Simple linear regression is the most commonly used technique for determining how one variable of interest (the response variable) is affected by changes in another variable (the explanatory variable). The terms "response" and "explanatory" mean the same thing as "dependent" and "independent", but the former terminology is preferred because the "independent" variable may actually be interdependent with many other variables as well.

Simple linear regression is used for three main purposes:

1. To describe the linear dependence of one variable on another
2. To predict values of one variable from values of another, for which more data are available
3. To correct for the linear dependence of one variable on another, in order to clarify other features of its variability.

Any line fitted through a cloud of data will deviate from each data point to greater or lesser degree. The vertical distance between a data point and the fitted line is termed a "residual". This distance is a measure of prediction error, in the sense that it is the discrepancy between the actual value of the response variable and the value predicted by the line. Linear regression determines the best-fit line through a scatterplot of data, such that the sum of squared residuals is minimized; equivalently, it minimizes the error variance. The fit is "best" in precisely that sense: the sum of squared errors is as small as possible. That is why it is also termed "Ordinary Least Squares" regression.

Derivation of linear regression equations

The mathematical problem is straightforward:

given a set of n points (X_i,Y_i) on a scatterplot,
find the best-fit line, $\hat{Y}_i = a + bX_i$
such that the sum of squared errors in Y, $\sum\left(Y_i - \hat{Y}_i\right)^2$ is minimized

The derivation proceeds as follows: for convenience, name the sum of squares "Q",

$$Q = \sum_{i=1}^{n}\left(Y_i - \hat{Y}\right)^2 = \sum_{i=1}^{n}\left(Y_i - a - bX_i\right)^2 \tag{1}$$

Then, Q will be minimized at the values of a and b for which $\partial Q / \partial a = 0$ and $\partial Q / \partial b = 0$. The first of these conditions is,

$$\frac{\partial Q}{\partial a} = \sum_{i=1}^{n} -2\left(Y_i - a - bX_i\right) = 2\left(na + b\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i\right) = 0 \tag{2}$$

which, if we divide through by 2 and solve for *a*, becomes simply,

$$a = \overline{Y} - b\overline{X} \tag{3}$$

which says that the constant *a* (the y-intercept) is set such that the line must go through the mean of x and y. This makes sense, because this point is the "center" of the data cloud. The second condition for minimizing Q is,

$$\frac{\partial Q}{\partial b} = \sum_{i=1}^{n} -2X_i\left(Y_i - a - bX_i\right) = \sum_{i=1}^{n} -2\left(X_iY_i - aX_i - bX_i^2\right) = 0 \tag{4}$$

If we substitute the expression for *a* from (3) into (4), then we get,

$$\sum_{i=1}^{n}\left(X_iY_i - X_i\overline{Y} + bX_i\overline{X} - bX_i^2\right) = 0 \tag{5}$$

We can separate this into two sums,

$$\sum_{i=1}^{n}\left(X_iY_i - X_i\overline{Y}\right) - b\sum_{i=1}^{n}\left(X_i^2 - X_i\overline{X}\right) = 0 \tag{6}$$

which becomes directly,

$$b = \frac{\sum_{i=1}^{n}\left(X_iY_i - X_i\overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i^2 - X_i\overline{X}\right)} = \frac{\sum_{i=1}^{n}\left(X_iY_i\right) - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n}\left(X_i^2\right) - n\overline{X}^2} \tag{7}$$

We can translate (7) into a more intuitively obvious form, by noting that

$$\sum_{i=1}^{n}\left(\overline{X}^2 - X_i\overline{X}\right) = 0 \qquad \text{and} \qquad \sum_{i=1}^{n}\left(\overline{X}\,\overline{Y} - Y_i\overline{X}\right) = 0 \tag{8}$$

so that *b* can be rewritten as the ratio of Cov(x,y) to Var(x):

$$b = \frac{\sum_{i=1}^{n}\left(X_iY_i - X_i\overline{Y}\right) + \sum_{i=1}^{n}\left(\overline{X}\,\overline{Y} - Y_i\overline{X}\right)}{\sum_{i=1}^{n}\left(X_i^2 - X_i\overline{X}\right) + \sum_{i=1}^{n}\left(\overline{X}^2 - X_i\overline{X}\right)} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \frac{Cov(X,Y)}{Var(X)} \tag{9}$$

