# Linear Regression

## Simple Linear Regression Model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

**Notation:**

RANDOM

$y_i$ = $i^{th}$ observed value of the response variable

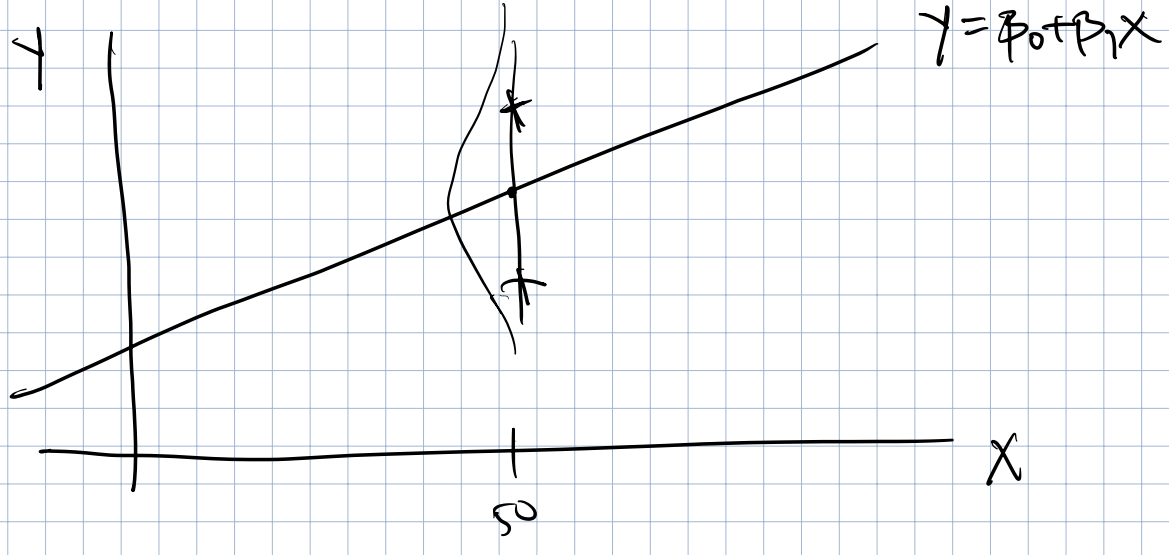
$x_i$ = $i^{th}$ " " of the predictor variable — FIXED

$n$ = sample size — FIXED

$\beta_0$ = unknown intercept parameter — FIXED

$\beta_1$ = unknown slope parameter — FIXED

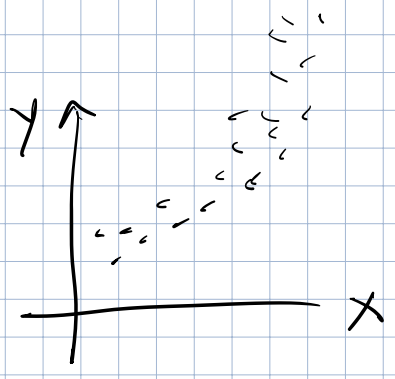
$\varepsilon_i$ = random error terms — RANDOM



Observations:

① "Simple": one predictor (x) variable

② "linear": $y_i$ is a linear f'n of $\beta_0$ & $\beta_1$

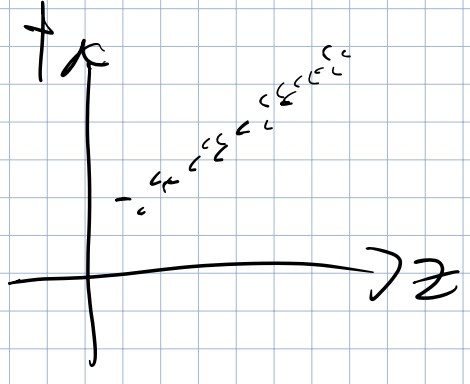



Ex:

① $y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$

↓ $z_i = x_i^2$

$y_i = \beta_0 + \beta_1 z_i + \varepsilon_i$

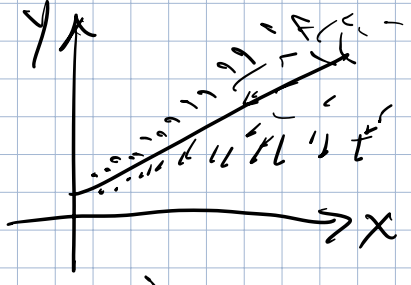



② $y_i = \beta_0 + \beta_1 \log x_i + \varepsilon_i$

③ $y_i = \frac{\beta_0}{\beta_1 + x_i} + \varepsilon_i$

## Model Assumptions

- $x_i$'s are fixed
  - ↳ (if not fixed, need random-effects model.)




