky. So we should predict that in the second half their performance will be closer to the mean. Meanwhile, players whose performance was merely average in the first half probably had skill and luck working in opposite directions for them. We should therefore expect their performance in the second half to move away from the mean in one direction or another, as we get another independent test of their skill. We don't know *which* direction they will move, though, so even for them we should predict that their second half performance will be closer to the mean than their first half performance. However, the *actual* performance of the players should be expected to have an *equally large variance* in the second half of the year as in the first half, because it merely results from a redistribution of independently random luck among players with the same distribution of skill as before.

A nice discussion of regression to the mean in the broader context of social science research can be found [here](#). [\(Return to top of page.\)](#)

---

## Justification for regression assumptions

Why should we assume that relationships between variables are **linear**?

1. Because linear relationships are the *simplest non-trivial relationships* that can be imagined (hence the easiest to work with), and.....
2. Because the "true" relationships between our variables are often at least *approximately* linear over the range of values that are of interest to us, and...
3. Even if they're not, we can often *transform* the variables in such a way as to linearize the relationships.

This is a strong assumption, and the first step in regression modeling should be to look at scatterplots of the variables (and in the case of time series data, plots of the variables vs. time), to make sure it is reasonable a priori. And after fitting a model, plots of the errors should be studied to see if there are unexplained nonlinear patterns. This is especially important when the goal is to make predictions for scenarios outside the range of the historical data, where departures from perfect linearity are likely to have the biggest effect. If you see evidence of nonlinear relationships, it is possible (though not guaranteed) that transformations of variables will straighten them out in a way that will yield useful inferences and predictions via linear regression. [\(Return to top of page.\)](#)

---

And why should we assume that the effects of different independent variables on the expected value of the dependent variable are **additive**? This is a very strong assumption, stronger than most people realize. It implies that the marginal effect of one independent variable (i.e., its slope coefficient) does not depend on the current values of other independent variables. But... why shouldn't it? It's conceivable that one independent variable could amplify the effect of another, or that its effect might vary systematically over time. In a multiple regression model, the estimated coefficient of a given independent variable supposedly measures its effect while "controlling" for the presence of the others. However, the way in which controlling is performed is extremely simplistic: *multiples of other variables are merely added or subtracted*.

Many users just throw a lot of independent variables into the model without thinking carefully about this issue, as if their software will automatically figure out exactly how they are related. It won't! Even

automatic model-selection methods (e.g., stepwise regression) require you to have a good understanding of your own data and to use a guiding hand in the analysis. They work only with the variables they are given, in the form that they are given, and then they look only for linear, additive patterns among them in the context of each other. **A regression model does not merely assume that Y is "some function" of the X's. It assumes that it is a very special kind of function of the X's.** A common practice is to include independent variables whose predictive effects logically cannot be additive, say, some that are totals and others that are rates or percentages. Sometimes this can be rationalized by local first-order-approximation arguments, and sometimes it can't.

You need to collect the relevant data, understand what it measures, clean it up if necessary, perform descriptive analysis to look for patterns before fitting any models, and study the diagnostic tests of model assumptions afterward, especially statistics and plots of the errors. You should also try to apply the appropriate economic or physical reasoning to determine whether an additive prediction equation makes sense. Here too, it is possible (but not guaranteed) that transformations of variables or the inclusion of interaction terms might separate their effects into an additive form, if they do not have such a form to begin with, but this requires some thought and effort on your part. ([Return to top of page.](#))

---

And why should we assume the *errors* of linear models are ***independently and identically normally distributed***?