The quantities that result from regression analyses can be written in many different forms that are mathematically equivalent but superficially distinct. All of the following forms of the regression slope *b* are mathematically equivalent:

$$b = \frac{Cov(X,Y)}{Var(X)} \text{ or } \frac{\sum xy}{\sum x^2} \text{ or } \frac{\sum_{i=1}^{n}\left(X_iY_i\right) - \frac{\sum_{i=1}^{n}X_i\sum_{i=1}^{n}Y_i}{n}}{\sum_{i=1}^{n}\left(X_i\right)^2 - \frac{\left(\sum_{i=1}^{n}X_i\right)^2}{n}} \text{ or } \frac{\sum_{i=1}^{n}\left(X_iY_i\right) - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n}\left(X_i^2\right) - n\overline{X}^2} \text{ or } \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_iY_i\right) - \overline{X}\,\overline{Y}}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i^2\right) - \overline{X}^2} \text{ or } \frac{\overline{(XY)} - \overline{X}\,\overline{Y}}{\overline{\left(X_i^2\right)} - \overline{X}^2} \tag{10}$$

A common notational shorthand is to write the "sum of squares of X" (that is, the sum of squared deviations of the X's from their mean), the "sum of squares of Y", and the "sum of XY cross products" as,

$$\sum x^2 = SS_x = (n-1)Var(X) = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n}\left(X_i^2\right) - n\overline{X}^2 \tag{11}$$

$$\sum y^2 = SS_y = (n-1)Var(Y) = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(Y_i^2\right) - n\overline{Y}^2 \tag{12}$$

$$\sum xy = S_{xy} = (n-1)Cov(X,Y) = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) = \sum_{i=1}^{n}\left(X_iY_i\right) - n\overline{X}\,\overline{Y} \tag{13}$$

It is important to recognize that Σx^2, Σy^2, and Σxy, as used in equations (10)-(13) and in several textbooks, including Sokal and Rohlf, are not summations; instead, they are *symbols* for the sums of squares and cross products. Note also that S and SS in (11)-(13) are uppercase S's rather than standard deviations.

Besides the regression slope *b* and intercept *a*, the third parameter of fundamental importance is the correlation coefficient *r* or the coefficient of determination r^2. r^2 is the ratio between the variance in Y that is "explained" by the regression (or, equivalently, the variance in $\hat{Y}$), and the total variance in Y. Like *b*, r^2 can be calculated many different ways:

$$r^2 = \frac{Var(\hat{Y})}{Var(Y)} = \frac{b^2Var(X)}{Var(Y)} = \frac{\left(Cov(x,y)\right)^2}{Var(X)Var(Y)} = \frac{Var(Y) - Var(Y - \hat{Y})}{Var(Y)} = \frac{S_{xy}^2}{SS_xSS_y} \tag{14}$$

Equation (14) implies the following relationship between the correlation coefficient, *r*, the regression slope, *b*, and the standard deviations of X and Y (s_X and s_Y):

$$r = b\frac{s_X}{s_Y} \qquad \text{and} \qquad b = r\frac{s_Y}{s_X} \tag{15}$$

The residuals e_i are the deviations of each response value Y_i from its estimate $\hat{Y}_i$. These residuals can be summed in the sum of squared errors (SSE). The mean square error (MSE) is just what the name implies, and can also be considered the "error variance" ($s_{Y\bullet X}^2$). The root-mean-square-error (RMSE), also termed the "standard error of the regression" ($s_{Y\bullet X}$) is the standard deviation of the residuals. The mean square error and RMSE are calculated by dividing by n-2, because linear regression removes two degrees of freedom from the data (by estimating two parameters, *a* and *b*).

