ضروریات آمار برای علم داده

گمشدگی داده‌ها و برازش رگرسیون

گمشدگی داده‌ها

$$D = \{A_1, A_2, \ldots, A_r\}$$

$$A_j = \{A_j^{obs}, A_j^{mis}\}$$

$$D = \{D^{obs}, D^{mis}\}$$

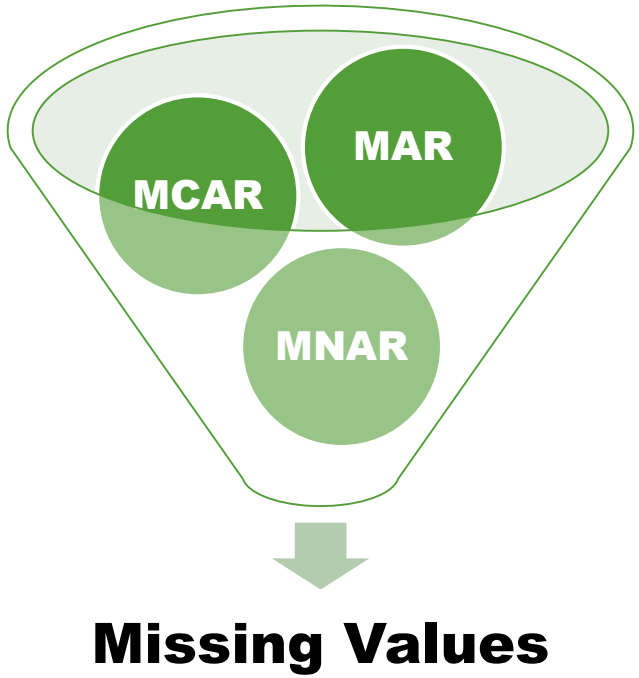
$$R_{ij} = \begin{cases} 0 & if\ v_{ij} \text{ is missing} \\ 1 & if\ v_{ij} \text{ is observed} \end{cases}$$

انواع گمشدگی داده‌ها

$Pr(R|D^{mis},D^{obs})=Pr(R)$ گمشدگی کاملا تصادفی

$Pr(R|D^{mis},D^{obs})=Pr(R|D^{obs})$ گمشدگی تصادفی

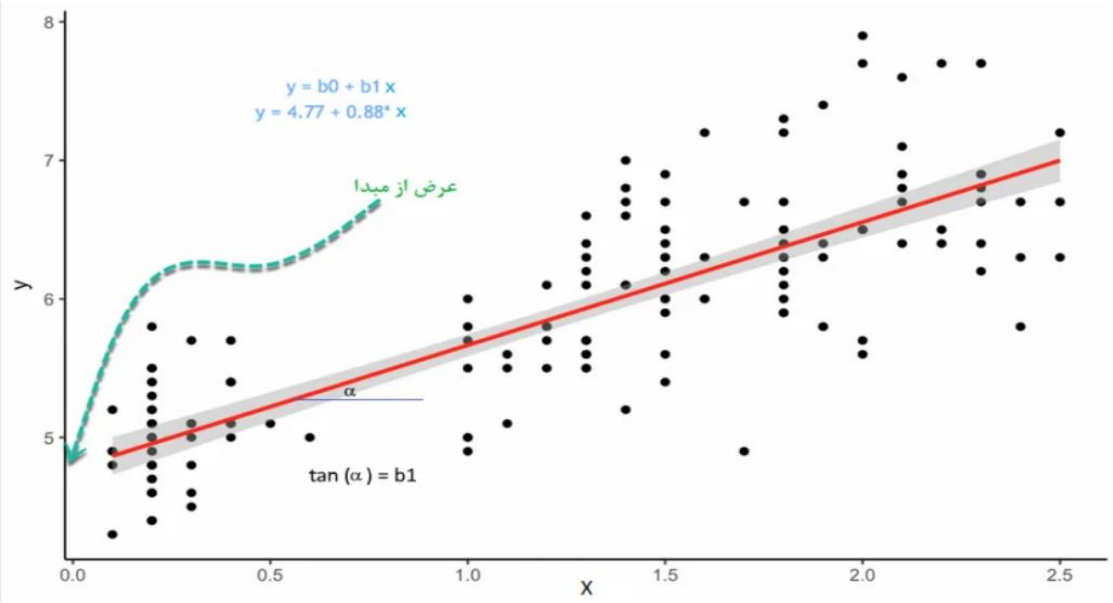
$Pr(R|D^{mis},D^{obs})$ گمشدگی غیرتصادفی



