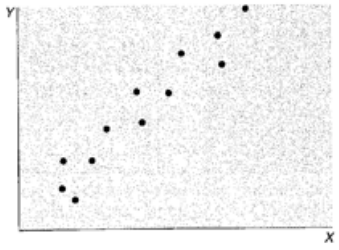
Statistics

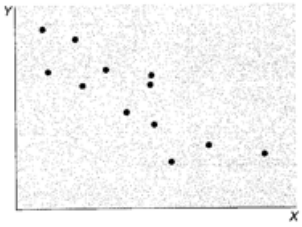
March 4, 2013 Debdeep Pati

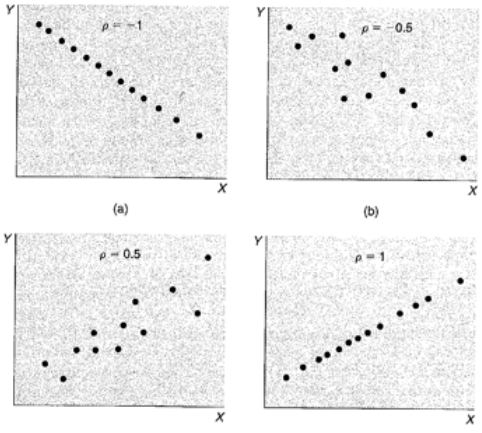
Correlation methods

1. Understand the nature and strength of the association between two measurement variables X and Y.

2. Population correlation coefficient, ρ, quantifies the linear relationship between X and Y.







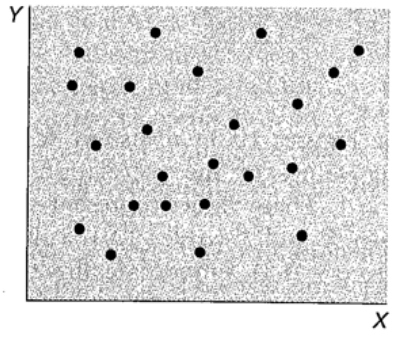
$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

$$-1 \le \rho(X,Y) \le 1$$

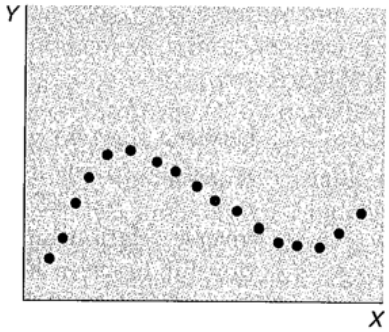
$$\mathrm{Cov}(X,Y) = E((X-\mu)(Y-\nu))$$

3

3. Correlation coefficient is a measure of linear association between X and Y. So corr=0 implies that either there isn't any relationship between X and Y or a possibly non-linear relationship between X and Y.



(a) No relationship between X and Y



(b) Nonlinear relationship between X and Y

4. Correlation coefficient can be affected by truncation, e.g. Relationship between SAT scores measured during the senior year of high school and GPA measured at the completion of the freshman year in college.

5. Correlation coefficient may be affected by confounding variables Relationship between size of a home and its selling price. Association between age of first job and starting salary. Sample correlation coefficient r

6. $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$ Compute the sample correlation coefficient r. Compute the test statistic $t = \frac{r(n-2)^{1/2}}{(1-r^2)^{1/2}}$ which under H_0 follows a t distribution with $n-2$ df. For a two-sided level α test if $t > t_{n-2,1-\alpha/2}$ or $t < -t_{n-2,1-\alpha/2}$ reject H_0, otherwise accept H_0. The p-value is given by

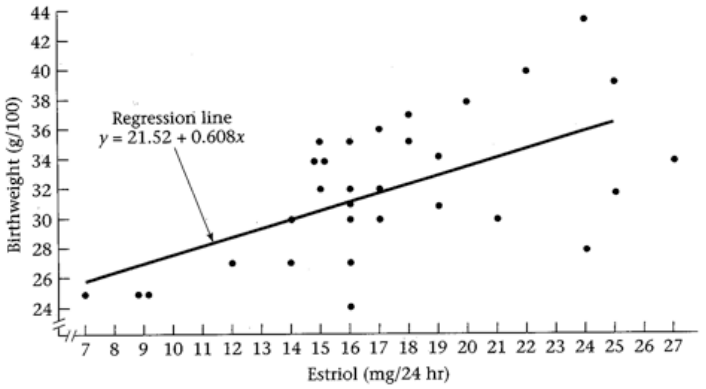
$$\begin{aligned} p &= 2 \times (\text{area to the left of} \, t \;\; \text{under a} \, t_{n-2} \, \text{distribution}) \, if \, t < 0 \\ p &= 2 \times (\text{area to the right of} \, t \;\; \text{under a} \, t_{n-2} \, \text{distribution}) \, if \, t < 0 \end{aligned}$$

7. Suppose serum-cholesterol levels in spouse pairs are measured to determine whether or not there is a correlation between cholesterol levels in spouses. Specifically, we wish to test the hypothesis $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$. Suppose that r = .25 based on 100 spouse pairs. Is this evidence enough to warrant rejecting H_0? We have n = 100, r = .25 t = .25(98)1/2/(1-.252)1/2 = 2.56 qt(.975,98) = 1.98 $t_{60,.99} = 2.39, t_{60,.995} = 2.66, t_{120,.99} = 2.358, t_{120,.995} = 2.617$ .005 ¡ p/2 ¡ .01 or .01 ¡ p ¡ .02. P-value: 2*(1-pt(2.56,98)) = 0.012 H_0 is rejected.

