



Notes on linear regression analysis

Robert Nau

Fuqua School of Business, Duke University

1. Introduction to linear regression
2. Correlation and regression-to-mediocrity
3. The simple regression model (formulas)
4. Take-aways

1. Introduction to linear regression

Regression analysis is the art and science of fitting straight lines to patterns of data. In a linear regression model, the variable of interest (the so-called “dependent” variable) is predicted from k other variables (the so-called “independent” variables) using a linear equation. If Y denotes the dependent variable, and X_1, \dots, X_k , are the independent variables, then the assumption is that the value of Y at time t (or row t) in the data sample is determined by the linear equation

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \dots + \beta_k X_{kt} + \varepsilon_t$$

where the betas are constants and the epsilons are independent and identically distributed (i.i.d.) normal random variables with mean zero (the “noise” in the system). β_0 is the so-called *intercept* of the model—the expected value of Y when all the X ’s are zero—and β_i is the *coefficient* (multiplier) of the variable X_i . The betas together with the mean and standard deviation of the epsilons are the *parameters* of the model. The corresponding equation for *predicting* Y_t from the corresponding values of the X ’s is therefore

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

where the b ’s are estimates of the betas obtained by least-squares, i.e., minimizing the squared prediction error within the sample. This is about the simplest possible model for predicting one variable from a group of others, and it rests on the following assumptions.

1. The expected value of Y is a *linear function* of the X variables. This means:
 - a. if X_i changes by an amount ΔX_i , holding other variables fixed, then the expected value of Y changes by a *proportional* amount $\beta_i \Delta X_i$, for some constant β_i (which in general could be a positive or negative number).
 - b. The value of β_i is always the same, regardless of values of the other X ’s.
 - c. The total effect of the X ’s on the expected value of Y is the *sum* of their separate effects.

(c) 2014 by Robert Nau, all rights reserved. Last updated on 11/26/2014.

The material presented in this handout can also be found in html form on the [introduction-to-linear-regression page](#) and the [mathematics-of-simple-regression page](#) on the main web site: people.duke.edu/~rnau/411/home.htm

2. The unexplained variations of Y are *independent* random variables (in particular, not “autocorrelated” if the variables are time series)
3. They all have the *same variance* (“homoscedasticity”).
4. They are *normally distributed*.

These are strong assumptions. You can easily imagine situations in which Y might be a *nonlinear* function of the X's (e.g., if there are diminishing marginal effects), or in which there might be *interactions* among the X's in their effects on Y (e.g., if the sensitivity of Y to one of the X's depends on the values of other X's), or in which the size of the random deviations of Y from its expected value might depend on the values of the X's (e.g., if there is greater or lesser uncertainty under some conditions), or in which the random deviations might be correlated in time, or in which the errors are not normally distributed (e.g., the error distribution might not be bell-shaped and/or might have some really extreme values). A regression model assumes that there are no such nonlinearities or interactions or changing volatility or autocorrelation or non-normality in the random variations. (Further discussion of the assumptions of regression models and how to test them are given on the [introduction-to-regression](#) web page and the [testing-model-assumptions](#) page on the main web site.)

Of course, no model is perfect—these assumptions will never be *exactly* satisfied by real-world messy data—but you hope that they are not badly wrong. Just to be clear about all this:

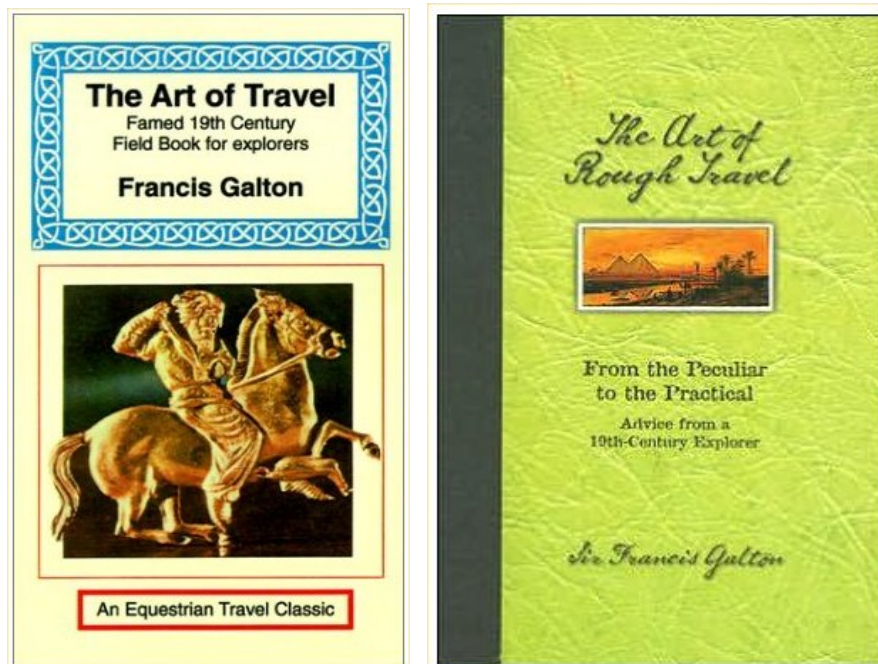
A regression model does not assume that Y merely “depends in some way” on the X's. If you have a variable Y that you wish to predict, and you have some other variables X_1 , X_2 , etc., that you believe have “some sort of effect” on Y or “some sort of predictive value” with respect to future values of Y, this does NOT suffice to justify using a linear regression model to predict Y from the X's. The regression model makes very strong assumptions about the WAY in which Y depends on the X's, namely that the causal or predictive effects of the X's with respect to Y are *linear* and *additive* and *non-interactive* and that any variations in Y that are not explained by the X's are *statistically independent* of each other and *identically normally distributed* under all conditions.