$$\begin{gathered} e_i = Y_i - \hat{Y}_i \qquad SSE = \sum_{i=1}^{n} e_i^2 \qquad MSE = s_{Y\bullet X}^2 = \frac{SSE}{n-2} = Var(Y)\left(1-r^2\right)\frac{n-1}{n-2} \\ RMSE = s_{Y\bullet X} = \sqrt{\frac{SSE}{n-2}} = s_Y\sqrt{\frac{n-1}{n-2}}\sqrt{1-r^2} \end{gathered} \tag{16}$$

where Var(Y) is the sample, not population, variance of Y, and the factors of n-1/n-2 serve only to correct for changes in the number of degrees of freedom between the calculation of variance (d.f.=n-1) and $s_{Y\bullet X}$ (d.f.=n-2).

Uncertainty in regression parameters

The standard error of the regression slope *b* can be expressed many different ways, including:

$$s_b = \sqrt{\frac{SS_Y/SS_X - b^2}{n-2}} = \frac{s_{Y\bullet X}}{\sqrt{SS_X}} = \frac{1}{\sqrt{n-1}}\frac{s_{Y\bullet X}}{s_X} = \frac{s_Y}{s_X}\frac{\sqrt{1-r^2}}{\sqrt{n-2}} = \frac{b}{r}\frac{\sqrt{1-r^2}}{\sqrt{n-2}} = \frac{b}{\sqrt{n-2}}\sqrt{\frac{1}{r^2}-1} \tag{17}$$

If all of the assumptions underlying linear regression are true (see below), the regression slope *b* will be approximately *t*-distributed. Therefore, confidence intervals for *b* can be calculated as,

$$CI = b \pm t_{\alpha(2),n-2}s_b \tag{18}$$

To determine whether the slope of the regression line is statistically significant, one can straightforwardly calculate t, the number of standard errors that *b* differs from a slope of zero:

$$t = \frac{b}{s_b} = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}} \tag{19}$$

and then use the t-table to evaluate the α for this value of t (and n-2 degrees of freedom). The uncertainty in the elevation of the regression line at the mean X (that is, the uncertainty in $\hat{Y}$ at the mean X) is simply the standard error of the regression $s_{Y\bullet X}$, divided by the square root of n. Thus the standard error in the predicted value of $\hat{Y}_i$ for some X_i is the uncertainty in the elevation at the mean X, plus the uncertainty in *b* times the distance from the mean X to X_i, added in quadrature:

$$s_{\hat{Y}_i} = \sqrt{\left(s_{Y\bullet X}/\sqrt{n}\right)^2 + \left(s_b\left(X_i - \overline{X}\right)\right)^2} = s_{Y\bullet X}\sqrt{\frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{SS_X}} = \frac{s_{Y\bullet X}}{\sqrt{n}}\sqrt{1 + \frac{\left(X_i - \overline{X}\right)^2}{Var(X)}} \tag{20}$$

where Var(X) is the population, (not sample) variance of X (that is, it is calculated with *n* rather than *n*-1). $\hat{Y}_i$ is also *t*-distributed, so a confidence interval for $\hat{Y}_i$ can be estimated by multiplying the standard error of $\hat{Y}_i$ by $t_{\alpha(2),n-2}$. Note that this confidence interval grows as X_i moves farther and farther from the mean of X. Extrapolation beyond the range of the data assumes that the underlying relationship continues to be linear beyond that range. Equation (20) gives the standard error of the $\hat{Y}_i$, that is, the Y-values predicted by the regression line. The uncertainty in a new individual value of Y (that is, the prediction interval rather than the confidence interval) depends not only on the uncertainty in where the regression line is, but also the uncertainty in where the individual data point Y lies in relation to the regression line. This latter uncertainty is simply the standard deviation of the residuals, or $s_{Y\bullet X}$, which is added (in quadrature) to the uncertainty in $\hat{Y}_i$, as follows:

$$\left(s_{\hat{Y}_i}\right)_1 = \sqrt{{s_{Y\bullet X}}^2 + s_{\hat{Y}_i}^2} = s_{Y\bullet X}\sqrt{1+\frac{1}{n}+\frac{\left(X_i-\overline{X}\right)^2}{SS_X}} \qquad (21)$$

