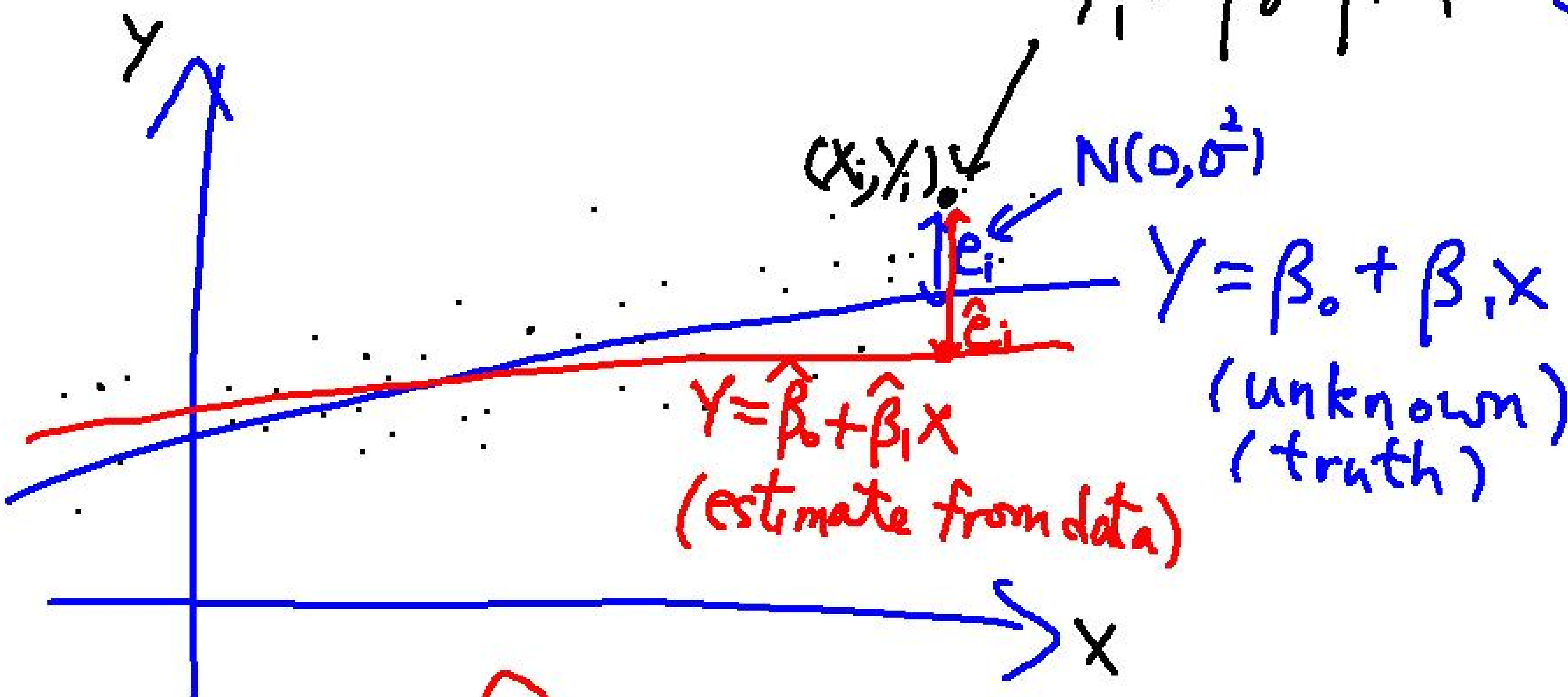
$$Y_i = \beta_0 + \beta_1 X_i + e_i$$




Least Sq

$$RSS = \sum (Y_i - \beta_0 - \beta_1 X_i)^2$$

minimize RSS w.r.t $\beta_0$ $\beta_1$

$$\Rightarrow \begin{cases} \dfrac{\partial RSS}{\partial \beta_0} = 0 \\ \dfrac{\partial RSS}{\partial \beta_1} = 0 \end{cases}$$

