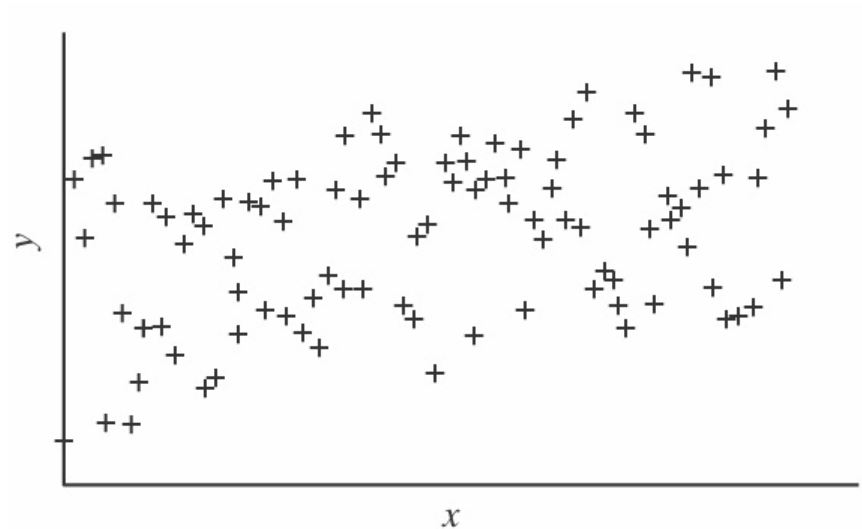
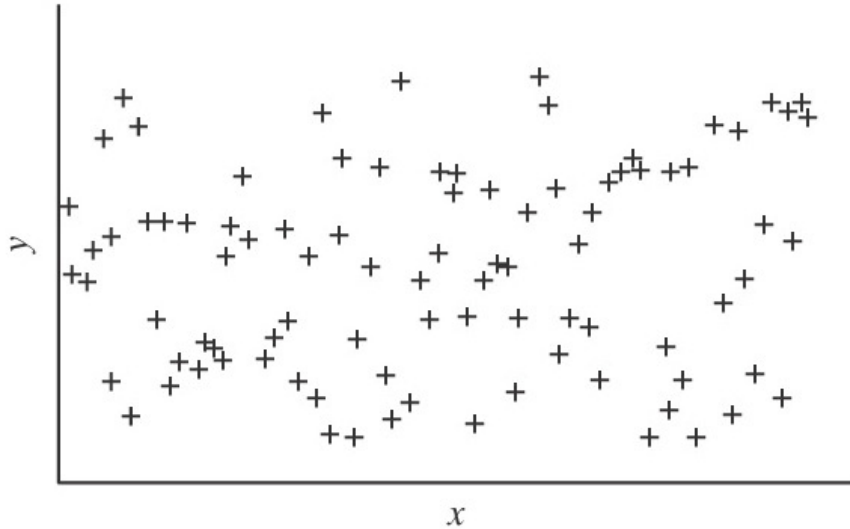

Correlation and Regression

Ananda Mysore
SJSU

Correlated or Not?



- ❑ All experimental data is subject to random error.
- ❑ Random error obscures the relationship between an independent variables x and a dependent variables y .
- ❑ How does one quantitatively express whether or not there is a correlation between variables?

Correlation Coefficient

- ❑ A **correlation coefficient** is a calculated value that determines the extent to which two variables follow a particular trend.
- ❑ The linear correlation coefficient r_{xy} is defined to express the extent to which a set of n pairs of variables x and y follows a line.

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

- ❑ Values for the r_{xy} range from -1 to +1. The farther it is from zero, the stronger the correlation.
- ❑ Question: How does r_{xy} relate to the slope of a fitted line through the (x, y) data?

How Correlated is Correlated?

- If the scatter of y at a particular x is random, then statistical methods can be applied to identify a threshold value of $|r_{xy}|$ for any selected significance level α .
- Beyond the threshold, correlation may be claimed with $100(1-\alpha)\%$ confidence.