1. This assumption is often justified by appeal to the **Central Limit Theorem** of statistics, which states that the *sum or average* of a sufficiently large number of independent random variables--whatever their individual distributions--approaches a normal distribution. Much data in business and economics and engineering and the natural sciences is obtained by adding or averaging numerical measurements performed on many different persons or products or locations or time intervals. Insofar as the activities that generate the measurements may occur somewhat randomly and somewhat independently, we might expect the variations in the totals or averages to be somewhat normally distributed.
2. It is (again) mathematically convenient: it implies that the optimal coefficient estimates for a linear model are those that minimize the *mean squared error* (which are easily calculated), and it justifies the use of a host of statistical tests based on the normal family of distributions. (This family includes the t distribution, the F distribution, and the Chi-square distribution.)
3. Even if the "true" error process is not normal in terms of the original units of the data, it may be possible to transform the data so that your model's prediction errors are approximately normal.

But here too caution must be exercised. Even if the unexplained variations in the dependent variable are approximately normally distributed, it is not guaranteed that they will also be *identically* normally distributed for all values of the independent variables. Perhaps the unexplained variations are larger under some conditions than others, a condition known as "heteroscedasticity". For example, if the dependent variable consists of daily or monthly total sales, there are probably significant day-of-week patterns or seasonal patterns. In such cases the variance of the total will be larger on days or in seasons with greater business activity--another consequence of the central limit theorem. (Variable transformations such as logging and/or seasonal adjustment are often used to deal with this problem.) It is also not guaranteed that the random variations will be statistically independent. This is an especially important question when the data consists of *time series*: if the model is not correctly specified, it is possible that consecutive errors (or errors separated by some other number of periods) will have a systematic tendency to have the same sign or a systematic tendency to have opposite signs, a phenomenon known as "autocorrelation" or "serial correlation".

A very important special case is that of **stock price data**, in which percentage changes rather than absolute changes tend to be normally distributed. This implies that over moderate to large time scales, movements in stock prices are *lognormally* distributed rather than normally distributed. A [log transformation](#) is typically applied to historical stock price data when studying growth and volatility. Caution: although simple regression models are often fitted to historical stock returns to estimate "betas", which are indicators of relative risk in the context of a diversified portfolio, I do not recommend that you use regression to try to predict *future* stock returns. See the [geometric random walk](#) page instead.

You still might think that variations in the values of *portfolios* of stocks would tend to be normally distributed, by virtue of the central limit theorem, but the central limit theorem is actually rather slow to bite on the lognormal distribution because it is so asymmetrically long-tailed. **A sum of 10 or 20 independently and identically lognormally distributed variables has a distribution that is still quite close to lognormal.** If you don't believe this, try testing it by Monte Carlo simulation: you'll be surprised. (I was.)

---

Because the assumptions of linear regression (linear, additive relationships with i.i.d. normally distributed errors) are so strong, it is very important to test their validity when fitting models, a topic discussed in more detail on [the testing-model-assumptions page](#), and be alert to the possibility that you may need more or better data to accomplish your objectives. You can't get something from nothing. All too often, naïve users of regression analysis view it as a black box that can automatically predict any given variable from any other variables that are fed into it, when in fact a regression model is a very special and very transparent kind of prediction box. Its output contains no more information than is provided by its inputs, and its inner mechanism needs to be compared with reality in each situation where it is applied. ([Return to top of page.](#))

---

## Correlation and simple regression formulas

A **variable** is, by definition, a *quantity that may vary* from one measurement to another in situations where different samples are taken from a population or observations are made at different points in time. In fitting statistical models in which some variables are used to predict others, what we hope to find is that the different variables do *not* vary *independently* (in a statistical sense), but that they tend to vary *together*.

In particular, when fitting *linear* models, we hope to find that one variable (say,  $Y$ ) is varying as a *straight-line* function of another variable (say,  $X$ ). In other words, if all other possibly-relevant variables could be held fixed, we would hope to find the *graph* of  $Y$  versus  $X$  to be a straight line (apart from the inevitable random errors or "noise").

A measure of the absolute amount of variability in a variable is (naturally) its **variance**, which is defined as its *average squared deviation from its own mean*. Equivalently, we can measure variability in terms of the **standard deviation**, which is defined as the square root of the variance. The standard deviation has the advantage that it is measured in the same units as the original variable, rather than squared units.