$$\Rightarrow \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{SXY}{SXX}$$

$E(\hat{\beta}_0) = \beta_0$ $\quad E(\hat{\beta}_1) = \beta_1$

$Var(\hat{\beta}_0) = ?$ $\quad Var(\hat{\beta}_1) = \dfrac{\sigma^2}{SXX}$

☆☆ express

$\hat{\beta}_1 = \sum a_i Y_i$

depends on X

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-2}$$

$$E(\hat{\sigma}^2) \overset{Ex\,2}{=} \sigma^2$$

# Analysis of variance (ANOVA)

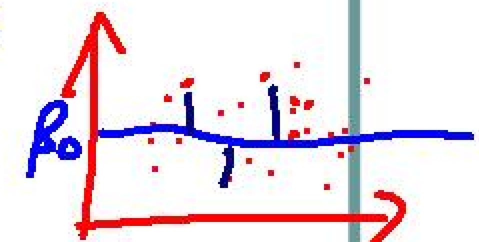
- Regression is the study of dependence of variables
  - $y_i = \beta_0 + \beta_1 x_i + e_i$
    - $\beta_1 = 0 \rightarrow$ x and y are not dependent
    - $\beta_1 \neq 0 \rightarrow$ x and y are ~~not~~ dependent
- Question:
  - Are x and y dependent?
- Answer:
  - Method 1) test whether $\beta_1 = 0$
  - Method 2) Compare the two models

Model1 $\beta_1 = 0$ → • $E(y|x) = \beta_0$ i.e. $y_i = \beta_0 + e_i$

• $E(y|x) = \beta_0 + \beta_1 x$ i.e. $y_i = \beta_0 + \beta_1 x_i + e_i$ ← Model 2 : $\beta_1$ free

# Analysis of variance (ANOVA)

- Analysis of variance (ANOVA) is a method that compares two models of mean functions
  - $E(y|x)=\beta_0$
  - $E(y|x)=\beta_0+\beta_1 x$




- For the first model: $E(y|x)=\beta_0$
  - $\beta_0$ can be estimated by minimizing $\sum(y_i-\beta_0)^2$ ← $RSS(\beta_0)$
    - differentiate w.r.t. $\beta_0$ gives $\hat{\beta}_0=\bar{y}$
  - Residual sum of square RSS is

$$\sum(y_i-\hat{\beta}_0)^2=\sum(y_i-\bar{y})^2=SYY$$

$$\frac{\partial RSS}{\partial \beta_0}=-2\sum(y_i-\beta_0)=0$$

$$\Rightarrow \hat{\beta}_0=\frac{\sum y_i}{n}=\bar{y}$$

$Sxx = \Sigma(x_i - \bar{x})^2$

- Compare
  - $E(y|x)=\beta_0$
  - $E(y|x)=\beta_0 + \beta_1 x$
- For the first model: $E(y|x)=\beta_0$
  - Residual sum of square, $RSS_1$, is

$$\sum (y_i - \hat{\beta}_0)^2 = \sum (y_i - \bar{y})^2 = SYY$$

- For the second model: $E(y|x)=\beta_0 + \beta_1 x$
  - Residual sum of square, $RSS_2$, is

$$\sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 = SYY - \frac{(SXY)^2}{SXX}$$

- $RSS_1 > RSS_2$… Is the 2nd model always better?

# Analysis of variance (ANOVA)

- Difference sum of square due to regression (SSreg)
  - SSreg = RSS$_1$-RSS$_2$= $\frac{(SXY)^2}{SXX}$ $= \left[\frac{\sum(x_i-\bar{x})(y_i-\cancel{\bar{y}})}{\sqrt{SXX}}\right]^2$
  - Large SSreg ➔ 2$^{nd}$ model explains much more variation
  - How large is large?
- Study the distribution of SSreg under model 1 (idea)
  - After some algebra, $\boxed{SSreg = [\sum(\frac{x_i-\bar{x}}{\sqrt{SXX}})y_i]^2}$
  - By CLT, $\sum(\frac{x_i-\bar{x}}{\sqrt{SXX}})y_i$ is approximately $N(0,\sigma^2)$
  - SSreg ~ $\sigma^2$ $(N(0,1))^2 = \sigma^2 X^2_1$ (Chi-square with d.f. 1)
  - $\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{n-2}$ ~ $\sigma^2 X^2_{n-2}/(n-2)$ (Chi-square with d.f. n-2)
  - SSreg/$\hat{\sigma}^2$ ~ $X^2_1 /(X^2_{n-2}/(n-2))$=F(1,n-2)

# Analysis of Variance (ANOVA)

- ANOVA table: a break-down of squares (variation)

$SYY - SXY^2/SXX$

| Source | df | SS | MS | F | p-value |
|---|---|---|---|---|---|
| Regression | 1 | $SSreg$ | $SSreg/1$ | $MSreg/\hat{\sigma}^2$ | |
| Residual | $n-2$ | $RSS$ | $\hat{\sigma}^2 = RSS/(n-2)$ | | |
| Total | $n-1$ | $SYY$ | | | |

checking $F(1, n-2)$

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

$$TSS = SYY = RSS + SSreg$$

**Variation of the data** = **Variation not explained by regression** + **Variation explained by regression**