The standard error of the Y-intercept, *a*, is just a special case of (20) for X_i=0,

$$s_a = \sqrt{\left(\frac{s_{Y\bullet X}}{\sqrt{n}}\right)^2 + \left(s_b\overline{X}\right)^2} = s_{Y\bullet X}\sqrt{\frac{1}{n}+\frac{\overline{X}^2}{SS_X}} \qquad (22)$$

The standard error of the correlation coefficient *r* is,

$$s_r = \sqrt{\frac{1-r^2}{n-2}} \qquad (23)$$

We can test whether the correlation between X and Y is statistically significant by comparing r to its standard error,

$$t = \frac{r}{s_r} = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}} \qquad (24)$$

and looking up this value in a t-table. Note that $t=r/s_r$ has the same value as $t=b/s_b$; that is, the statistical significance of the correlation coefficient *r* is equivalent to the statistical significance of the regression slope *b*.

Assumptions behind linear regression

The assumptions that must be met for linear regression to be valid depend on the purposes for which it will be used. Any application of linear regression makes two assumptions:

(A) The data used in fitting the model are representative of the population.

(B) The true underlying relationship between X and Y is linear.

All you need to assume to predict Y from X are (A) and (B). To estimate the standard error of the prediction $s_{\hat{Y}_i}$, you also must assume that:

(C) The variance of the residuals is constant (homoscedastic, not heteroscedastic).

For linear regression to provide the best linear unbiased estimator of the true Y, (A) through (C) must be true, and you must also assume that:

(D) The residuals must be independent.

To make probabilistic statements, such as hypothesis tests involving *b* or *r*, or to construct confidence intervals, (A) through (D) must be true, and you must also assume that:

(E) The residuals are normally distributed.

Contrary to common mythology, linear regression does *not* assume *anything* about the distributions of either X or Y; it only makes assumptions about the distribution of the residuals e_i. As with many other statistical techniques, it is *not* necessary for the data themselves to be normally distributed, only for the errors (residuals) to be normally distributed. And this is only required for the statistical significance tests (and other probabilistic statements) to be valid; regression can be applied for many other purposes even if the errors are non-normally distributed.

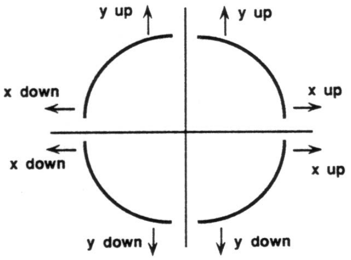
Steps in constructing good regression models

1. Plot and examine the data.

2. If necessary, transform the X and/or Y variables so that:
   -the relationship between X and Y is linear, and
   -Y is homoskedastic (that is, the scatter in Y is constant from one end of the X data to the other)

If (as is often the case), the scatter in Y increases with increasing Y, the heteroscedasticity can be eliminated by transforming Y downward on the "ladder of powers" (see the toolkit on transforming distributions). Conversely, if the scatter in Y is greater for smaller Y, transform Y upward on the ladder of powers.

Curvature in the data can be reduced by transforming Y and/or X up or down the ladder of powers according to the "bulging rule" of Mosteller and Tukey (1977), which is illustrated in the following diagram:



The bulging rule for transforming curvature to linearity.
(after Mosteller and Tukey, 1977).

Note that transforming X will change the curvature of the data without affecting the variance of Y, whereas transforming Y will affect both the shape of the data and the heteroscedasticity of the data. Note that visual assessments of the "scatter" in the data are vulnerable to an optical illusion: if the data density changes with X, the spread in the Y values will look larger wherever there are more data, even if the error variance is constant throughout the range of X.

