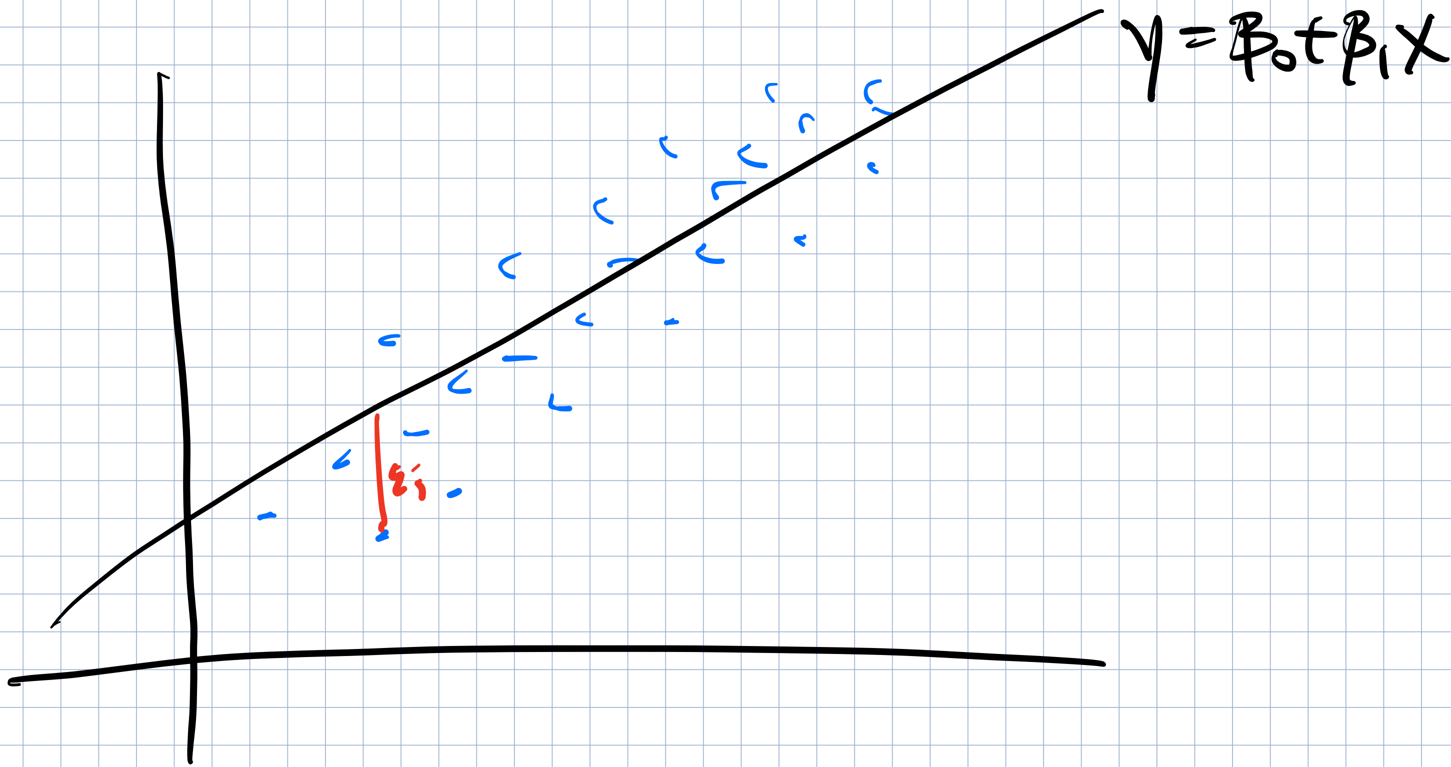
Regression

I. Simple Linear Regression Model

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{} + \varepsilon_i, \qquad i = 1, \ldots, n$$

- y_i: i^{th} observed value of response variable Y
- x_i: i^{th} observed value of predictor variable X
- n: sample size
- β_0: unknown intercept parameter
- β_1: unknown slope parameter
- ε_i: random error term