The *art* of regression modeling is to (most importantly!) collect data that is relevant and informative with respect to your decision or inference problem, and then define your variables and construct your model in such a way that the assumptions listed above are plausible, at least as a first-order approximation to what is really happening. There is no magic formula for doing this—you need to exercise your own judgment based on your own understanding of the situation and your own understanding of how a regression model works. Choosing a good regression model requires (a) gathering useful data and making sure you know where it came from and how it was measured, (b) performing descriptive analysis on it to understand its general patterns and to spot data-quality problems, (c) applying appropriate data transformations if you see strong evidence of relationships that are nonlinear or noise that is non-normal or time-dependent, (d) fitting and refining and comparing models, (e) checking to see whether a given model's assumptions are reasonably well satisfied or whether an alternative model is suggested, (f) choosing among reasonable models based on the appropriate bottom-line accuracy measure, and (g) deriving some useful insights from the whole process. What story does the model tell about the data, does it make sense to you, is it useful, and can you explain (or sell) it to someone else?

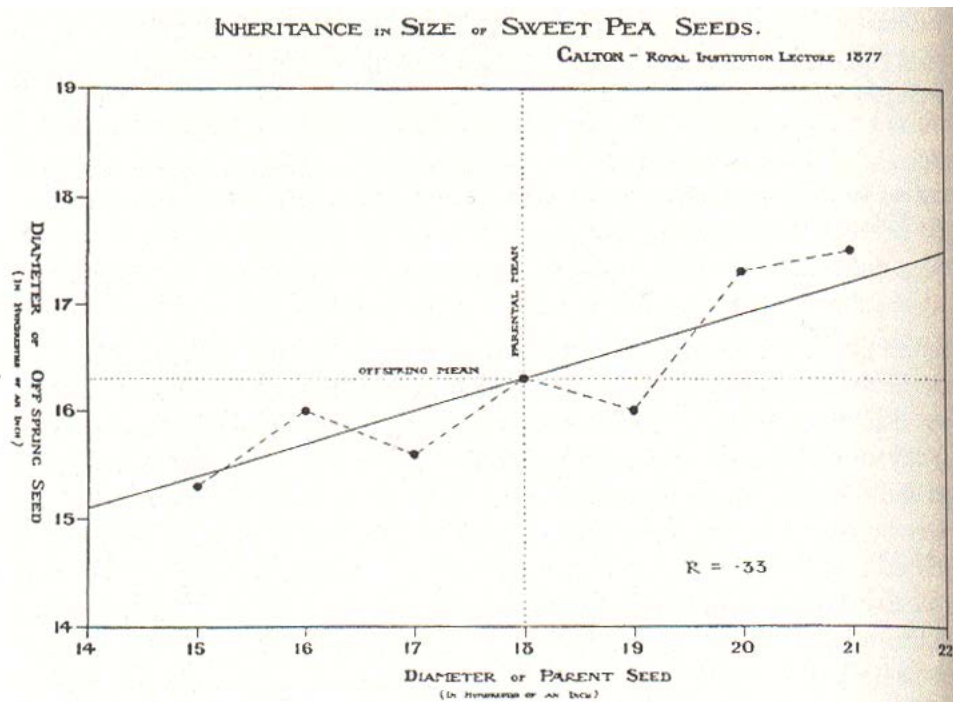
In decision-making settings the dependent variable might be some bottom-line measure of revenue or cost or productivity and the independent variables might be things that you control (such as your own prices or the timing of promotions) or that you don't control (such as your competitors' prices or the unemployment rate or the timing of holidays). For example, in an agricultural decision problem, the question of interest might be how crop yields are affected by decisions about chemical treatments and by weather conditions during the growing season. The dependent variable might be the crop yield in bushels per acre and independent variables might be pounds of fertilizers and pesticides applied per acre and amounts of rainfall and average temperatures during the months of the growing season. In a marketing decision problem, the dependent variable might be units of a product sold per week and the independent variables might be numbers of discount coupons distributed, numbers of spot advertisements on TV, and number of in-store displays. In other settings the question of interest may be one of inference or general knowledge, e.g., determining whether one variable has any significant effect on or association with another variable, in order to test a theory or a conjecture or to justify a claim about positive or negative effects of some activity or product or medical treatment.

2. Correlation and regression-to-mediocrity

The use of regression models in statistical analysis was pioneered by (Sir) Francis Galton, a 19th Century scientist and explorer who might be considered a model for the Indiana Jones character of the movies. Early in his career, after he inherited a fortune and quit medical school, he went on two expeditions to Africa, the first to the upper Nile Valley and the second through parts of south-west Africa, sandwiched around 5 years of enjoying the sporting life. Based on what he learned from these adventures he wrote two best-selling books—*The Art of Travel* and its sequel, *The Art of Rough Travel*—which offered practical advice to future explorers on topics such as how to treat spear wounds and pull your horse out of quicksand, and he introduced a new item of camping gear to the Western world: the sleeping bag. These authoritative books are still in print and you can order them from Amazon:



Galton went on to become a pioneer in the collection & analysis of biometric, anthropometric & psychometric data, as well as the invention of weather maps and the application of scientific methods to everyday activities such as tea-brewing. He was a half-cousin of Charles Darwin, and they struck up a long friendship and correspondence (initiated by Darwin) after they discovered each others' work. Darwin's theory had a revolutionary impact on Galton and inspired him to conduct experiments in plant-breeding to test some of its implications. (In fact, he corrected some errors in Darwin's thinking about whether acquired traits can be passed on to offspring.) In his experiments, Galton discovered a phenomenon whose universality took a while to become appreciated: an individual that is exceptional within its own generation with respect to some measurable trait (e.g., size) has offspring who *on average* are exceptional in the same way, but *less so in relative terms* within their own generation. More precisely, if the measured trait of the parent is X standard deviations from the mean within its own generation, then the same trait of the offspring should be predicted to be $r \times X$ deviations from the mean of *its* own generation, where r is a number less than 1 in magnitude: the *coefficient of correlation* between the trait of the parent and that of the offspring.² Here is the first published picture of a regression line, from a lecture presented by Galton in 1877:



Galton termed this phenomenon a “[regression towards mediocrity](#),” which we now call a “regression to the mean.”³ At first glance it appears to imply that individual differences will be

² Galton was an empirical scientist rather than a mathematician, and the formal details of regression models as we know them today were more fully worked out by his disciple, Karl Pearson, and other statistical theorists of the early 20th Century.

