Lecture 12

§1 Linear regression（线性回归）

1、模型选取

1° 由散点图推测变量 X 与 Y 呈 linear dependent（线性相关）

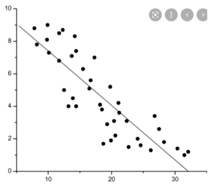
2° Y 的取值集中于 $\beta_0+\beta_1 X$ 附近（β_0, β_1 为未知系数）

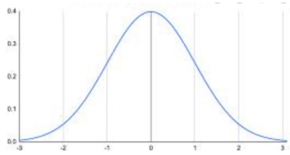
3° 选取正态分布模型

$$Y \sim N(\beta_0+\beta_1 X, \sigma^2)$$

From data, we observe that

- They are more likely to be linearly dependent with each other.
- Y is centralized at some value $\beta_0+\beta_1 X$.





2、选择最佳模型（利用 MLE 确定最佳的系数 β_0, β_1）

- 给定 σ^2（由观测可确定 σ^2 较小，视为已知量）

  Samples：$(X_1, Y_1), (X_2, Y_2), \cdots\cdots, (X_N, Y_N)$

  PDF for normal:

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

- 可得出 likehood 为

$$L(\beta_0, \beta_1) = \frac{1}{(\sqrt{2\pi})^n \cdot \sigma^n} \cdot \exp\left[-\frac{1}{2} \cdot \frac{\sum_i (Y_i-\beta_1 X_i-\beta_0)^2}{\sigma^2}\right]$$

  要求 $\max L(\beta_0, \beta_1)$，只需要 minimize $\sum_i (Y_i-\beta_1 X_i-\beta_0)^2$

- 分别对 β_0 与 β_1 求偏导，得：

$$\begin{cases} \sum_i (Y_i-\beta_1 X_i-\beta_0) \cdot X_i = 0 \\ \sum_i (Y_i-\beta_1 X_i-\beta_0) = 0 \end{cases}$$

  得：

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_i (X_i-\bar{X})(Y_i-\bar{Y})}{\sum_i (X_i-\bar{X})^2} \\ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \end{cases}$$

3、Residual（随机误差）analysis：检验假设

1° 模型的两种表达方式

① $Y \sim N(\beta_0+\beta_1 X, \sigma^2)$

② $Y-\beta_0-\beta_1 X \sim N(0, \sigma^2)$

2° Residual e_i / ε_i

$$e_i = Y_i - \beta_0 - \beta_1 X_i$$

3° 通过 residuals 检验回归分析成立的假设

① 假设一：X与Y为 linear relationship

检验：e_i does not depend on X_i

② 假设二：对任意X，$Y-\beta_0-\beta_1 X$ 的方差 (σ^2) 均相同

(homogeneity (均一性) of variances)

检验：variance of e_i does not depend on X_i

4° Graphical analysis of residuals