- $\varepsilon_i$'s must satisfy:

ASSUME

i. $E(\varepsilon_i) = 0$

ii. $Var(\varepsilon_i) = \sigma^2$ (constant in X)

iii. $Cor(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$

↳ errors are uncorrelated

CLASSICAL

iv. $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

↳ useful for inference

## 3. Regression Function

The regression function is

$$g(x) = E(Y \mid X = x)$$

↳ if simple & linear assumptions are made

$$= \beta_0 + \beta_1 X$$

The goal of estimating $g(x)$ can be reduced to

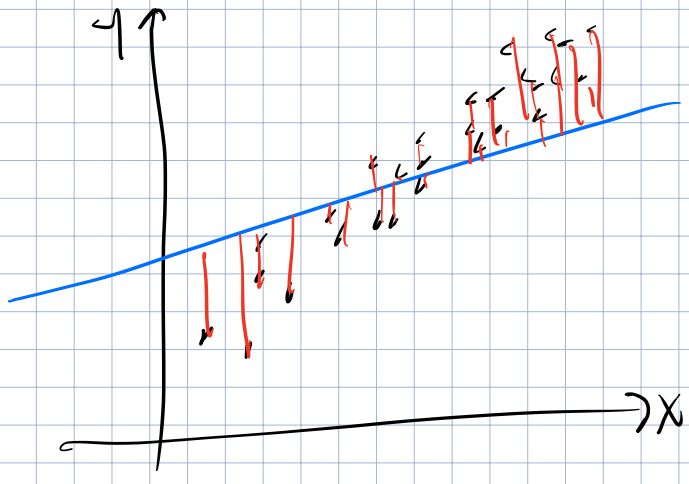
estimating $\beta_0$ & $\beta_1$.

How to estimate $\beta_0$ & $\beta_1$?

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

The Least Squares Principle is one way to do this.



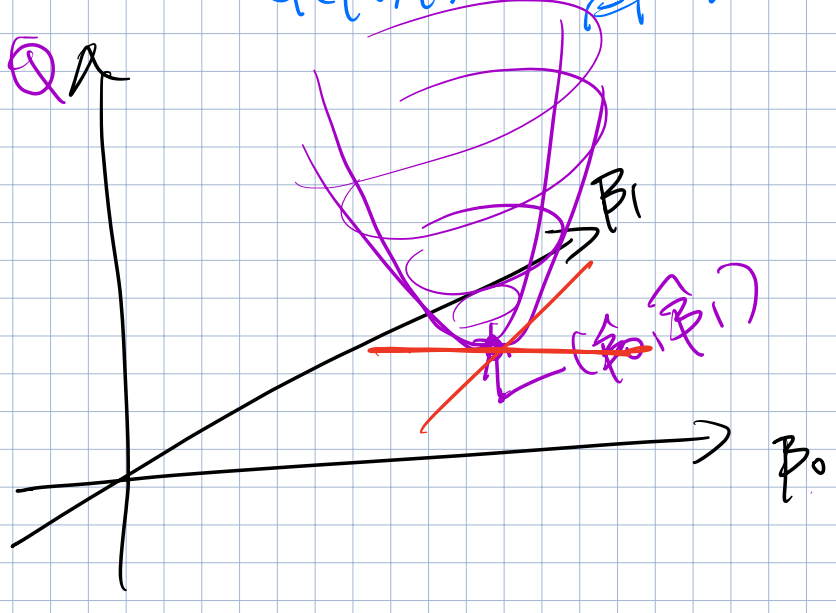

↑ Find the line $(\hat{\beta}_0, \hat{\beta}_1)$ that minimize

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2$$

$$= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

How can I minimize

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$




① $\frac{\partial Q}{\partial \beta_0} \overset{!}{=} 0$

② $\frac{\partial Q}{\partial \beta_1} \overset{!}{=} 0$

① $\frac{\partial Q}{\partial \beta_0} = \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

$= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_0} (y_i - \beta_0 - \beta_1 x_i)^2$

$= \sum_{i=1}^{n} 2 (y_i - \beta_0 - \beta_1 x_i)(0 - 1 - 0)$

$= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \stackrel{!}{=} 0$

what does this imply?