3. Calculate the linear regression statistics. Every standard statistics package does this, as do many spreadsheets, pocket calculators, etc. It is not difficult to do by hand, or via a custom spreadsheet. The steps are as follows:
   (a) for each data point, calculate X_i^2, Y_i^2, and X_iY_i
   (b) calculate the sums of the X_i, Y_i, X_i^2, Y_i^2, and X_iY_i
   (c) calculate the sums of squares SS_x, SS_y, and S_{xy} (also written Σx^2, Σy^2, and Σxy) via(11)-(13)
   (d) calculate a, b, and r^2 via (10), (3), and (14)
   (e) calculate s_b and s_r via (17) and (23)

4. (a) Examine the regression slope and intercept. Are they physically plausible? Within a plausible range of X values, does the regression equation predict reasonable values of Y? (b) Does r^2 indicate that the regression explains enough variance to make it useful? "Useful" depends on your purpose: if you seek to predict Y accurately, then you want to be able to explain a substantial fraction of the variance in Y. If, on the other hand, you want to simply clarify how X affects Y, a high r^2 is not important (indeed, part of your task of clarification consists in determining how much of the variation in Y is explainable by variation in X). (c) Does the standard error of the slope indicate that *b* is precise enough for your purposes? If you want to predict Y from X, are the confidence intervals for Y adequate for your purposes?

Important note: r^2 is often largely irrelevant to the task at hand, and slavishly seeking to obtain the highest possible r^2 is often counterproductive. In polynomial regression or multiple regression, adding more adjustable coefficients to the regression equation will always increase r^2, even though doing so may not improve the predictive validity of the fitted equation. Indeed, it may undermine the usefulness of the analysis, if one begins fitting to the *noise* in the data rather than the *signal*.

5. Examine the residuals, $e_i = Y_i - \hat{Y}_i$. The following residual plots are particularly useful:

5(a). Plot the residuals versus X. (see examples on pp. 7-8)
   -If the residuals increase or decrease with X, they are heteroscedastic. Transform Y to cure this.
   -If the residuals are curved with X, the relationship between X and Y is nonlinear. Either transform X, or fit a nonlinear curve to the data.

-If there are outliers, check their validity, and/or use robust regression techniques.

5(b) Plot the residuals versus $\hat{Y}$, again to check for heteroscedasticity (this step is redundant with 5(a) for simple one-variable linear regression, and can be skipped).

5(c) Plot the residuals against every other possible explanatory variable in the data set.

-If the residuals are correlated with another variable (call it Z), then check to see whether Z is also correlated with X. If both the residuals and X are correlated with Z, then the regression slope will *not* accurately reflect the dependence of Y on X. You <u>must</u> either: (1) correct both X and Y for changes in Z, before regressing Y on X, or preferably (2) use multiple regression, or another fitting technique that can account for the interactions between X and Z and their combined effect on Y. If the residuals are correlated with Z, but X is not, then multiple regression on both X and Z will allow you to predict Y more accurately (that is, explain more of the variance in Y), but ignoring Z will not bias the regression slope of Y on X.

5(d) Plot the residuals against time.

-Check for seasonal variation, or long-term trend. Again, they should be accounted for, if they are present.

5(e) Plot the residuals against their lags (that is, plot e_i versus e_{i-1}).

-If the residuals are strongly correlated with their lags, the residuals are serially correlated. Serial correlation (also called autocorrelation) means the true uncertainties in the relationship between X and Y will be larger (potentially *much* larger) than suggested by the calculated standard errors. Dealing with serial correlation requires special techniques such as the Hildreth-Lu procedure, which will be explained in a later toolkit.

5(f) Plot the distribution of the residuals (either as a histogram, or a normal quantile plot)

-If the residuals are not normally distributed, your estimates of statistical significance and confidence intervals will not be accurate.