Our task in predicting  $Y$  might be described as that of explaining some or all of its variance--i.e., *why*, or under what conditions, does it deviate from its mean? Why is it not constant? That is, we would like to be able to improve on the naïve predictive model:  $\hat{Y}_t = \text{CONSTANT}$ , in which the best value for the constant is presumably the historical mean of  $Y$ . More precisely, *we hope to find a model whose prediction errors are smaller, in a mean square sense, than the deviations of the original variable from its mean.*

---

In using *linear* models for prediction, it turns out very conveniently that the *only* statistics of interest (at least for purposes of estimating coefficients to minimize squared error) are the mean and variance of each variable and the **correlation coefficient** between each pair of variables. The coefficient of correlation between  $X$  and  $Y$  is commonly denoted by  $r_{XY}$ , and it measures the strength of the linear relationship between them on a relative (i.e., unitless) scale of -1 to +1. That is, it measures the extent to which a linear model can be used to predict the deviation of one variable from its mean given knowledge of the other's deviation from its mean at the same point in time.

The correlation coefficient is most easily computed if we first **standardize** the variables, which means to convert them to units of standard-deviations-from-the-mean, using the *population* standard deviation rather than the sample standard deviation, i.e., using the statistic whose formula has  $n$  rather than  $n-1$  in the denominator, where  $n$  is the sample size. The standardized version of  $X$  will be denoted here by  $X^*$ , and its value in period  $t$  is defined in Excel notation as:

$$X_t^* = \frac{X_t - \text{AVERAGE}(X)}{\text{STDEV.P}(X)}$$

...where **STDEV.P** is the Excel function for the population standard deviation. (Here and elsewhere I am going to use Excel functions rather than conventional math symbols in some of the formulas to illustrate how the calculations would be done on a spreadsheet.) For example, suppose that  $\text{AVERAGE}(X) = 20$  and  $\text{STDEV.P}(X) = 5$ . If  $X_t = 25$ , then  $X_t^* = 1$ , if  $X_t = 10$ , then  $X_t^* = -2$ , and so on.  $Y^*$  will denote the similarly standardized value of  $Y$ .

Now, **the correlation coefficient is equal to the average product of the standardized values** of the two variables within the given sample of  $n$  observations:

$$r_{XY} = (X_1^* Y_1^* + X_2^* Y_2^* + \dots + X_n^* Y_n^*) / n$$

Thus, for example, if  $X$  and  $Y$  are stored in columns on a spreadsheet, you can use the **AVERAGE** and **STDEV.P** functions to compute their averages and population standard deviations, then you can create two new columns in which the values of  $X^*$  and  $Y^*$  in each row are computed according to the formula above. Then create a third new column in which  $X^*$  is multiplied by  $Y^*$  in every row. The average of the values in the last column is the correlation between  $X$  and  $Y$ . Of course, in Excel, you can just use the formula **=CORREL(X,Y)** to calculate a correlation coefficient, where  $X$  and  $Y$  denote the cell ranges of the data for the variables. (Note: in some situations it might be of interest to standardize the data relative to the *sample* standard deviation, which is **STDEV.S** in Excel, but the population statistic is the correct one to use in the formula above.) [\(Return to top of page.\)](#)

If the two variables tend to vary on the *same sides* of their respective means at the same time, then the average product of their deviations (and hence the correlation between them) will be *positive*, since the product of two numbers with the same sign is positive. Conversely, if they tend to vary on *opposite* sides of their respective means at the same time, their correlation will be *negative*. If they vary *independently* with respect to their means--that is, if one is equally likely to be above or below its mean regardless of what the other is doing--then the correlation will be *zero*. And if  $Y$  is an exact linear function of  $X$ , then either  $Y_t^* = X_t^*$  for all  $t$  or else  $Y_t^* = -X_t^*$  for all  $t$ , in which case the formula for the correlation reduces to  $+1$  or  $-1$ .