³ The “R” on this chart is the estimated coefficient of the parent seed diameter in the regression, not the correlation. From my attempts to read the offspring diameter values off the Y axis, I get $r_{XY} \approx 0.9$, $STDEV.P(Y) \approx 0.75$, $STDEV.P(X) \approx 2$. The coefficient of X in a simple regression is given by the formula $r_{XY} \times (STDEV.P(Y)/STDEV.P(X))$, which is equal to $0.9 \times (0.75/2) \approx 0.338$ in this case.

smoothed out over generations, each one becoming more mediocre than the one before it, but this is not so. It is a purely statistical phenomenon that can be viewed as a form of selection bias. Every quantitative measurement is a combination of signal and noise. When a value above the mean is observed, it is probable that the value of the signal was above average *and* the value of the noise was above average. Now suppose there is some other quantity (say, some measurable trait of the offspring—not necessarily the same one) whose value depends only on the signal, not the noise, in the first quantity. Then a measurement of the second quantity should also be expected to be above the mean, but less so in relative terms, because only the above-average signal is passed on; the above-average noise is not. In fact, it is the independent noise in the second quantity that prevents variations from ever dying out over generations. For example, in the case of Galton’s sweet pea seeds, an unusually large parent seed probably occurred partly for genetic reasons (which were passed on to its offspring) and partly for environmental reasons (which were not), and the sizes of the offspring seeds would also have been affected to some extent by random forces unrelated to their parentage or simply unmeasured by the scientist.

Regression-to-mediocrity is a fact of life that you should come to terms with. If you take a course in which there is both a midterm and a final exam, then you should expect that your performance on the final exam will be more mediocre than your performance on the midterm, in the sense of being fewer standard deviations from the class mean, for better or worse. If you have a “hot hand” (or a cold one) in a game of pickup basketball, you should expect to have a less-hot (or less-cold) hand in the next one. The mutual fund managers who had hot hands over the last 5 years will probably have less-hot hands over the next 5 years. If you are exceptionally talented along some dimension (which probably you are, if you are reading these notes in the first place), you should expect that your children will also be talented *but less so than you are* along the same dimension, measured relative to their own cohort (sorry!).

This is not guaranteed, but *it’s the way to bet*. What it implies for forecasting is that **on a scatterplot of Y vs. X, the regression line for predicting Y from X is not the visual “axis of symmetry” of the joint distribution. Rather, it tilts toward the horizontal.**

If you *standardize* the values of Y and X by converting them to units of standard-deviations-from-the-mean, then the distribution of points will be roughly symmetric around a 45-degree line drawn through the origin, and the regression line for predicting Y from X will also pass through the origin (i.e., Y will be predicted to equal its own mean when X equals its own mean), but it will have a slope of less than 45 degrees, i.e., it will “regress” toward the X axis.

To make this argument precise, let X^* and Y^* denote the variables that are the standardized values of X and Y, whose observations in period t are defined as:

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

where STDEV.P(X) and STDEV.P(Y) are the *population* standard deviations.⁴ Mathematically, **the coefficient of correlation between X and Y is the average product of their standardized values,**⁵ i.e.,

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

where n is the sample size. Intuitively, the quantity determined by this equation should be positive if X and Y have a tendency to vary on the same sides of their respective means at the same time (i.e., if X_t^* and Y_t^* tend to have the same sign for any given t), and it should be negative if X and Y have a tendency to vary on opposite sides of their respective means at the same time (i.e., if X_t^* and Y_t^* tend to have opposite signs).

In general **the correlation coefficient is a number that lies somewhere between -1 and +1, where -1 indicates a perfect negative linear relationship, +1 indicates a perfect positive linear relationship, and zero indicates no linear relationship.**

Just to be clear: a correlation of zero between X and Y does not necessarily mean that there is no relationship, just that there is no *linear* relationship within the historical sample of data that is being analyzed, i.e., it means that an equation of the form $\hat{Y} = b_0 + b_1X$ is no better for predicting Y than the equation $\hat{Y} = b_0$ within that sample of data. For example, if Y is always exactly equal to X squared, and you look at a sample of data in which X is symmetrically distributed on both sides of zero, then the graph of Y versus X looks like a *parabola*, and the correlation between X and Y can be *zero* even though Y can be perfectly predicted from X—it just can't be perfectly predicted using a *linear* equation.

In Excel, correlations can be calculated using the CORREL function. For example, if X and Y are stored in the ranges A1:A100 and B1:B100, then their correlation is given by the formula =CORREL(A1:A100,B1:B100). RegressIt and programs such as Statgraphics and SPSS and Stata also have built-in functions for calculating correlations—in fact, they can compute all *pairwise* correlations between a specified group of variables, arranged in the form of a *correlation matrix*.

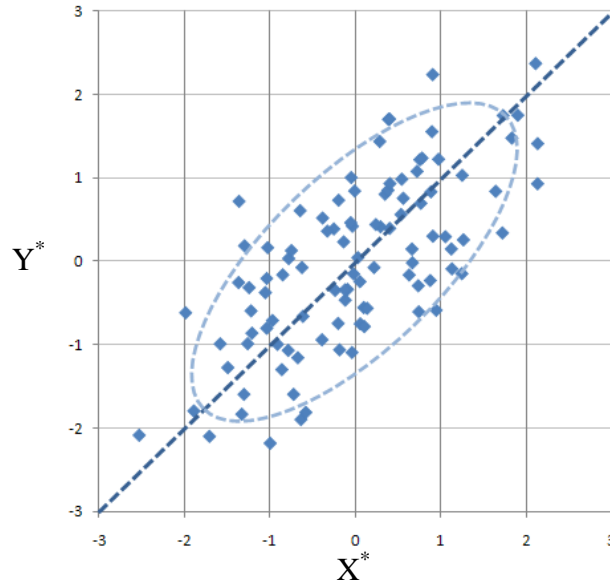
To continue the story, the phenomenon that Galton noted was that **the regression line for predicting Y^* from X^* is a line passes that through the origin and has a slope equal to the correlation between Y and X,** i.e., the regression equation in standardized units is:

⁴ The *population* standard deviation is the *actual* root-mean-square deviation of the variable from its mean value, which can be calculated when the data sample consists of the entire population and actual mean value is exactly known. The *sample* standard deviation, which is the function STDEV.S in Excel, is the *estimated* root-mean-square deviation of the variable from its mean value, when only a sample is observed and the mean value is estimated from the sample. The formula for the sample standard deviation includes a correction for the fact that the true mean is unknown. The two formulas are related by $STDEV.S = STDEV.P \times \sqrt{n/(n-1)}$ where n is the sample size. (In older version of Excel the sample and population standard functions were called STDEV and STDEVP.) Standardized values are sometimes computed relative to sample standard deviations, but the population standard deviations are the ones that are needed in the correlation calculation.

⁵ This is one of several equivalent mathematical formulas for computing the correlation coefficient.

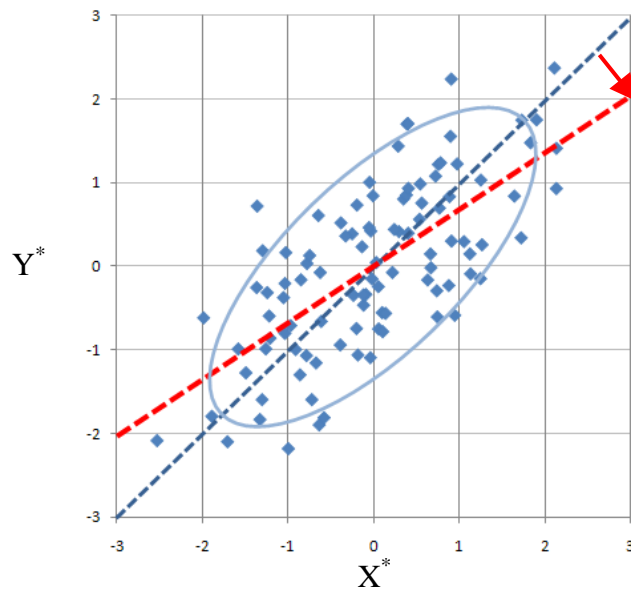
$$\hat{Y}^* = r_{XY}X^*$$

The plots below illustrate this equation. When the units of X and Y are standardized and both are also normally distributed, their values are distributed in an elliptical pattern that is symmetric around the 45-degree line, which has a slope equal to 1:



If X and Y are two jointly normally distributed random variables, then in *standardized units* (X^* , Y^*) their values are scattered in an elliptical pattern that is *symmetric around the 45-degree line*.

However, the *regression line* for predicting Y^* from X^* is NOT the 45-degree line. Rather, it is a line passing through the origin whose slope is r_{XY} —the dashed red line in the picture below—which is tilted toward the horizontal because the correlation is less than 1 in magnitude.⁶ In other words, it is a line that “regresses” (i.e., moves backward) toward the X-axis.



The *regression line* for predicting Y^* from X^* is NOT the 45-degree line. It has slope r_{XY} , which is less than 1. Hence it “regresses” toward the X-axis. For this data sample, $r_{XY} = 0.69$.

⁶ If Y and X are *negatively* correlated, the distribution of points on this chart will be symmetric around the *negative*-45-degree line, but the regression line will still tilt toward the horizontal, merely from below rather than from above.

This is actually true regardless of whether X and Y are normally distributed—the key assumption is that your goal is to minimize squared error in predicting Y from X. If Y is a time series, then the regression-to-the-mean effect implies that *the plot of the forecasts vs. time will look smoother (less variable) than the plot of the data versus time.*

What’s going on here is that the *mean* value of Y that is observed for any given value of X is the value around which the sum of squared deviations is minimized. (Recall that the mean of a sample of data has the property that it minimizes the sum of squared deviations from itself.) So, when our objective is to position the line so that its height at any point is the mean value of Y that should be expected for a given value of X, we want to orient the line so as to minimize the sum of squared errors measured *vertically* (i.e., the vertical distances of the points from the line). Now, here’s the interesting and important thing: **the sum of the squared errors measured vertically is minimized by tilting the prediction line toward the horizontal to some extent—in fact, tilting it precisely so that its slope is r_{XY} when the two variables are in standardized units.**

That’s the story of how the strange term “regression” came to be applied to linear statistical models. When we speak of “regressing” one variable on a group of others, we mean the fitting of a linear equation that minimizes the sum of squared errors in predicting that variable from the others.

3. Mathematics of the simple regression model

The remainder of this handout discusses the equations used to estimate parameters and calculate forecasts and confidence limits with simple regression models. Before reading it, you should have read the [handout on the mean model](#), because it lays the groundwork for the mathematics of regression models (although its key equations are restated below). The formulas are also demonstrated in the [simple regression Excel file](#) on the web site.

To set the stage for discussing the formulas used to fit a simple (one-variable) regression model, let’s briefly review the formulas for the mean model, which can be considered as an "intercept-only" (zero-variable) regression model. You can use regression software to fit this model and produce all of the standard table and chart output by merely not selecting any independent variables. R-squared will be zero in this case, because the mean model does not explain any of the variance in the dependent variable: it merely measures it.