Note:

1. classic = standard set of assumption
2. simple = only 1 predictor
3. Linear = y_i is a linear function of β_0 and β_1

Ex:

- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ (✓ linear)

- $y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$ (✓ linear)

$z_i = x_i^2$

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

- $y_i = \beta_0 + \beta_1 \log x_i + \varepsilon_i$ (✓ linear)

- $y_i = \dfrac{\beta_0}{\beta_1 + x_i} + \varepsilon_i$ (✗ not linear)

2. Model Assumptions

- ε_i's are RANDOM error terms that satisfy:

BASIC

(i) $E(\varepsilon_i)=0$

(ii) $Var(\varepsilon_i)=\sigma^2$ — constant

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

↳ no correlations b/w errors

CLASSICAL

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

↳ useful for inference

- X_i's are fixed
  ↳ we don't need to worry about X's dist.

(If X_i's are random, use random effects model.)

3. Regression Function

The regression function is

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

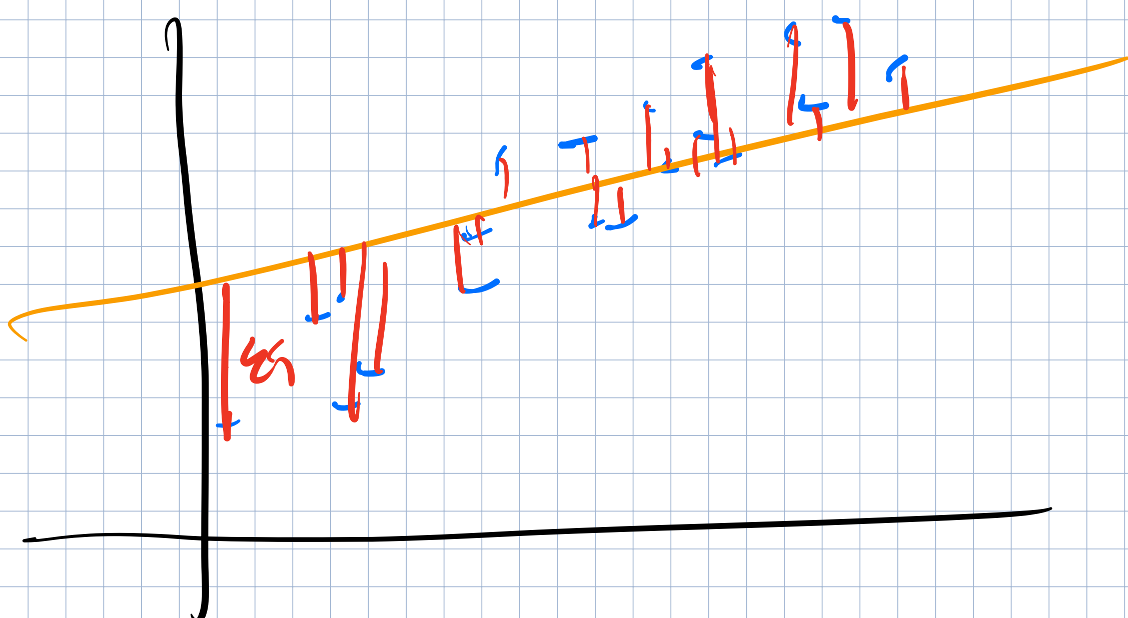
$$= \beta_0 + \beta_1 X$$

g is our target!

This is equivalent (under SLR assumptions) to the task of estimating:

① β_0 & ② β_1.

The <u>Least Squares Principle</u> is how we try to solve this problem!