Regression methods

Greene and Touchstone conducted a study to relate birthweight to the estriol level of pregnant women. How can this relationship be quantified?

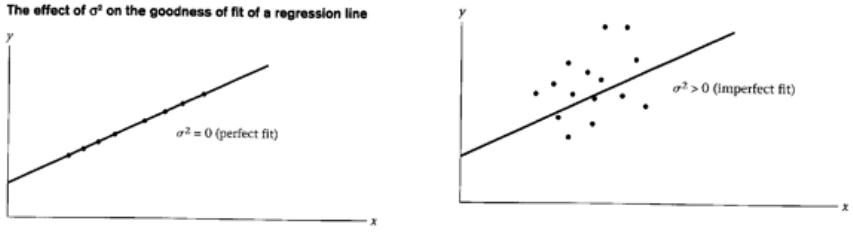
Data from the Greene-Touchstone study relating birthweight and estriol level in pregnant women near term



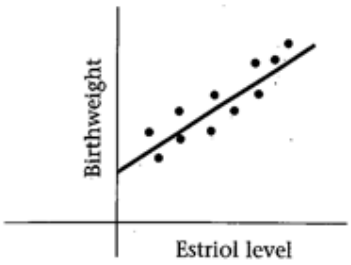
Source: Reprinted with permission of the *American Journal of Obstetrics and Gynecology, 85*(1), 1–9, 1963.

1. $y = \alpha + \beta x + \epsilon$.
2. ϵ is normally distributed with mean 0 and variance σ^2. y is called the dependent variable. x is called the independent variable.

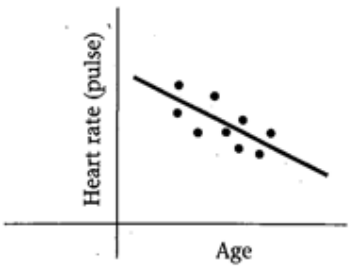
The effect of σ^2 on the goodness of fit of a regression line



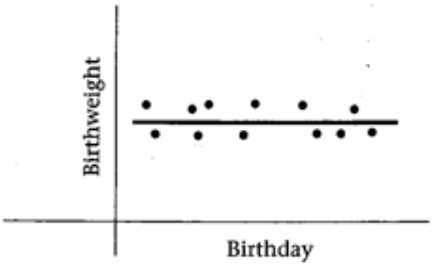
The interpretation of the regression line for different values of β



(a) $\beta > 0$



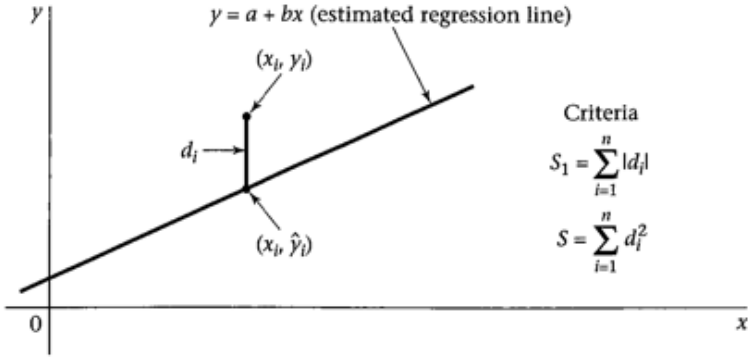
(b) $\beta < 0$



(c) $\beta = 0$

Criteria for fitting a regression line

Possible criteria for judging the fit of a regression line



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

$$a = \bar{y} - b\bar{x}$$

Prediction using Fitted Regression Line

The predicted, or average, value of y for a given value of x, as estimated from the fitted regression line, is denoted by $\hat{y} = a + bx$

R-squared

For any sample point (x_i, y_i), the residual, or residual component, of that point about the regression line is defined by $\hat{y}_i = a + bx_i$. For any sample point (x_i, y_i), the regression component of that point about the regression line is defined by $\hat{y}_i - \bar{y}$. Decompose the total sum of squares (TSS) into regression (Reg(SS) and residual components (Res(SS)).

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Coefficient of determination, $R^2 = Reg(SS)/TSS$. The proportion of variation in the dependent variable y that can be explained by the independent variable x.