The forecasting equation of the mean model is:

$$\hat{Y} = b_0$$

...where b_0 , the the least-squares estimate of β_0 , is

$$b_0 = \text{AVERAGE}(Y)$$

(The sample mean has the property that it is the least-squares estimate of the true mean.) Thus, every forecast is simply equal to the mean of the historical data. The error that the model makes for observation t is therefore the deviation of Y from its historical average value:

$$e_t = Y_t - \hat{Y}_t = Y_t - b_0$$

The **standard error of the model**, denoted by s , is *our estimate of the standard deviation of the true noise (unexplainable variation) in the data*. Smaller is better, other things being equal: we want the model to explain as much of the variation as possible. In the special case of the mean model, the standard error of the model is just the sample standard deviation of the data:

$$s = \sqrt{\frac{1}{n-1} \sum_{t=1}^n e_t^2} = \text{STDEV.S}(Y)$$

Note that the standard error of the model is *not* the square root of the *average* value of the squared errors within the historical sample of data. Rather, the sum of squared errors is divided by $n-1$ rather than n under the square root sign because this adjusts for the fact that a “degree of freedom for error” has been used up by estimating one model parameter (namely the mean) from the sample of n data points.

Now, the accuracy of the estimated mean is measured by the *standard error of the mean*, whose formula is:

$$SE_{mean} = \frac{s}{\sqrt{n}}$$

This is the estimated standard deviation of the error in estimating the mean. Notice that it is inversely proportional to the square root of the sample size, so it tends to go down as the sample size goes up. For example, if the sample size is increased by a factor of 4, the standard error of the mean goes down by a factor of 2, i.e., our estimate of the mean becomes twice as precise.

The accuracy of a *forecast* based on this model is measured by the *standard error of the forecast*, whose formula is

$$SE_{fcst} = \sqrt{s^2 + SE_{mean}^2}$$

This is the estimated standard deviation of the error in the forecast, which is not quite the same thing as the standard deviation of the unpredictable variations in the data (which is s). It takes into account both the unpredictable variations in Y and the error in estimating the mean.

The standard error of the forecast gets smaller as the sample size is increased, but only up to a point. More data yields a systematic reduction in the standard error of the mean, but it does not yield a systematic reduction in the standard error of the model. The standard error of the model will change to some extent if a larger sample is taken, due to sampling variation, but it could

equally well go up or down. The sort of variations in the data that were previously unexplainable remain unexplainable if we continue to believe in the model's assumptions, so the standard error of the model is always a lower bound on the standard error of the forecast.

Confidence intervals for the mean and for the forecast are equal to the forecast plus-or-minus the appropriate standard error multiplied by the appropriate 2-tailed critical value of the t distribution. The critical value that should be used depends on the number of degrees of freedom for error (the number data points minus number of parameters estimated, which is $n-1$ for this model) and the desired level of confidence. It can be computed in Excel using the T.INV.2T function. So, for example, a 95% confidence interval for the forecast is given by

$$b_0 \pm SE_{fcst} \times T.INV.2T(0.05, n-1)$$

In general, T.INV.2T(0.05, $n-1$) is fairly close to 2 except for very small samples, i.e., a 95% confidence interval for the forecast is roughly equal to the forecast plus-or-minus two standard errors. (In older versions of Excel, this function was just called TINV.)

Now, let's see how those formulas are generalized when we move from the mean model to a simple regression model...

A simple regression model is a regression model with a single independent variable, as in the pictures shown in the previous section, whose forecasting equation in real units is

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

It differs from the mean model merely by the addition of a multiple of X_t to the forecast. The estimated constant b_0 is the **Y-intercept** of the regression line (usually just called “the intercept”), which is the value that would be predicted for Y at $X = 0$. The estimated coefficient b_1 is the **slope** of the regression line, i.e., the *predicted change in Y per unit of change in X* . The simple regression model reduces to the mean model in the special case where the estimated slope is exactly zero. The estimated slope is almost never *exactly* zero (due to sampling variation), but if it is *not significantly different from zero* (as measured by its t -statistic), that suggests that the mean model should be preferred on grounds of simplicity.

Usually we do not care too much about the exact value of the intercept or whether *it* is significantly different from zero, unless we are really interested in what happens when X goes to “absolute zero” on whatever scale it is measured. Often X is a variable which logically can never go to zero, or even close to it, given the way it is defined. So, attention usually focuses mainly on the slope coefficient in the model, which measures the change in Y to be expected per unit of change in X as both variables move up or down relative to their historical mean values on their own natural scales of measurement.

The coefficients, standard errors, and forecasts for this model are given by the following equations.

The least-squares estimate of the slope coefficient (b_1) is equal to the correlation between X and Y times the ratio of the standard deviation of Y to the standard deviation of X:

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

The ratio of standard deviations on the RHS of this equation merely serves to scale the correlation coefficient appropriately for the real units in which the variables are measured.

The least-squares estimate of the intercept is the mean of Y minus the slope coefficient times the mean of X:

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

This equation implies *that Y must be predicted to be equal to its own average value whenever X is equal to its own average value.*⁷

The standard error of the model (denoted again by s) is usually referred to as the **standard error of the regression** (or sometimes the “standard error of the estimate”) in this context, and it is equal to **the square root of {the sum of squared errors divided by $n-2$ }**, or equivalently, the **standard deviation of the errors multiplied by the square root of $(n-1)/(n-2)$** , where the latter factor is a number slightly larger than 1:

$$s = \sqrt{\frac{1}{n-2} \sum_{t=1}^n e_t^2} = \text{STDEV.S}(\text{errors}) \times \text{SQRT}((n-1)/(n-2))$$

The sum of squared errors is divided by $n-2$ in this calculation rather than $n-1$ because an additional degree of freedom for error has been used up by estimating *two* parameters (a slope and an intercept) rather than only one (the mean) in fitting the model to the data. **The standard error of the regression is an unbiased estimate of the standard deviation of the true noise in the data, i.e., the variations in Y that are not explained by the model.**

⁷These formulas for b_0 and b_1 can be derived from the standardized regression equation $\hat{Y}_t^* = r_{XY} \times X_t^*$ by $X_t^* = (X_t - \text{AVERAGE}(X))/\text{STDEV.P}(X)$ and $\hat{Y}_t^* = (\hat{Y}_t - \text{AVERAGE}(Y))/\text{STDEV.P}(Y)$ and rearranging terms. The sample standard deviation, STDEV.S in Excel, can also be used in the equation for b_1 because it differs from STDEV.P only by a multiplicative factor that cancels between numerator and denominator.