The correlation coefficient can be said to measure the strength of the *linear* relationship between  $Y$  and  $X$  for the following reason. The linear equation for predicting  $Y^*$  from  $X^*$  that *minimizes mean squared error* is simply:

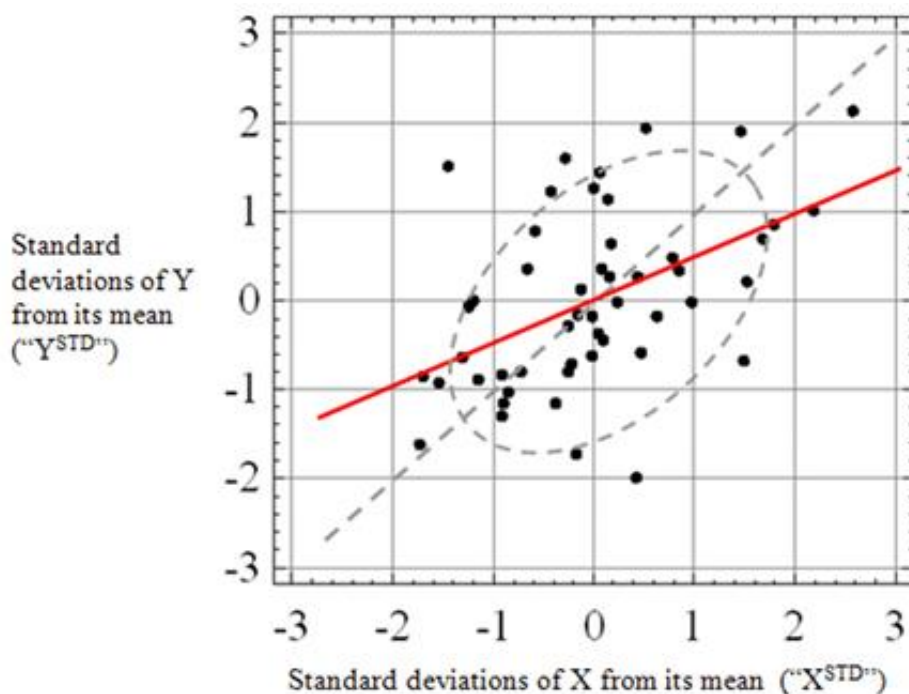
$$\hat{Y}_t^* = r_{XY} X_t^*$$

Thus, if  $X$  is observed to be 1 standard deviation above its own mean, then we should predict that  $Y$  will be  $r_{XY}$  standard deviations above its own mean; if  $X$  is 2 standard deviations below its own mean, then we should be predict that  $Y$  will be  $2r_{XY}$  standard deviations below its own mean, and so on.

In graphical terms, this means that, **on a scatterplot of  $Y^*$  versus  $X^*$ , the line for predicting  $Y^*$  from  $X^*$  so as to minimize mean squared error is the line that passes through the origin and has slope  $r_{XY}$** . This fact is not supposed to be obvious, but it is easily proved by elementary differential calculus.

Here is an example: on a scatterplot of  $Y^*$  versus  $X^*$ , the visual axis of symmetry is a line that passes through the origin and whose slope is equal to 1 (i.e., a 45-degree line), which is the gray dashed line on

the plot below. It passes through the origin because the means of both standardized variables are zero, and its slope is equal to 1 because their standard deviations are both equal to 1. (The latter fact means that the points are equally spread out horizontally and vertically in terms of mean squared deviations from zero, which forces their pattern to appear roughly symmetric around the 45-degree line if the relationship between the variables really is linear.) However, the gray dashed line is not the best line to use for predicting the value of  $Y^*$  for a given value of  $X^*$ . The best line for predicting  $Y^*$  from  $X^*$  has a slope of less than 1: *it regresses toward the X axis*. The regression line is shown in red, and its slope is the correlation between X and Y, which is 0.46 in this case. Why is this true? Because, that's the way to bet if you want to minimize the mean squared error *measured in the Y direction*. If instead you wanted to predict  $X^*$  from  $Y^*$  so as to minimize mean squared error measured in the X direction, the line would regress in the other direction relative to the 45-degree line, and by exactly the same amount.