6 Check for outliers, both visually and statistically (see Helsel and Hirsch, section 9.5 for more information). One of the simplest measures of influence is Cook's "D", calculated for each data point as

$$D_i = \frac{e_i^2}{2s_{Y\bullet X}^2}\left(\frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{SS_X}\right)$$

Points with D_i values greater than $F_{0.1,3,n-2}$ are considered to have a large influence on the regression line; for n greater than about 30 this corresponds to $D_i \approx 2.4$. High influence does not automatically make a point an outlier, but it does mean that it makes a substantial difference whether the point is included or not. If the point can be shown to be a mistake, then it should be either corrected or deleted. If it is not a mistake, or if you are unsure, then a more robust regression technique is often advisable.

Common pitfalls in simple linear regression

<u>Mistakenly attributing causation.</u> Regression *assumes* that X causes Y; it cannot *prove* that X causes Y. X and Y may be strongly correlated either because X causes Y, or because Y causes X, or because some other variable(s) causes variation in both X and Y. Which brings us to:

<u>Overlooking hidden variables.</u> As mentioned above, hidden variables that are correlated with both X and Y can obscure, or even distort, the dependence of Y on X.

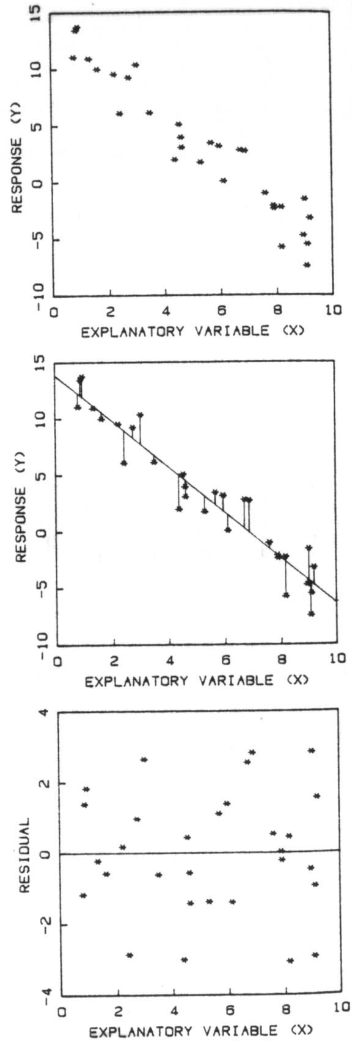
<u>Overlooking serial correlation.</u> Strong serial correlation can cause you to seriously underestimate the uncertainties in your regression results (since successive measurements are not independent, the true number of degrees of freedom is much smaller than n suggests). In time-series data, it can also produce spurious but impressive-looking trends.

<u>Overlooking artifactual correlation.</u> Whenever some part of the X-axis variable also appears on the Y-axis, there is an artifactual correlation between X and Y in addition to (or in opposition to) the real correlation between X and Y.

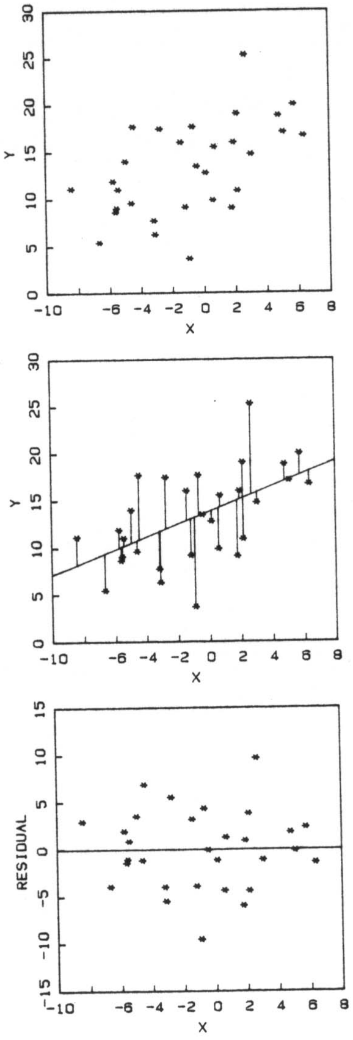
Note: all of pitfalls listed above can occur even though r^2 is large; indeed, all of them can sometimes serve to inflate r^2. This is yet another reason why r^2 should rarely be the "holy grail" of regression analysis.