Each of the two model parameters, the slope and intercept, has its own standard error, which is the estimated standard deviation of the error in estimating *it*. The standard error of the intercept is

$$SE_{b_0} = \frac{s}{\sqrt{n}} \times \sqrt{1 + \frac{(\text{AVERAGE}(X))^2}{\text{VARP}(X)}}$$

...which looks exactly like the formula for the standard error of the mean in the mean model, except for the additional term of $(\text{AVERAGE}(X))^2/\text{VARP}(X)$ under the square root sign.⁸ The standard error of the slope coefficient is given by:

$$SE_{b_1} = \frac{s}{\sqrt{n}} \times \frac{1}{\text{STDEV.P}(X)}$$

...which also looks very similar, except for the factor of $\text{STDEV.P}(X)$ in the denominator. Note that s is measured in units of Y and $\text{STDEV.P}(X)$ is measured in units of X , so SE_{b_1} is measured (necessarily) in “units of Y per unit of X ”, the same as b_1 itself. The terms in these equations that involve the variance or standard deviation of X mainly play a role in scaling the units of the coefficients and standard errors.

You don’t need to memorize all these equations, but there is one important thing to note: **the standard errors of the coefficients are directly proportional to the standard error of the regression and inversely proportional to the square root of the sample size.** This means that “noise” in the data (whose intensity if measured by s) tends to “fuzz up” all the coefficient estimates in exactly the same way, and it also means that 4 times as much data will tend to reduce the standard errors of the all coefficients by approximately a factor of 2, assuming the data is really all generated from the same model, and a really huge of amount of data will reduce them to zero.

However, more data will not systematically reduce the standard error of the *regression*. As with the mean model, variations that were considered inherently unexplainable before are still not going to be explainable with more of the same kind of data under the same model assumptions. *As the sample size gets larger, the standard error of the regression merely becomes a more accurate estimate of the standard deviation of the noise.*

⁸ Note that the *population* (.P) statistic rather than the sample (.S) statistic is used in the formulas for standard errors of coefficients, as well as in the formula for standardized values that was used to calculate the correlation coefficient earlier. In the rest of the equations of the regression model, the *sample* rather than the population statistics are used.

And now, introducing R-squared....

The fraction of the variance of Y that is “explained” by the simple regression model, i.e., the percentage by which the sample variance of the errors (“residuals”⁹) is less than the sample variance of Y itself, is equal to the square of the correlation between them, i.e., “R squared”:

$$\text{R-squared} = 1 - \frac{\text{VAR.S(errors)}}{\text{VAR.S(Y)}} = r_{XY}^2$$

Equivalently:

$$\text{VAR.S(errors)} = (1 \text{ minus R-squared}) \times \text{VAR.S(Y)}$$

Thus, for example, if the correlation is $r_{XY} = 0.5$, then $r_{XY}^2 = 0.25$, so the simple regression model explains 25% of the variance in Y in the sense that the sample variance of the errors of the simple regression model is 25% less than the sample variance of Y. This is not supposed to be obvious—it is a “strange but true” fact that can be proved with a little bit calculus.

By taking square roots everywhere, the same equation can be rewritten in terms of standard deviations to show that the standard deviation of the errors is equal to the standard deviation of Y multiplied by the *square root* of 1 minus the correlation squared:

$$\text{STDEV.S(errors)} = \text{SQRT}(1 \text{ minus R-squared}) \times \text{STDEV.S(Y)}$$

However, the sample variance and standard deviation of the errors are not *unbiased* estimates of the variance and standard deviation of the unexplained variations in the data, because they do not into account the fact that 2 degrees of freedom for error have been used up in the process of estimating the slope and intercept. The fraction by which the square of the standard error of the regression is less than the sample variance of Y (which is the fractional reduction in unexplained variation compared to using the mean model) is the “adjusted” R-squared of the model, and in a simple regression model it is given by the formula

$$\text{Adjusted R-squared} = 1 - \frac{n-1}{n-2} \times (1 \text{ minus R-squared}).$$

The factor of $(n-1)/(n-2)$ in this equation is the same adjustment for degrees of freedom that is made in calculating the standard error of the regression.¹⁰

⁹ The term “residuals” refers to the prediction errors made *within the sample of data to which the model was fitted*. Their statistical properties are not quite the same as those of the errors made in predicting outside the sample. The residuals tend to be a little smaller than the true random variations in the data because the coefficient estimates are adjusted precisely so as to minimize their sum of squared values (thereby reducing the degrees of freedom for error), while out-of-sample forecast errors tend to be a little larger than the true random variations due to errors in the coefficient estimates.

¹⁰ In a *multiple* regression model with k independent variables plus an intercept, the number of degrees of freedom for error is $n-(k+1)$, and the formulas for the standard error of the regression and adjusted R-squared remain the same except that the $n-2$ term is replaced by $n-(k+1)$.

In fact, adjusted R-squared can be used to determine the standard error of the regression from the sample standard deviation of Y in exactly the same way that R-squared can be used to determine the sample standard deviation of the errors as a fraction of the sample standard deviation of Y:

$$\text{Standard error of regression} = \text{SQRT}(1 \text{ minus } \textit{Adjusted R-squared}) \times \text{STDEV.S}(Y)$$

You can apply this equation without even calculating the model coefficients or the actual errors!

It follows from this equation that if you fit simple regression models to the same sample of the same dependent variable Y with different choices of X as the independent variable, then **adjusted R-squared necessarily goes up as the standard error of the regression goes down, and vice versa**. Hence, it is equivalent to say that your goal is to minimize the standard error of the regression or to maximize adjusted R-squared through your choice of X, other things being equal. However, as I will keep saying, the standard error of the regression is the real “bottom line” in your analysis: it measures the variations in the data that are not explained by the model *in real economic or physical terms*.

Adjusted R-squared can actually be *negative* if X has no measurable predictive value with respect to Y. In particular, if the correlation between X and Y is exactly zero, then R-squared is exactly equal to zero, and adjusted R-squared is equal to $1 - (n-1)/(n-2)$, which is negative because the ratio $(n-1)/(n-2)$ is greater than 1. If this is the case, then the mean model is clearly a better choice than the regression model. Some regression software will not even display a negative value for adjusted R-squared and will just report it to be zero in that case.

And finally, the standard errors of means and forecasts...

The **standard error of the mean** for Y for a given value of X is the estimated standard deviation of the error in measuring the height of the regression line at that location, given by the formula

$$SE_{mean}(X) = \frac{s}{\sqrt{n}} \times \sqrt{1 + \frac{(X - \text{AVERAGE}(X))^2}{\text{VAR.P}(X)}}$$

This looks like a lot like the formula for the standard error of the mean in the mean model: it is proportional to the standard error of the regression and inversely proportional to the square root of the sample size, so it gets steadily smaller as the sample size gets larger, approaching zero in the limit even in the presence of a lot of noise. However, in the regression model the standard error of the mean also depends to some extent on the value of X, so the term s/\sqrt{n} is scaled up by a factor that is greater than 1 and is larger for values of X that are farther from its mean. Hence, *there is relatively greater uncertainty about the true height of the regression line for values of X that are farther from its historical mean value*.

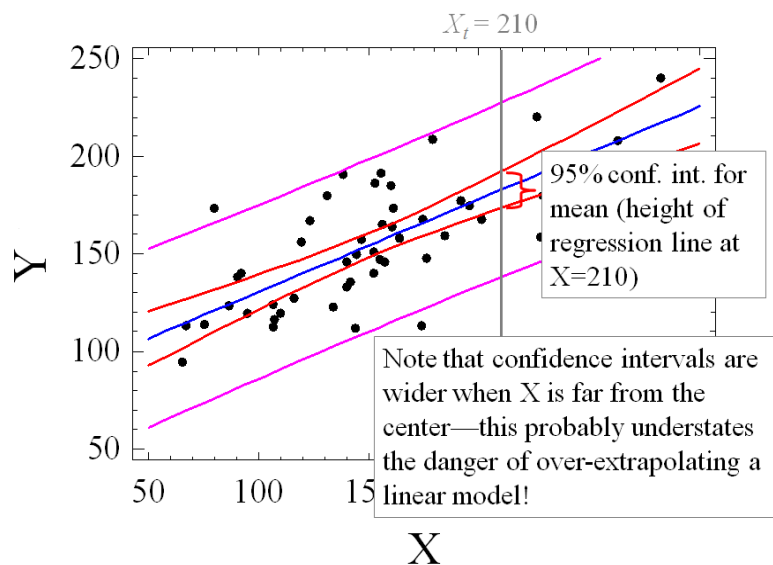
The **standard error for the forecast** for Y for a given value of X is then computed in exactly the same way as it was for the mean model:

$$SE_{fcst}(X) = \sqrt{s^2 + (SE_{mean}(X))^2}$$

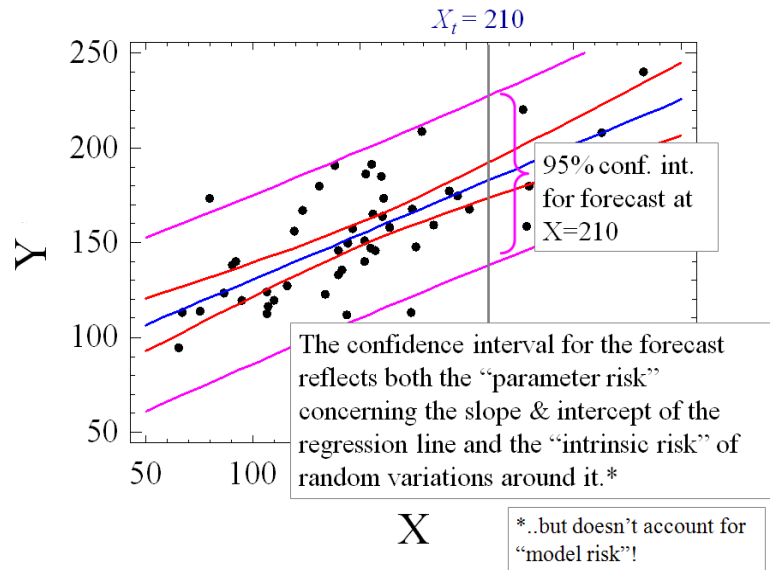
In the regression model it is larger for values of X that are farther from the mean—i.e., you expect to make bigger forecast errors when extrapolating the regression line farther out into space—because $SE_{mean}(X)$ is larger for more extreme values of X. The standard error of the forecast is not quite as sensitive to X in relative terms as is the standard error of the mean, because of the presence of the noise term s^2 under the square root sign. (Remember that s^2 is the estimated variance of the noise in the data.) In fact, s is usually much larger than $SE_{mean}(X)$ unless the data set is very small or X is very extreme, so usually the standard error of the forecast is not too much larger than the standard error of the regression.

Finally, **confidence limits for means and forecasts** are calculated in the usual way, namely as *the forecast plus or minus the relevant standard error times the critical t-value for the desired level of confidence and the number of degrees of freedom*, where the latter is $n-2$ for a simple regression model. For all but the smallest sample sizes, a 95% confidence interval is approximately equal to the point forecast plus-or-minus two standard errors, although there is nothing particularly magical about the 95% level of confidence—you can choose your own, or just report the standard error along with the point forecast.

Here are a couple of additional pictures that illustrate the behavior of the standard-error-of-the-mean and the standard-error-of-the-forecast in the special case of a simplex regression model. Because the standard error of the mean gets larger for extreme (farther-from-the-mean) values of X, the confidence intervals for the mean (the height of the regression line) widen noticeably at either end.