Try to minimize

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

"sum of squared errors"

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

Idea:

Solve the system:

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

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

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

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

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

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

$$\Rightarrow \quad \sum_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\Rightarrow \quad \sum_i y_i - n\beta_0 - \beta_1 \sum_i x_i = 0$$

$$\Rightarrow \quad \sum_i y_i - \beta_1 \sum_i x_i = n\beta_0$$

$$\Rightarrow \quad \beta_0 = \frac{\sum_i y_i}{n} - \beta_1 \frac{\sum_i x_i}{n}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \quad (\text{cond. } 1)$$

$$\Rightarrow \quad \boxed{\bar{y} = \beta_0 + \beta_1 \bar{x}}$$

$$\frac{\partial Q}{\partial \beta_1} = \frac{\partial}{\partial \beta_1} \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

$$= \sum_i \frac{\partial}{\partial \beta_1} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$= \sum_i 2(y_i - \beta_0 - \beta_1 x_i)(-x_i)$$

$$= -2 \sum_i (x_i y_i - \beta_0 x_i - \beta_1 x_i^2) \overset{!}{=} 0$$

$$\Rightarrow \sum_i x_i y_i - \beta_0 \underline{\sum_i x_i} - \underline{\beta_1 \sum_i x_i^2} = 0$$

(condition 2)

Plug (C1) into (C2)

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

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

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

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

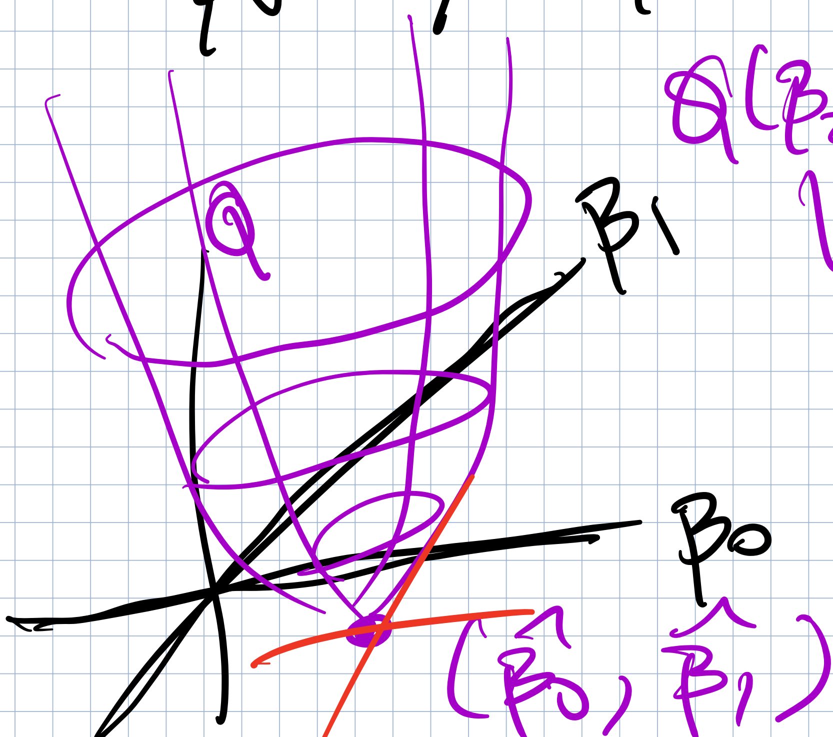
$$\Rightarrow \hat{\beta}_1 = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2}$$

We want to denote the specific minimizers of $Q(\beta_0, \beta_1)$

as:

$$\underline{\underline{(\hat{\beta}_0, \hat{\beta}_1)}} = \underset{\underline{\underline{\beta_0, \beta_1}}}{\operatorname{argmin}}\, Q(\beta_0, \beta_1)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$ (C1 w/ $\hat{\beta}_1$ plugged in)



$$Q(\beta_0, \beta_1) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

The solution to this problem is called the ordinary least squares ests.

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

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

HW:

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

$$= \frac{\frac{1}{n-1} \sum_i (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1} \sum_i (x_i - \bar{x})^2}$$

$$= \frac{\widehat{Cor}(X, Y)}{\widehat{Var}(X)}$$