② $\frac{\partial Q}{\partial \beta_1} = \frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

$= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_1} (y_i - \beta_0 - \beta_1 x_i)^2$

$= \sum_{i=1}^{n} 2 (y_i - \beta_0 - \beta_1 x_i)(0 - 0 - x_i)$

$= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i \stackrel{!}{=} 0$

① $-2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)\overset{!}{=}0$

$\bar{y}=\frac{\Sigma_i y_i}{n}$

$n\bar{y}=\Sigma_i y_i$

$\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)=0$

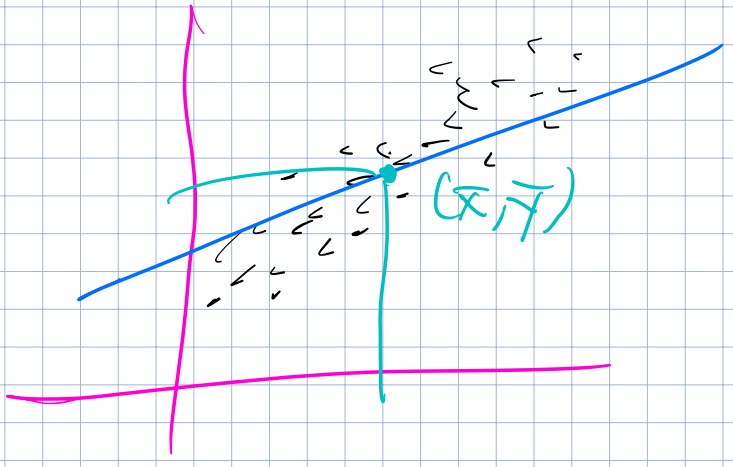
$\Sigma_i y_i-\Sigma_i\beta_0-\Sigma_i\beta_1 x_i=0$

$n\bar{y}-n\beta_0-\beta_1 n\bar{x}=0$

$n\bar{y}=n\beta_0+n\beta_1\bar{x}$

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

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




② $-2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)(x_i)\overset{!}{=}0$

$\sum_{i=1}^{n}(x_i y_i-\beta_0 x_i-\beta_1 x_i^2)=0$

$\sum_{i=1}^{n}x_i y_i-\beta_0\sum_{i=1}^{n}x_i-\beta_1\Sigma_i x_i^2=0$

$\Sigma_i x_i y_i-(\bar{y}-\beta_1\bar{x})(\underbrace{\Sigma_i x_i}_{n\bar{x}})-\beta_1\Sigma_i x_i^2=0$

$$\sum_i x_i y_i - (n\bar{x}\bar{y} - n\beta_1 \bar{x}^2) - \beta_1 \sum_i x_i^2 = 0$$

$$\sum_i x_i y_i - n\bar{x}\bar{y} + n\beta_1 \bar{x}^2 - \beta_1 \sum_i x_i^2 = 0$$

$$\sum_i x_i y_i - n\bar{x}\bar{y} = \beta_1 \sum_i x_i^2 - \beta_1 n\bar{x}^2$$

$$= \beta_1 \left(\sum_i x_i^2 - n\bar{x}^2\right)$$

$$\hat{\beta}_1 = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2}$$

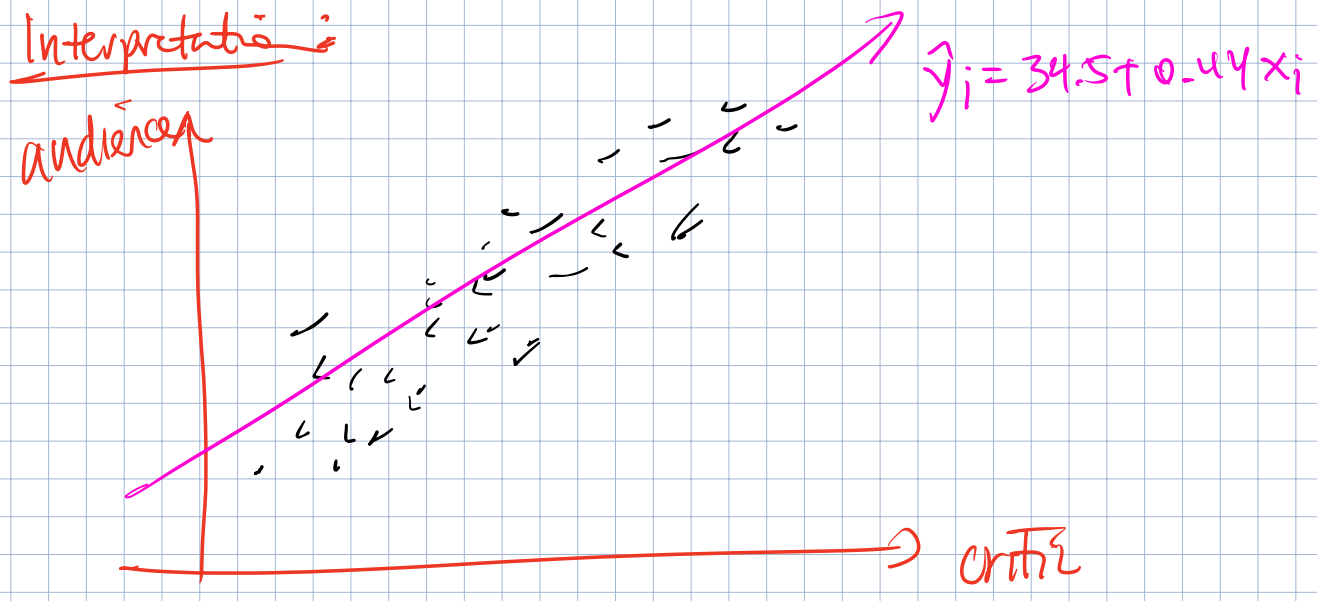
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Least Squares Estimates of $\beta_0$ & $\beta_1$