[\(Return to top of page.\)](#)

If we want to obtain the linear regression equation for predicting Y from X in *unstandardized terms*, we just need to substitute the formulas for the standardized values in the preceding equation, which then becomes:

$$\frac{\hat{Y}_t - \text{AVERAGE}(Y)}{\text{STDEV.P}(Y)} = r_{XY} \times \frac{X_t - \text{AVERAGE}(X)}{\text{STDEV.P}(X)}$$

By rearranging this equation and collecting constant terms, we obtain:

$$\hat{Y}_t = b_0 + b_1 X_t$$

where:

$$b_1 = r_{XY} \frac{\text{STDEV.P}(Y)}{\text{STDEV.P}(X)}$$

is the estimated slope of the regression line, and

$$b_0 = \text{AVERAGE}(Y) - b_1 \times \text{AVERAGE}(X)$$



is the estimated Y-intercept of the line.

Notice that, as we claimed earlier, the coefficients in the linear equation for predicting Y from X depend only on the means and standard deviations of X and Y and on their coefficient of correlation.

The additional formulas that are needed to compute *standard errors*, *t-statistics*, and *P-values* (statistics that measure the precision and significance of the estimated coefficients) are given in the [notes on mathematics of simple regression](#) and also illustrated in [this spreadsheet file](#).

Perfect positive correlation ( $r_{XY} = +1$ ) or perfect negative correlation ( $r_{XY} = -1$ ) is only obtained if one variable is an exact linear function of the other, without error, in which case they aren't really "different" variables at all.

In general we find less-than-perfect correlation, which is to say, we find that  $r_{XY}$  is less than 1 in absolute value. Therefore our prediction for  $Y^*$  is typically *smaller* in absolute value than our observed value for  $X^*$ . That is, **the prediction for Y is always closer to its own mean, in units of its own standard deviation, than X was observed to be, which is Galton's phenomenon of regression to the mean.**

So, the technical explanation of the regression-to-the-mean effect hinges on two mathematical facts: (i) the correlation coefficient, calculated in the manner described above, happens to be the coefficient that minimizes the squared error in predicting  $Y^*$  from  $X^*$ , and (ii) the correlation coefficient is never larger than 1 in absolute value, and it is only equal to 1 when  $Y^*$  is an exact (noiseless) linear function of  $X^*$ .

The term "regression" has stuck and has even mutated from an intransitive verb into a transitive one since Galton's time. We don't merely say that the predictions for Y "regress to the mean"--we now say that we are "regressing Y on X" when we estimate a linear equation for predicting Y from X, and we refer to X as a "regressor" in this case.

When we have fitted a linear regression model, we can compute the variance of its errors and compare this to the variance of the dependent variable (the latter being the error variance of an intercept-only model). The relative amount by which the regression model's error variance is less than the variance of the dependent variable is referred to as the *fraction* of the variance that was *explained* by the independent variable(s). For example, if the error variance is 20% less than the original variance, we say we have "explained 20% of the variance."

It turns out that **in a simple regression model, the fraction of variance explained is precisely the square of the correlation coefficient**--i.e., the square of  $r$ . Hence, the fraction-of-variance-explained has come to be known as "**R-squared**". The interpretation and use of R-squared are discussed in more detail [here](#).

In a *multiple* regression model (one with two or more X variables), there are many correlation coefficients that must be computed, in addition to all the means and variances. For example, we must consider the correlation between *each* X variable and the Y variable, and also the correlation between each *pair* of X variables. In this case, it still turns out that the model coefficients and the fraction-of-variance-explained statistic can be computed entirely from knowledge of the means, standard deviations, and correlation coefficients among the variables--but the computations are no longer easy. We will leave those details to the computer. [\(Return to top of page.\)](#)

Go on to a nearby topic:

· [Mathematics of simple regression](#)

- [Example #1: baseball batting averages](#)
  - [What to look for in regression output](#)
-