POINTS OF SIGNIFICANCE

Simple linear regression

"The statistician knows…that in nature there never was a normal distribution, there never was a straight line, yet with normal and linear assumptions, known to be false, he can often derive results which match, to a useful approximation, those found in the real world."[1]

We have previously defined association between X and Y as meaning that the distribution of Y varies with X. We discussed correlation as a type of association in which larger values of Y are associated with larger values of X (increasing trend) or smaller values of X (decreasing trend)[2]. If we suspect a trend, we may want to attempt to predict the values of one variable using the values of the other. One of the simplest prediction methods is linear regression, in which we attempt to find a 'best line' through the data points.

Correlation and linear regression are closely linked—they both quantify trends. Typically, in correlation we sample both variables randomly from a population (for example, height and weight), and in regression we fix the value of the independent variable (for example, dose) and observe the response. The predictor variable may also be randomly selected, but we treat it as fixed when making predictions (for example, predicted weight for someone of a given height). We say there is a regression relationship between X and Y when the mean of Y varies with X.

In simple regression, there is one independent variable, X, and one dependent variable, Y. For a given value of X, we can estimate the average value of Y and write this as a conditional expectation $E(Y|X)$, often written simply as $\mu(X)$. If $\mu(X)$ varies with X, then we say that Y has a regression on X (**Fig. 1**). Regression is a specific kind of association and may be linear or nonlinear (**Fig. 1c,d**).

The most basic regression relationship is a simple linear regression. In this case, $E(Y|X) = \mu(X) = \beta_0 + \beta_1 X$, a line with intercept β_0 and slope β_1. We can interpret this as Y having a distribution with mean $\mu(X)$ for any given value of X. Here we are not interested in the shape of this distribution; we care only about its mean. The deviation of Y from $\mu(X)$ is often called the error, $\varepsilon = Y - \mu(X)$. It's important to realize that this term arises not because of any kind of error but because Y has a distribution for a given value of X. In other words, in the expression $Y = \mu(X) + \varepsilon$, $\mu(X)$ specifies the location of the distribution, and ε captures its shape. To predict Y at unobserved values of X, one substitutes the desired values of X in the estimated regression equation. Here X is referred to as the predictor, and Y is referred to as the predicted variable.

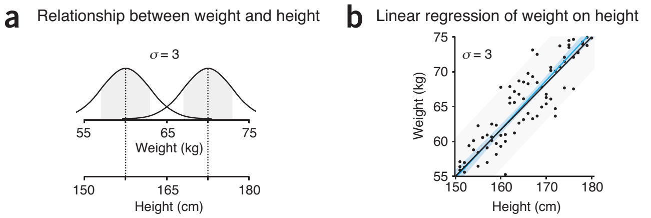

Figure 2 | In a linear regression relationship, the response variable has a distribution for each value of the independent variable. (**a**) At each height, weight is distributed normally with s.d. $\sigma = 3$. (**b**) Linear regression of $n = 3$ weight measurements for each height. The mean weight varies as $\mu(\text{Height}) = 2 \times \text{Height}/3 - 45$ (black line) and is estimated by a regression line (blue line) with 95% confidence interval (blue band). The 95% prediction interval (gray band) is the region in which 95% of the population is predicted to lie for each fixed height.

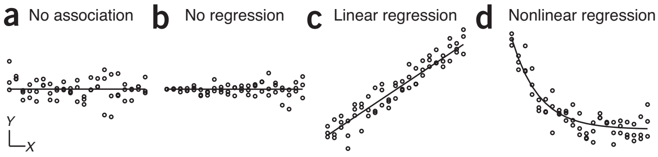

Figure 1 | A variable Y has a regression on variable X if the mean of Y (black line) $E(Y|X)$ varies with X. (**a**) If the properties of Y do not change with X, there is no association. (**b**) Association is possible without regression. Here $E(Y|X)$ is constant, but the variance of Y increases with X. (**c**) Linear regression $E(Y|X) = \beta_0 + \beta_1 X$. (**d**) Nonlinear regression $E(Y|X) = \exp(\beta_0 + \beta_1 X)$.

Consider a relationship between weight Y (in kilograms) and height X (in centimeters), where the mean weight at a given height is $\mu(X) = 2X/3 - 45$ for $X > 100$. Because of biological variability, the weight will vary—for example, it might be normally distributed with a fixed $\sigma = 3$ (**Fig. 2a**). The difference between an observed weight and mean weight at a given height is referred to as the error for that weight.

