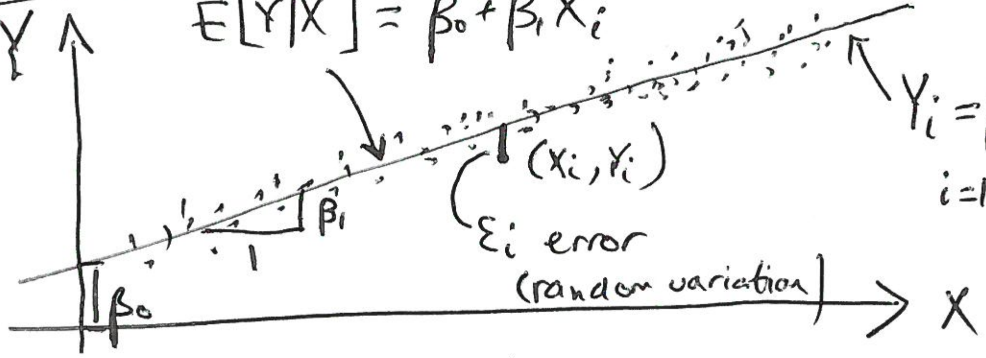
Simple Linear Regression models

Population

Mean model

$E[Y|X] = \beta_0 + \beta_1 X_i$

population size N

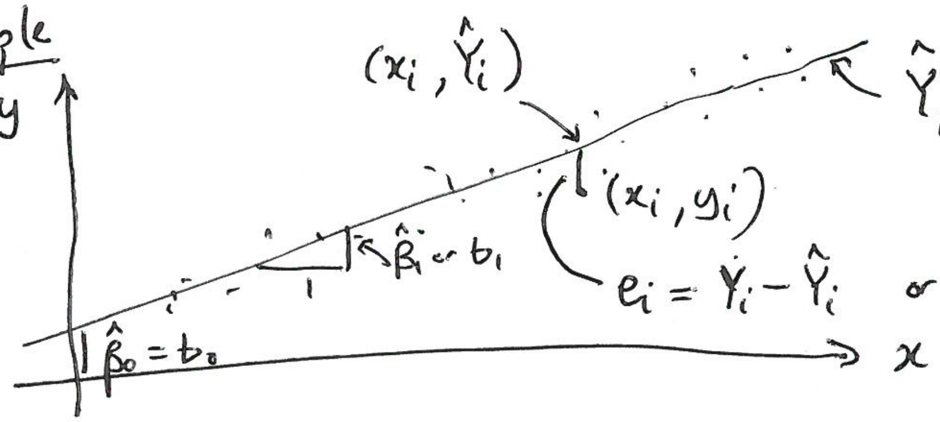


$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

$i = 1, 2, \ldots N$

"representative" sampling process

Sample



$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

$i = 1, 2, \ldots n$ sample size

$e_i = Y_i - \hat{Y}_i$ or $y_i - \hat{Y}_i$ residual

So, how do we estimate $\hat{\beta}_0 = b_0$ & $\hat{\beta}_1 = b_1$?

Gauss – method of least squares [see extract for STAT2001/6039 text]

find b_0 & b_1 that minimise the sum of squares of the errors

population $\sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$

or sample $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$

$\hat{\beta}_0 = b_0$ & $\hat{\beta}_1 = b_0$ are the estimates that minimise this!

To calculate b_0 & b_1 in practice we need means & variances of the x & y sample variables & we also need the covariance of X, Y:

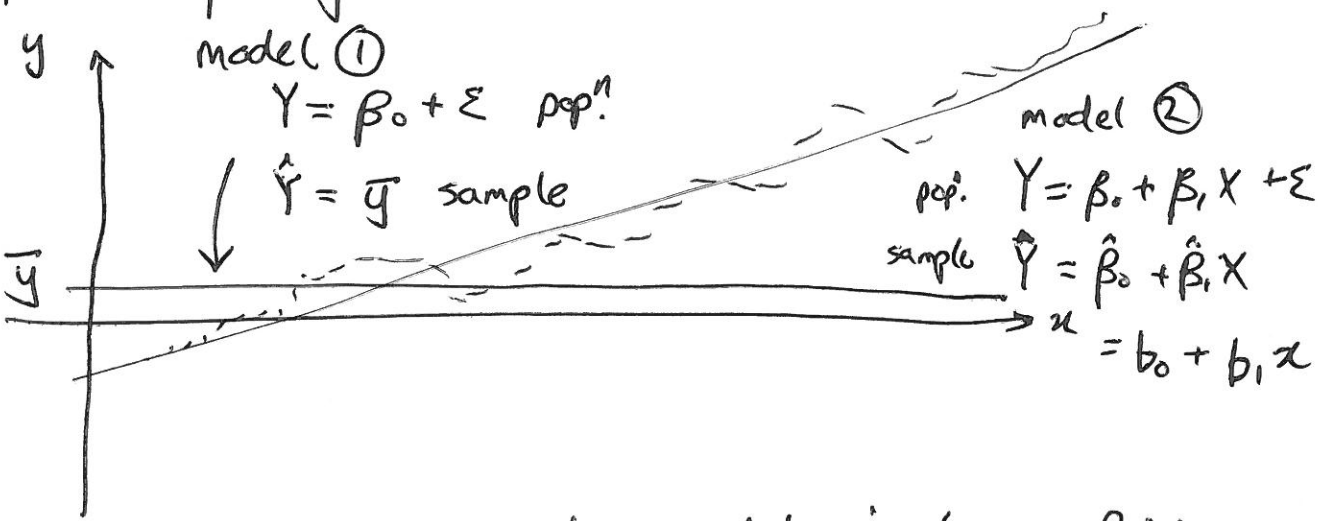
to estimate this in the sample we use

$$\frac{S_{xy}}{(n-1)} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

then $$\hat{\beta}_1 = b_1 = \frac{\frac{1}{n-1}\sum(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n-1}\sum(x_i-\bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = b_0 = \bar{y} - b_1\bar{x}$$

Two "competing" models



Difference between the two models is term $\beta_1 X$

Do we need this term?

→ if we don't then $\beta_1 = 0$ & we have model (1)

→ if we are convinced that a positive linear trend is a better fit then $\beta_1 > 0$ & we have model (2)

⇒ Hypothesis test of $H_0: \beta_1 = 0$ vs $H_A: \beta_1 > 0$

Assumptions underlying a simple linear regression (SLR) model

General assumptions (applicable to most statistical models)

(1) that the sample is representative of the population of interest

(2) that the explanatory (X) variables are measured without error (or at least minimal error cf Y) → all the error is in the Y direction (vertical on the earlier plots)

(3) that a model of the proposed form (eg a linear model) is appropriate

Model-specific assumptions (most regression-type models including SLR)

(population) $$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{deterministic model for the mean}} + \underbrace{\varepsilon_i}_{\text{stochastic model for the variance}} \quad i = 1, 2, \ldots N$$

deterministic model for the mean

$$E[Y_i | X] = \beta_0 + \beta_1 X_i$$

stochastic model for the variance

the assumptions, specific to this model, are about ε_i

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

Errors (ε_i) are independent & identically (Normally) distributed with mean 0 & constant variance σ^2

[This in a nutshell is the variance model]