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{(n-1)\,\widehat{Cov}(X,Y)}{(n-1)\,\widehat{Var}(X)}$$

$$= \frac{\frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2} = \frac{\widehat{Cov}(X,Y)}{\widehat{Var}(X)}$$

Interpretation:

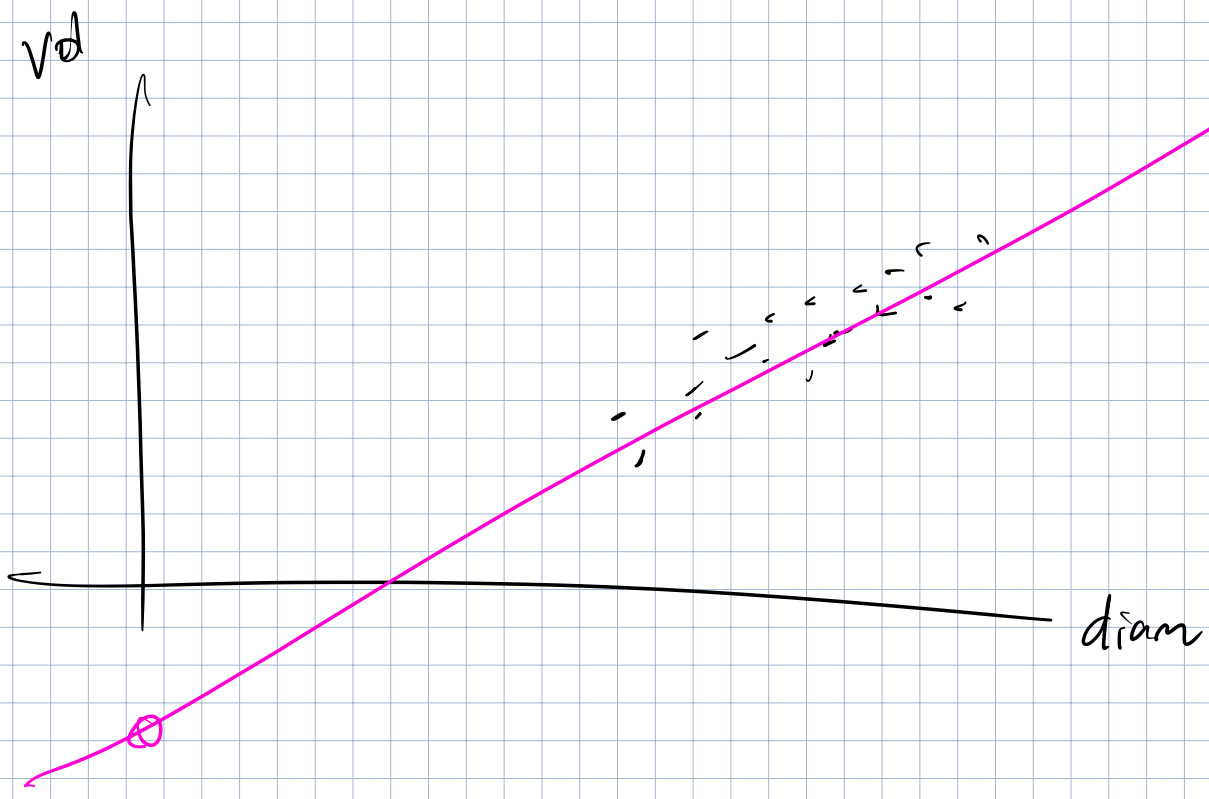



Slope: Interpret 0.44 in the movies ex

"For every additional point a critic give a movie, we expect the audience rating to increase by 0.44 points, on average."

Intercept: 34.5

We expect on average a movie rated as a 0 by a critic to have an audience rating of 34.5.


Vd
diam