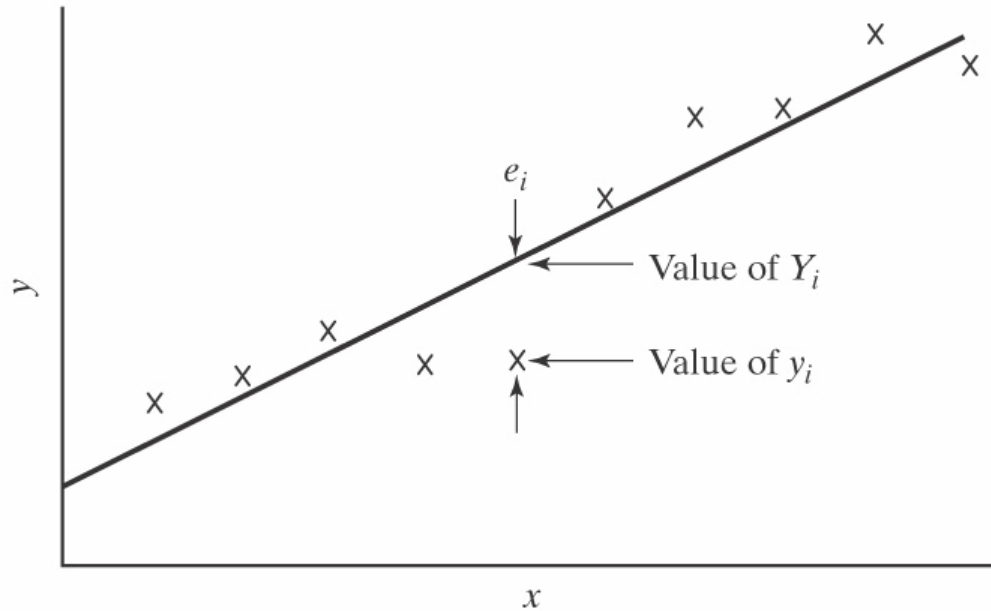
TABLE 6.9 Minimum Values of the Correlation Coefficient for Significance Level α

n	α				
	0.20	0.10	0.05	0.02	0.01
3	0.951	0.988	0.997	1.000	1.000
4	0.800	0.900	0.950	0.980	0.990
5	0.687	0.805	0.878	0.934	0.959
6	0.608	0.729	0.811	0.882	0.917
7	0.551	0.669	0.754	0.833	0.875
8	0.507	0.621	0.707	0.789	0.834
9	0.472	0.582	0.666	0.750	0.798
10	0.443	0.549	0.632	0.715	0.765

Limitations of Correlation Coefficient

- ❑ The linear correlation coefficient does not demand that the relationship between x and y be linear; it merely quantifies the extent to which it is.
- ❑ Outliers and systematic errors may greatly influence the value of r_{xy} and need to be eliminated for greatest effectiveness.
- ❑ Correlation is purely an analytical geometry observation, and does not imply causality.

Linear Regression



$$Y = ax + b$$

$$e_i = Y_i - y_i$$

- ❑ The general concept of **regression** is to use data to generate an analytical expression.
- ❑ Linear regression fits a straight line $Y = ax + b$ through data, using the method of “least squares” to choose slope a and y -intercept b , such that they minimize the sum of the squared errors e .

Coefficient of Determination

- The coefficient of determination r^2 expresses the extent to which a regression equation matches the actual (x, y) data:

$$r^2 = 1 - \frac{\sum_{i=1}^n (Y_i - y_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} = 1 - \frac{\sum_{i=1}^n [(ax_i + b) - y_i]^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2}$$

- The closer r^2 is to 1, the better the fit.
- If the data is reasonably certain to go through the origin (e.g. zero offset error has already been addressed), the fitted line may enforce that y-intercept b goes through zero.

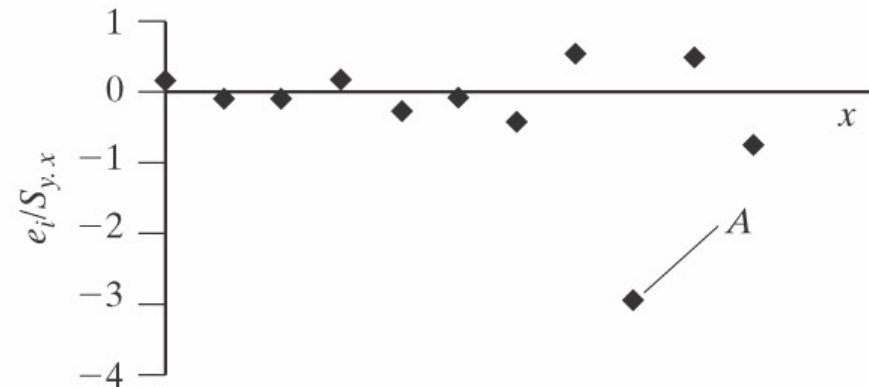
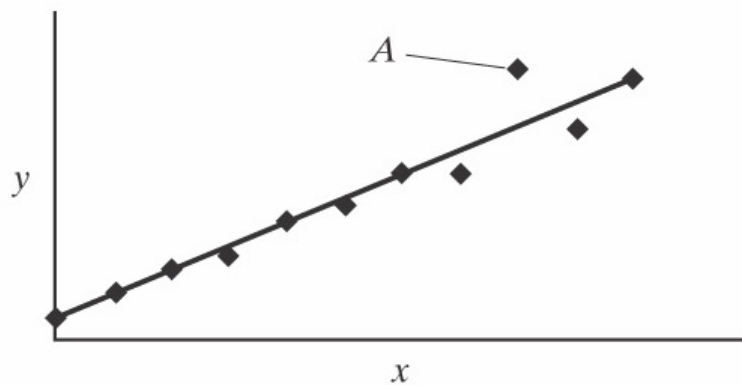
Standard Error of Estimate

- The **standard error of estimate** is analogous to standard deviation in general, but specifically tracks how data points over the entire (x, y) set differ from the best-fit line:

$$S_{y,x} = \sqrt{\frac{\sum y_i^2 - b\sum y_i - a\sum x_i y_i}{n - 2}}$$

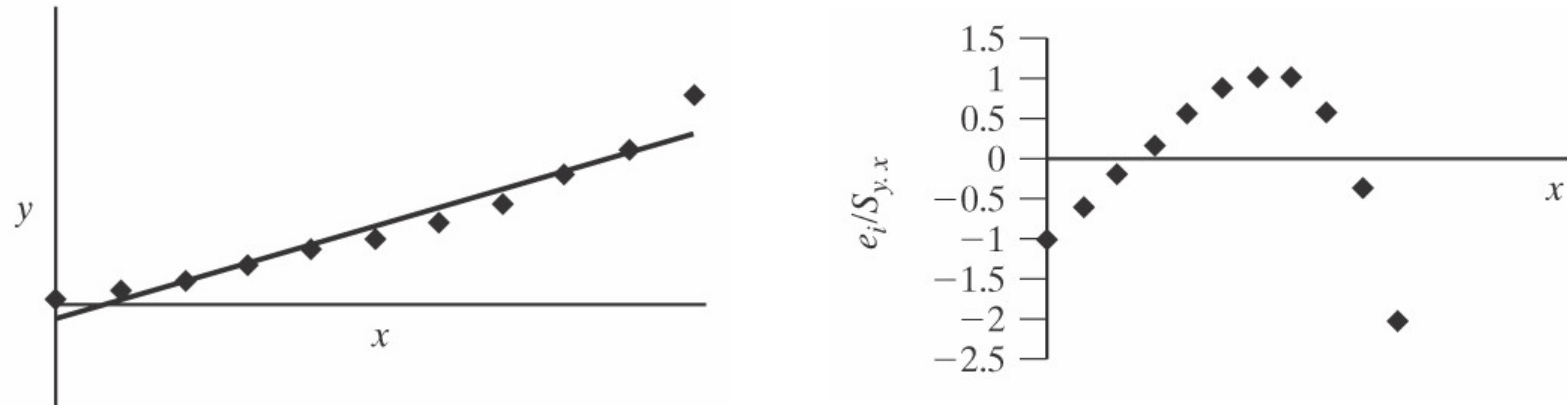
- The coefficient of determination r^2 is more typically used to express goodness-of-fit, but $S_{y,x}$ is useful for other purposes (explained subsequently).

Outliers and “Standardized Residuals”



- ❑ One effective way to identify outliers is to compare each individual residuals $e_i = Y_i - y_i$ in context against the standard error of estimate $S_{y,x}$.
- ❑ This relative comparison will express the residuals in relative context over the whole data set, and make a more “fair” determination whether the observed deviation is extreme or not.

Problems for Identifying Outliers



- ❑ Identifying outliers is a non-trivial problem and does not always obey one straightforward methodology alone. For example, if the “physics” is inherently nonlinear, some data may be incorrectly perceived as outliers.
- ❑ Also, outlier identification using standardized residuals can be misleading if there are insufficient data to have a meaningful standard error of estimate $S_{y,x}$, so other techniques (e.g. Thompson τ) for identifying and rejecting outliers also exist.

Limitations of Linear Regression

- ❑ Assumes random error and normally-distributed variation; ignores systematic error.
- ❑ Is “one-dimensional” along y , and assumes x values are error-free.
- ❑ May be sensitive to outliers, especially for small data sets.
- ❑ Does not apply (and should not be applied) to data that is nonlinear, unless addressed properly by other regression models.

Other Types of Regression

- ❑ Linear regression is not the only approach to obtaining an analytical fit to experimental data.
- ❑ For nonlinear data (such as $y = ae^{bx}$), one approach is to first perform mathematical transformation, followed by a linear fit on the transformed data, as in $\ln(y) = bx + \ln(a)$
- ❑ Regression is not limited to a single (x, y) pair. Higher-dimensional fits can be used to fit analytical expression for y as a function of multiple inputs, as in:
 - $Y = a_0 + a_1x_1 + a_2x_2 + \dots$
- ❑ Regression equations do not have to be restricted to linear fits, it can use higher-order polynomial equations, as in
 - $Y = a_0 + a_1x + a_2x^2 + a_3x^3 + \dots$