<u>Overlooking uncertainty in X.</u> Linear regression assumes that X is known precisely, and only Y is uncertain. If there are significant uncertainties in X, the regression slope will be lower than it would have been otherwise. The regression line will still be an unbiased estimator of the value of Y that is likely to accompany a given X measurement, but it will be a biased estimator of the Y values that would arise if X could be controlled precisely.

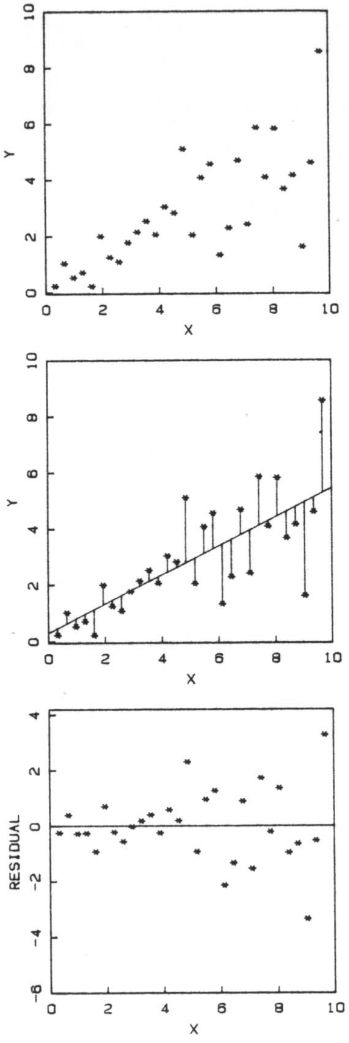
Scatterplot, fitted regression line, and residual plot for linear relationship with well-behaved residuals



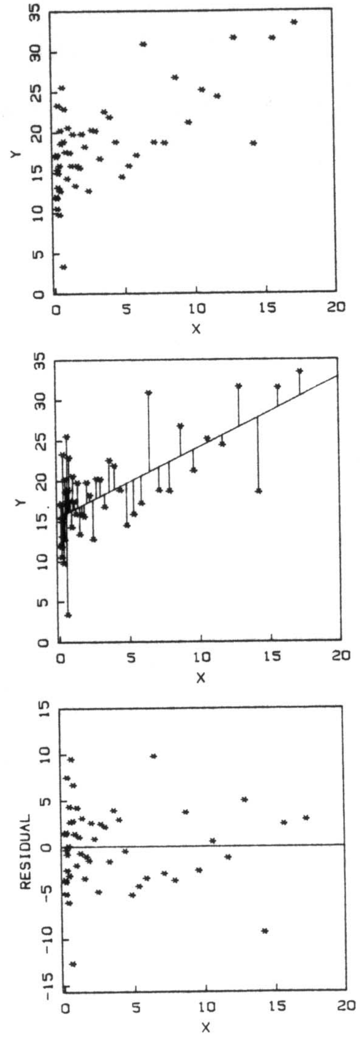
Another linear relationship with well-behaved residuals, but with more scatter