روش‌های مواجهه با داده‌های ناکامل

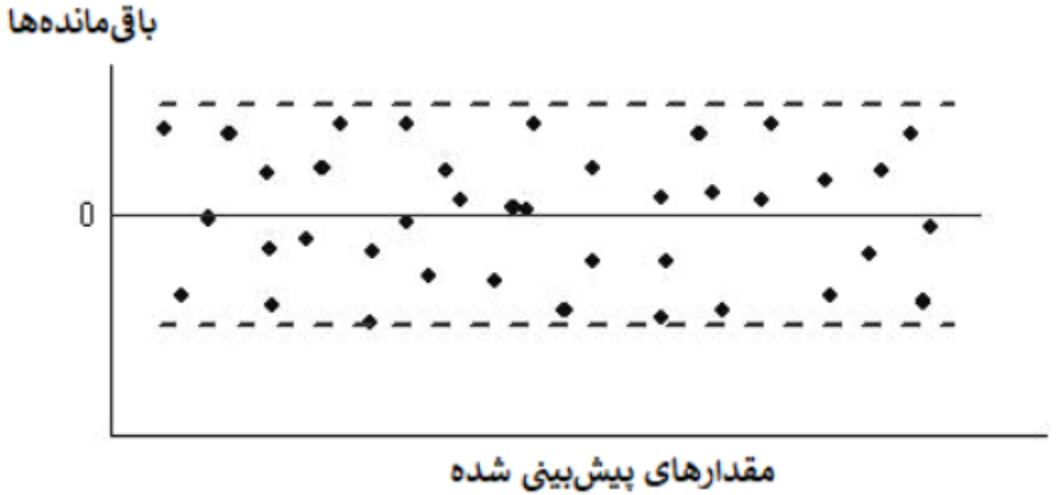
- تحلیل موارد کامل
- تحلیل موارد موجود وزن‌دار شده
- روش‌های مبنی بر درستنمایی
- روش جانهی
  میانگین
  میانگین + عامل تصادفی
  رگرسیون

رگرسیون خطی

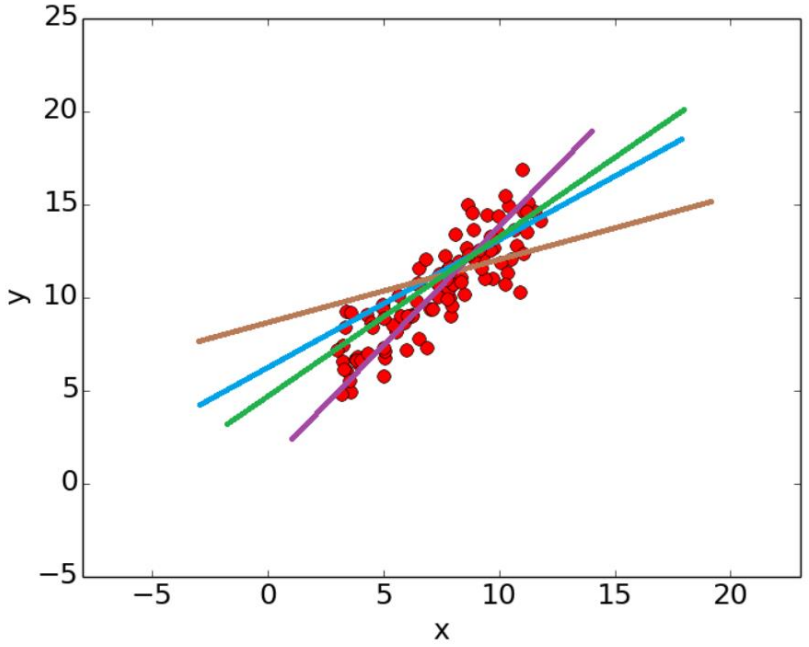


$$y = \beta_0 + \beta_1 x + \epsilon$$

- ✓ جملات خطا از یکدیگر مستقل و دارای توزیع نرمال هستند.
- ✓ میانگین جمله خطا صفر است.
- ✓ واریانس هر مولفه از جمله خطا ثابت و برابر با $\sigma 2$ است.



براورد پارامترهای رگرسیون خطی



$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \qquad \beta_1 = \frac{cov(x, y)}{var(x)}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

مینیم سازی مجموع مربعات خطا به منظور براورد پارامترهای مدل

$$\sum \epsilon^2$$

$$\sum (y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i))^2$$

$$\sum (-y_i + \widehat{\beta}_0 + \widehat{\beta}_1 x_i) = 0$$

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

$$-\sum (x_i y_i + (\bar{y} - \widehat{\beta}_1 \bar{x}) \sum x_i + \widehat{\beta}_1 \sum x_i^2) = 0$$

$$\widehat{\beta_1} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

مجموعه داده ۵۰ خانه شامل قیمت (به میلیون ریال) و متراژ (متر مربع) در شهر تهران جمع‌آوری شده است.

از مجموعه داده mtcarsکه اطلاعات و ویژگی‌هایی مختلف ۳۲ خودرو در آن ثبت شده کمک می‌گیریم و مدل رگرسیونی چند گانه را برای پیش‌بینی مسافت طی شده با یک گالن سوخت ایجاد می‌کنیم. این مجموعه داده به طور خودکار در Rبارگذاری شده است.

رگرسیون لجستیک

$$p(x) = \hat{Y} = E(Y = 1|X = x) = \frac{e^{b_0+b_1x}}{1 + e^{b_0+b_1x}}$$