To discover the linear relationship, we could measure the weight of three individuals at each height and apply linear regression to model the mean weight as a function of height using a straight line, $\mu(X) = \beta_0 + \beta_1 X$ (**Fig. 2b**). The most popular way to estimate the intercept β_0 and slope β_1 is the least-squares estimator (LSE). Let (x_i, y_i) be the ith pair of X and Y values. The LSE estimates β_0 and β_1 by minimizing the residual sum of squares (sum of squared errors), $\text{SSE} = \Sigma(y_i - \hat{y}_i)^2$, where $\hat{y}_i = m(x_i) = b_0 + b_1 x_i$ are the points on the estimated regression line and are called the fitted, predicted or 'hat' values. The estimates are given by $b_0 = \bar{Y} - b_1 \bar{X}$ and $b_1 = r s_X/s_Y$, and where $\bar{X}$ and $\bar{Y}$ are means of samples X and Y, s_X and s_Y are their s.d. values and $r = r(X,Y)$ is their correlation coefficient[2].

The LSE of the regression line has favorable properties for very general error distributions, which makes it a popular estimation method. When Y values are selected at random from the conditional distribution $E(Y|X)$, the LSEs of the intercept, slope and fitted values are unbiased estimates of the population value regardless of the distribution of the errors, as long as they have zero mean. By "unbiased," we mean that although they might deviate from the population values in any sample, they are not systematically too high or too low. However, because the LSE is very sensitive to extreme values of both X (high leverage points) and Y (outliers), diagnostic outlier analyses are needed before the estimates are used.

In the context of regression, the term "linear" can also refer to a linear model, where the predicted values are linear in the parameters. This occurs when $E(Y|X)$ is a linear function of a known function $g(X)$, such as $\beta_0 + \beta_1 g(X)$. For example, $\beta_0 + \beta_1 X^2$ and $\beta_0 + \beta_1 \sin(X)$ are both linear regressions, but $\exp(\beta_0 + \beta_1 X)$ is nonlinear because it is not a linear function of the parameters β_0 and β_1. Analysis of variance (ANOVA) is a special case of a linear model in which the t treatments are labeled by indicator variables $X_1 \ldots X_t$, $E(Y|X_1 \ldots X_t) = \mu_i$ is the ith treatment mean, and the LSE predicted values are the corresponding sample means[3].

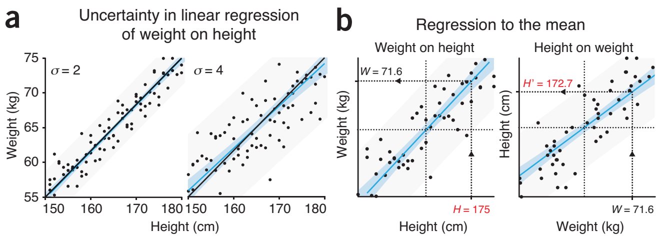

Figure 3 | Regression models associate error to response which tends to pull predictions closer to the mean of the data (regression to the mean). (**a**) Uncertainty in a linear regression relationship can be expressed by a 95% confidence interval (blue band) and 95% prediction interval (gray band). Shown are regressions for the relationship in **Figure 2a** using different amounts of scatter (normally distributed with s.d. σ). (**b**) Predictions using successive regressions $X \rightarrow Y \rightarrow X'$ to the mean. When predicting using height H = 175 cm (larger than average), we predict weight W = 71.6 kg (dashed line). If we then regress H on W at W = 71.6 kg, we predict H' = 172.7 cm, which is closer than H to the mean height (64.6 cm). Means of height and weight are shown as dotted lines.

Recall that in ANOVA, the SSE is the sum of squared deviations of the data from their respective sample means (i.e., their predicted values) and represents the variation in the data that is not accounted for by the treatments. Similarly, in regression, the SSE is the sum of squared deviations of the data from the predicted values that represents variation in data not explained by regression. In ANOVA we also compute the total and treatment sum of squares; the analogous quantities in linear regression are the total sum of squares, SST = $(n-1)s^2_Y$, and the regression sum of squares, SSR = $\Sigma(\hat{y}_i - \bar{Y})^2$, which are related by SST = SSR + SSE. Furthermore, SSR/SST = r^2 is the proportion of variance of Y explained by the linear regression of X (ref. 2).

When the errors have constant variance σ^2, we can model the uncertainty in regression parameters. In this case, b_0 and b_1 have means β_0 and β_1, respectively, and variances $\sigma^2(1/n + \bar{X}^2/s_{XX})$ and σ^2/s_{XX}, where $s_{XX} = (n-1)s^2_X$. As we collect X over a wider range, s_{XX} increases, so the variance of b_1 decreases. The predicted value $\hat{y}(x)$ has a mean $\beta_0 + \beta_1 x$ and variance $\sigma^2(1/n + (x - \bar{X})^2/s_{XX})$. Additionally, the mean square error (MSE) = SSE/(n – 2) is an unbiased estimator of the error variance (i.e., σ^2). This is identical to how MSE is used in ANOVA to estimate the within-group variance, and it can be used as an estimator of σ^2 in the equations above to allow us to find the standard error (SE) of b_0, b_1 and $\hat{y}_x$. For example, SE$(\hat{y}(x)) = \sqrt{}(\text{MSE}(1/n + (x - \bar{X})^2/s_{XX}))$.