The confidence intervals for predictions also get wider when X goes to extremes, but the effect is not quite as dramatic because the standard error of the regression (which is usually a bigger component of forecast error) is a constant. Note that the inner set of confidence bands widens more in relative terms at the far left and far right than does the outer set of confidence bands.



But remember: *the standard errors and confidence bands that are calculated by the regression formulas are all based on the assumption that the model is correct, i.e., that the data really is described by the assumed linear equation with normally distributed errors.* If the model assumptions are not correct—e.g., if the wrong variables have been included or important variables have been omitted or if there are non-normalities in the errors or nonlinear relationships among the variables—then the predictions and their standard errors and confidence limits may all be suspect. So, when we fit regression models, we don't just look at the printout of the model coefficients—we look at various other statistics and charts that shed light on the validity of the model assumptions.

4. Take-aways:

1. The **coefficients** and error measures for a regression model are entirely determined by the following summary statistics: *means*, *standard deviations*, and *correlations* of the variables, and the *sample size*.

2. The correlation between Y and X, denoted by r_{XY} , is equal to **the average product of their standardized values**, i.e., the average of {the number of standard deviations by which Y deviates from its mean} times {the number of standard deviations by which X deviates from its mean}, using the population (rather than sample) standard deviation in the calculation. This statistic measures the strength of the linear relation between Y and X on a relative scale of -1 to +1. The correlation between Y and X is positive if they tend to move in the same direction relative to their respective means and negative if they tend to move in opposite directions, and it is zero if their up-or-down movements with respect to their own means are statistically independent.

3. The *slope coefficient* in a simple regression of Y on X is *the correlation between Y and X multiplied by the ratio of their standard deviations*: $b_1 = r_{XY} \times (\text{STDEV.S}(Y)/\text{STDEV.S}(X))$.

4. In a simple regression model, the percentage of variance “explained” by the model, which is called R-squared, is the square of the correlation between Y and X. That is, **R-squared = r_{XY}^2** , and that’s why it’s called R-squared. This means that the *sample standard deviation of the errors is equal to {the square root of 1-minus-R-squared} times the sample standard deviation of Y*:

$$\text{STDEV.S(errors)} = (\text{SQRT}(1 \text{ minus R-squared})) \times \text{STDEV.S}(Y).$$

So, if you know the standard deviation of Y, and you know the correlation between Y and X, you can figure out what the standard deviation of the errors would be if you regressed Y on X. However....

5. The sample standard deviation of the errors is a *downward-biased* estimate of the size of the true unexplained deviations in Y because it does not adjust for the additional “degree of freedom” used up by estimating the slope coefficient. An *unbiased* estimate of the standard deviation of the true errors is given by the *standard error of the regression*, denoted by *s*. In the special case of a simple regression model, it is:

$$\text{Standard error of regression} = \text{STDEV.S(errors)} \times \text{SQRT}((n-1)/(n-2))$$

This is the real bottom line, because the standard deviations of the errors of all the forecasts and coefficient estimates are directly proportional to it (if the model’s assumptions are correct!!)

6. *Adjusted R-squared*, which is obtained by adjusting R-squared for the degrees of freedom for error in exactly the same way, is an unbiased estimate of the amount of variance explained:

$$\text{Adjusted R-squared} = 1 - ((n-1)/(n-2)) \times (1 - \text{R-squared}).$$

For large values of *n*, there isn’t much difference.

In a *multiple* regression model in which k is the number of independent variables, the $n-2$ term that appears in the formulas for the standard error of the regression and adjusted R-squared merely becomes $n-(k+1)$.

7. The important thing about adjusted R-squared is that:

Standard error of the regression

$$= (\text{SQRT}(1 \text{ minus } \textit{adjusted-R-squared})) \times \text{STDEV.S}(Y).$$

So, for models fitted to the same sample of the same dependent variable, *adjusted R-squared always goes up when the standard error of the regression goes down.*

A model does not *always* improve when more variables are added: adjusted R-squared can go down (even go negative) if *irrelevant* variables are added.

8. The **standard error of a coefficient estimate** is the estimated standard deviation of the error in measuring it. And the estimated height of the regression line for a given value of X has its own standard error, which is called the **standard error of the mean at X**. *All of these standard errors are proportional to the standard error of the regression divided by the square root of the sample size.* So, a greater amount of noise in the data (as measured by s) makes all the estimates proportionally *less* accurate, and a larger sample size makes all of *them more* accurate (4 times as much data reduces all standard errors by a factor of 2). However, more data does *not* systematically reduce the standard error of the *regression*. Rather, the standard error of the regression merely becomes a more accurate estimate of the true standard deviation of the noise.

9. The **standard error of the forecast** for Y for a given value of X is **the square root of the sum of squares of the standard error of the regression and the standard error of the mean at X**. The standard error of the mean is usually a lot smaller than the standard error of the regression except when the sample size is very small and/or you are trying to predict what will happen under very extreme conditions, so *the standard error of the forecast is usually only slightly larger than the standard error of the regression.* (Recall that under the mean model, the standard error of the mean is a constant. In a simple regression model, the standard error of the mean depends on the value of X and is larger for values of X that are farther from its own mean.)

10. **Two-sided confidence limits for coefficient estimates, means, and forecasts are all equal to their point estimates plus-or-minus the appropriate critical t -value times their respective standard errors.** For simple regression model, in which two degrees of freedom are *used* up in estimating both the intercept and the slope coefficient, the appropriate critical t -value is **T.INV.2T(1 - C, $n - 2$)** where C is the desired level of confidence and n is the sample size. The usual default value for the confidence level is 95%, for which the critical t -value is T.INV.2T(0.05, $n - 2$)

For the case in which there are two or more independent variables, the mathematics is not too much harder if you are familiar with how to do arithmetic with vectors and matrices. Here is another spreadsheet file that gives a demonstration: people.duke.edu/~rnau/simpreg.xls.