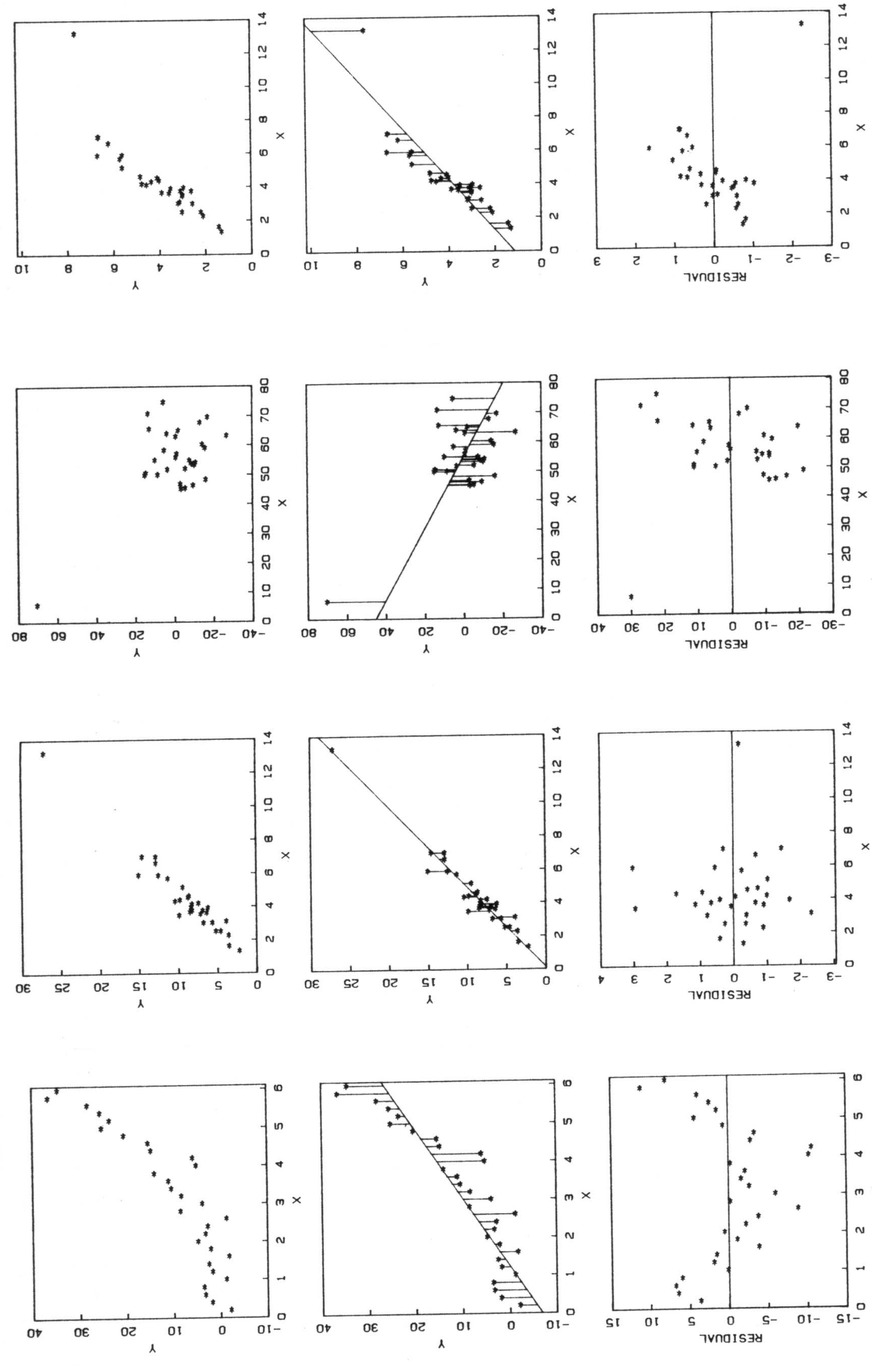


Linear relationship with heteroscedasticity: variance of residuals changes with X



Linear relationship without heteroscedasticity, but with optical illusion from change in density of X_i values





Nonlinear relationship between X and Y

Linear relationship with outlier that has high *leverage*, but little *influence* (it could profoundly alter the regression line, but doesn't, because it is consistent with the rest of the data)

Data that have virtually no linear relationship between X and Y, except for a highly influential outlier

Linear relationship with one highly influential outlier that alters the fitted regression line

References:

Chambers, J. M., W. S. Cleveland, B. Kleiner and P. A. Tukey, *Graphical Methods for Data Analysis*, 395 pp., Wadsworth & Brooks/Cole Publishing Co., 1983.

Helsel, D. R. and R. M. Hirsch, *Statistical Methods in Water Resources*, 522 pp., Elsevier, 1992.

Mosteller, F. and J. W. Tukey, *Data Analysis and Regression*, 588 pp., Addison-Wesley, 1977.