Correlation Coefficient

- $r = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \dfrac{SSXY}{\sqrt{SSX \bullet SSY}}$
  - recall that $\hat{\beta}_1 = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \dfrac{SSXY}{SSX}$
  - $\rightarrow \hat{\beta}_1 = r \dfrac{S_Y}{S_X}$

- Properties
  - $-1 \le r \le 1$
  - r does not depend on units of measurement (dimensionless)
  - $r \approx 0 \rightarrow$ no *linear* relationship between X and Y
  - larger $|r|$ means a stronger *linear* relationship
  - $r^2 = \dfrac{\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = \dfrac{SSY - SSE}{SSY}$

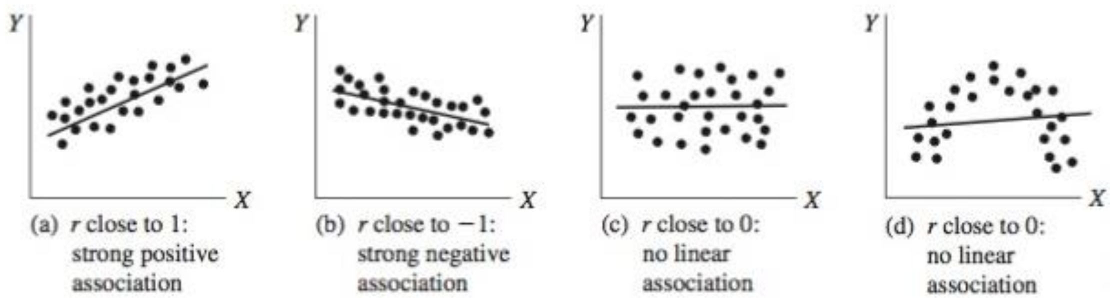


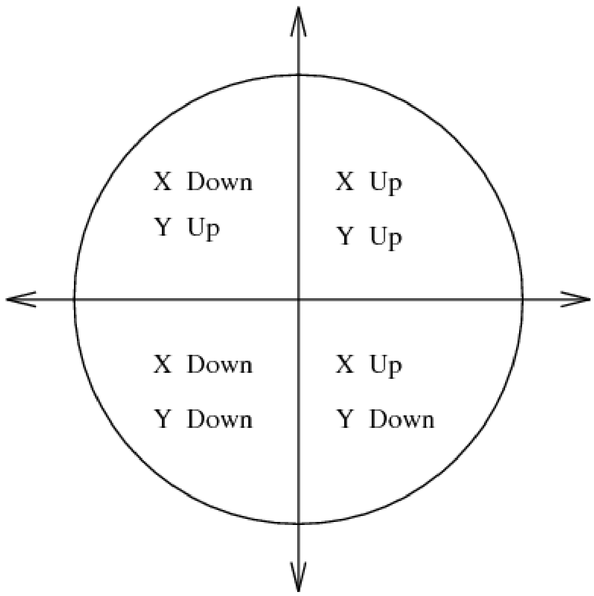
FIGURE 6.1 Correlation coefficient as a measure of association

Transformations and Tukey's "Rule of the Bulge"

- Observe which way the curve bulges as suggested by a scatterplot of the data
- Transform ***y*** or ***x*** (or both) according to the signs of the corresponding quadrant:

  up → powers >1

  down → powers <1 (including logarithm)



Correlation Coefficient - Inference

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{SSXY}{\sqrt{SSX \bullet SSY}} \rightarrow \hat{\beta}_1 = r \frac{S_Y}{S_X}$$

- Testing Hypotheses about **ρ**, the *population* correlation coefficient
- H_0: **ρ=0**
  - Sampling distribution of ***r*** is symmetric & approx. normal
  - $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ with *n-2* d.f. is equivalent to testing H_0: $\beta_1 = 0$ with $t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$

- H_0: **ρ= ρ₀** where $\rho_0 \neq 0$
  - Sampling distribution of ***r*** is *NOT* symmetric or approx. normal
  - *Fisher's "Z-transformation"* gives an approx. normally distributed statistic

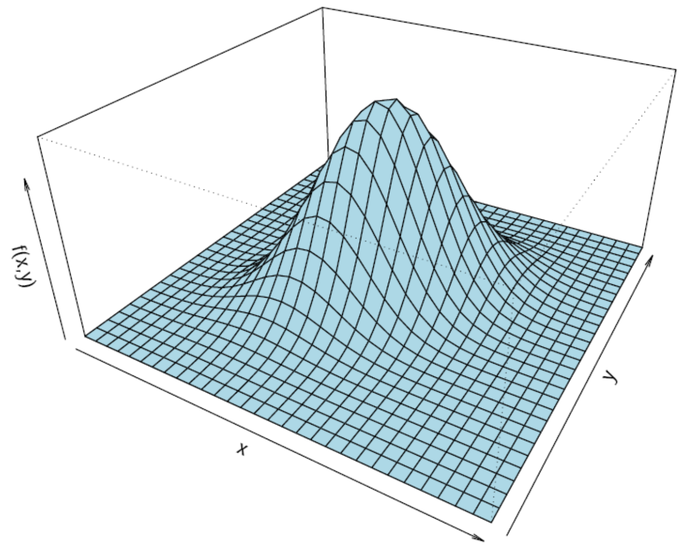
$$\frac{1}{2} \bullet \ln\left(\frac{1+r}{1-r}\right) \sim N\left(\frac{1}{2} \bullet \ln\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{n-3}\right)$$

  - Testing H_0: $\rho = \rho_0$ $Z = \frac{\frac{1}{2} \bullet \ln\left(1+r/1-r\right) - \frac{1}{2} \bullet \ln\left(1+\rho_0/1-\rho_0\right)}{1/\sqrt{n-3}} \sim N(0,1)$
  - 100(1-α)% CI for $\frac{1}{2} \bullet \ln\left(\frac{1+\rho}{1-\rho}\right)$ **:** $\frac{1}{2} \bullet \ln\left(\frac{1+r}{1-r}\right) \pm z_{\alpha/2} \bullet \frac{1}{\sqrt{n-3}} = \left(L_z, U_z\right)$

  - 100(1-α)% CI for **ρ:** (L_ρ, U_ρ) = (lower endpoint, upper endpoint)
    - Find L_z & U_z above and solve for L_ρ & U_ρ in $L_z = \frac{1}{2} \bullet \ln\left(\frac{1+L_\rho}{1-L_\rho}\right)$ $U_z = \frac{1}{2} \bullet \ln\left(\frac{1+U_\rho}{1-U_\rho}\right)$
    - → $\left(L_\rho, U_\rho\right) = \left(\frac{e^{2L_z}-1}{e^{2L_z}+1}, \frac{e^{2U_z}-1}{e^{2U_z}+1}\right)$
- H_0: $\rho_1 = \rho_2$ (for 2 *independent* samples)
  - let $W_1 = \frac{1}{2} \bullet \ln\left(\frac{1+r_1}{1-r_1}\right)$ and $W_2 = \frac{1}{2} \bullet \ln\left(\frac{1+r_2}{1-r_2}\right)$
  - $Z = \frac{W_1 - W_2}{\sqrt{1/(n_1-3) + 1/(n_2-3)}} \sim N(0,1)$

Bivariate Normal Distribution

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}e^{\frac{-1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_x}{\sigma_x}\right)^2-2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)+\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right)}$$



$$E(Y\mid X=x)=\mu_y+\rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)$$

$$\hat{\mu}_{Y|X=x}=\bar{y}+r\frac{s_y}{s_x}(x-\bar{x})$$