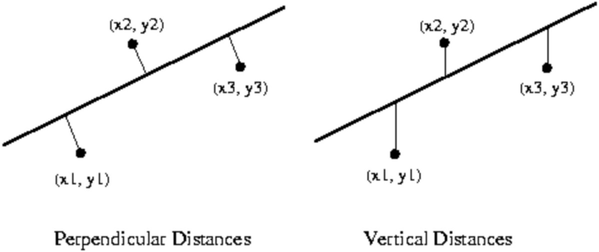
Correlation

Miaoyan Wang

Department of Statistics
UW Madison

Recall: brainstorm

Why do we use vertical distance to define the fitted line?



Other choices?

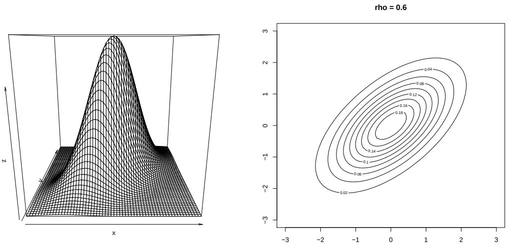
- The sum of the squares of perpendicular distance
- The sum of absolute value of the distance

Correlation Model

- In linear regression, we model and predict Y given $X = x$.
- If interested in how two variables are related to each other, X and Y are to be treated symmetrically.
- Let X and Y both be random and have a bivariate distribution.
- A useful distribution is a **bivariate normal distribution** with a probability density that is parameterized by
  - μ_Y and σ_Y: the mean and the SD of the marginal distribution of Y
  - μ_X and σ_X: the mean and the SD of the marginal distribution of X
  - ρ_{YX} (or ρ): the **coefficient of correlation** between Y and X

Correlation Model

The probability density surface can be plotted using a 3D or contour plot.



Properties for bivariate normal (homework):

- Marginal distribution: $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$.
- Conditional distribution: $Y|X = x \sim N(\alpha + \beta x, \sigma_{Y|x}^2)$ where

$$\alpha = \mu_Y - \mu_X \rho \frac{\sigma_Y}{\sigma_X}, \beta = \rho \frac{\sigma_Y}{\sigma_X}, \sigma_{Y|x}^2 = \sigma_Y^2 (1 - \rho^2).$$

Population Correlation Coefficient

- The **population correlation coefficient** (also called Pearson correlation coefficient) between X and Y is

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

- ρ is a measure of linear relationship between X and Y.
- $-1 \leq \rho \leq 1$.
- $\rho = 1$ indicates perfect positive correlation.
- $0 < \rho < 1$ indicates modest positive correlation.
- $\rho = 0$ indicates no linear relationship.
- $-1 < \rho < 0$ indicates modest negative correlation.
- $\rho = -1$ indicates perfect negative correlation.

Example 1

Correlation coefficients

Example 2

Correlation coefficients

Example 3

Correlation coefficients

Pearson's Sample Correlation Coefficient

- Based on the data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, the **sample correlation coefficient**

$$\hat{\rho} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

  estimates ρ.
- Note the symmetry between X and Y in $\hat{\rho}$.
- Sample covariance

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

- $\hat{\rho} = \frac{s_{xy}}{\sqrt{s_x s_y}}$ estimates the Pearson correlation coefficient, where s_x, s_y denote the sample covariance for X, Y, respectively.
- Connection to slope in simple linear regression? Recall $\hat{\beta}_1 = \frac{s_{xy}}{s_x}$.

Independence

Let (X, Y) be a bivariate random variable in $\mathbb{R}^2$.

- Independence:

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x)\mathbb{P}(Y \leq y), \quad \text{for all } x, y \in \mathbb{R}.$$

- Uncorrelated:

$$\mathrm{Cov}(X, Y) = 0.$$

- Independence $\longrightarrow$ uncorrelated, but not vice versa.
- Pearson correlation coefficient is a measure of the strength of linear dependence between two random variables.
- If $Y = aX + b$, then $\rho(X, Y) = 1$ when $a > 0$, and $\rho(X, Y) = -1$ when $a < 0$.

Linear independence

Correlation coefficients

Statistical Inference on ρ

- Assume X and Y are from a bivariate normal distribution.
- Define Fisher's transformation

$$\lambda(\rho) = \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right) = \text{arctanh}(\rho),$$

- Fisher R.A. (1915) shows that

$$\lambda(\hat{\rho}) = \frac{1}{2}\ln\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right) \approx N\left(\lambda(\rho), \frac{1}{n-3}\right).$$

- An approximate $(1-\alpha)$ CI for $\lambda(\rho)$ is

$$\lambda(\hat{\rho}) \pm z_{\alpha/2}\sqrt{\frac{1}{n-3}} = \left[\hat{\lambda}_1, \hat{\lambda}_2\right].$$

- An approximate $(1-\alpha)$ CI for ρ is

$$\left(\frac{e^{2\hat{\lambda}_1}-1}{e^{2\hat{\lambda}_1}+1} \equiv\right) tanh(\hat{\lambda}_1) \le \rho \le tanh(\hat{\lambda}_2) \left(\equiv \frac{e^{2\hat{\lambda}_2}-1}{e^{2\hat{\lambda}_2}+1}\right).$$

Example: Wetland Species Richness

- From the summary statistics, we have

$$\hat{\rho} = \frac{-10.775}{\sqrt{2.3316}\sqrt{528.84}} = -0.307.$$

- Find the Fisher's transformation

$$\lambda(\hat{\rho}) = \frac{1}{2}\log\left\{\frac{1+(-0.307)}{1-(-0.307)}\right\} = -0.3172$$

- An approximate 95% CI for $\lambda(\rho)$ is

$$(-0.3172) \pm 1.96 \times \sqrt{\frac{1}{55}} = [-0.582, -0.0529]$$

- An approximate 95% CI for ρ is

$$\frac{e^{2(-0.582)}-1}{e^{2(-0.582)}+1} \leq \rho \leq \frac{e^{2(-0.0529)}-1}{e^{2(-0.0529)}+1}$$

which is $[-0.524, -0.0528]$.

Remarks on $\hat{\rho}$

Correlation $\neq$ Causation