If the errors are normally distributed, so are b_0, b_1 and $(\hat{y}(x))$. Even if the errors are not normally distributed, as long as they have zero mean and constant variance, we can apply a version of the central limit theorem for large samples[4] to obtain approximate normality for the estimates. In these cases the SE is very helpful in testing hypotheses. For example, to test that the slope is β_1 = 2/3, we would use $t^* = (b_1 - \beta_1)/\text{SE}(b_1)$; when the errors are normal and the null hypothesis true, t^* has a t-distribution with d.f. = n – 2. We can also calculate the uncertainty of the regression parameters using confidence intervals, the range of values that are likely to contain β_i (for example, 95% of the time)[5]. The interval is $b_i \pm t_{0.975}\text{SE}(b_i)$, where $t_{0.975}$ is the 97.5% percentile of the t-distribution with d.f. = n – 2.

When the errors are normally distributed, we can also use confidence intervals to make statements about the predicted value for a fixed value of X. For example, the 95% confidence interval for $\mu(x)$ is $b_0 + b_1x \pm t_{0.975}\text{SE}(\hat{y}(x))$ (**Fig. 2b**) and depends on the error variance (**Fig. 3a**). This is called a point-wise interval because the 95% coverage is for a single fixed value of X. One can compute a band that covers the entire line 95% of the time by replacing $t_{0.975}$ with $W_{0.975} = \sqrt{}(2F_{0.975})$, where $F_{0.975}$ is the critical value from the $F_{2,n-2}$ distribution. This interval is wider because it must cover the entire regression line, not just one point on the line.

To express uncertainty about where a percentage (for example, 95%) of newly observed data points would fall, we use the prediction interval $b_0 + b_1\text{x} + t_{0.975}$ (MSE$(1 + 1/n + (x - \bar{X})^2/s_{XX})$). This interval is wider than the confidence interval because it must incorporate both the spread in the data and the uncertainty in the model parameters. A prediction interval for Y at a fixed value of X incorporates three sources of uncertainty: the population variance σ^2, the variance in estimating the mean and the variability due to estimating σ^2 with the MSE. Unlike confidence intervals, which are accurate when the sampling distribution of the estimator is close to normal, which usually occurs in sufficiently large samples, the prediction interval is accurate only when the errors are close to normal, which is not affected by sample size.

Linear regression is readily extended to multiple predictor variables $X_1, \ldots, X_p$, giving E$(Y|X_1, \ldots, X_p) = \beta_0 + \Sigma\beta_i X_i$. Clever choice of predictors allows for a wide variety of models. For example, $X_i = X^i$ yields a polynomial of degree p. If there are p + 1 groups, letting X_i = 1 when the sample comes from group i and 0 otherwise yields a model in which the fitted values are the group means. In this model, the intercept is the mean of the last group, and the slopes are the differences in means.

A common misinterpretation of linear regression is the 'regression fallacy'. For example, we might predict weight W = 71.6 kg for a larger than average height H = 175 cm and then predict height H' = 172.7 cm for someone with weight W = 71.6 kg (**Fig. 3b**). Here we will find $H' < H$. Similarly, if H is smaller than average, we will find $H' > H$. The regression fallacy is to ascribe a causal mechanism to regression to the mean, rather than realizing that it is due to the estimation method. Thus, if we start with some value of X, use it to predict Y, and then use Y to predict X, the predicted value will be closer to the mean of X than the original value (**Fig. 3b**).

Estimating the regression equation by LSE is quite robust to non-normality of and correlation in the errors, but it is sensitive to extreme values of both predictor and predicted. Linear regression is much more flexible than its name might suggest, including polynomials, ANOVA and other commonly used statistical methods.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Naomi Altman & Martin Krzywinski

1. Box, G. *J. Am. Stat. Assoc.* **71**, 791–799 (1976).
2. Altman, N. & Krzywinski, M. *Nat. Methods* **12**, 899–900 (2015).
3. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 699–700 (2014).
4. Krzywinski, M. & Altman, N. *Nat. Methods* **10**, 809–810 (2013).
5. Krzywinski, M. & Altman, N. *Nat. Methods* **10**, 1041–1042 (2013).

Naomi Altman is a Professor of Statistics at The Pennsylvania State University. Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre.