Linear Regression

In engineering we often have more than one variable in an application and it is known that there exists some inherent relationship among the variables.

We will study the case with 2 variables.

<u>Example</u>: Variable 1: Distance to transmitter : X

Variable 2: Wireless signal strength : Y

Lets assume a linear relationship between Y and x is reasonable:

$$Y = \alpha + \beta x$$

Dependent variable — intercept — slope — Independent variable (Regressor)

<u>Note</u>: Don't confuse α, β with the type I & II error probabilities in hypothesis testing.

If the relationship between Y and X is exact, then it is a <u>deterministic</u> relationship.

In real applications, there are many sources of randomness:

- measurement noise
- the linear regression model might be an approximation to a much more complicated and possibly unknown relationship
- other factors (variables) not considered in model

Randomness means the same value of X does not always give the same value of Y (non-deterministic)

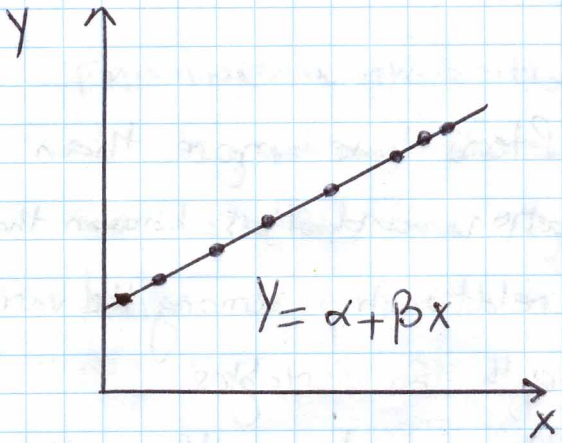
<u>Simple Linear Regression Model</u>:

$$Y = \alpha + \beta x + \varepsilon$$

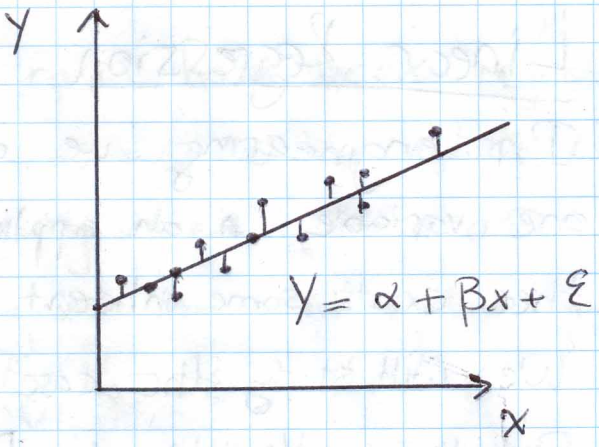
Dependent variable → Y; α: Unknown intercept; β: Unknown slope; x: Regressor; ε ← Random variable

$E[\varepsilon] = 0$

σ^2 variance



A deterministic relationship



A non-deterministic relationship

<u>Important</u>: ε is not a fixed number, it is a random variable which takes on a different value at each data point. The ε value at each point is shown as bars to the line $Y = \alpha + \beta x$ in the plot on the right above.

The only conditions on ε are $E[\varepsilon] = 0$
$Var[\varepsilon] = \sigma^2$

* $E[\varepsilon]$ implies that at a specific x, the Y values are distributed around the true (population) regression line.
* In practice α and β are unknown and must be estimated from data.

More advanced topics:

a) If there are more than one regressor variables

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

<u>Example</u>: Electromagnetic absorption in head due to cell phone use : Y
Signal frequency : x_1
Signal strength : x_2
Head size : x_3

b) If the relationship is nonlinear. There are many non-linear models, but a simple one would look like:

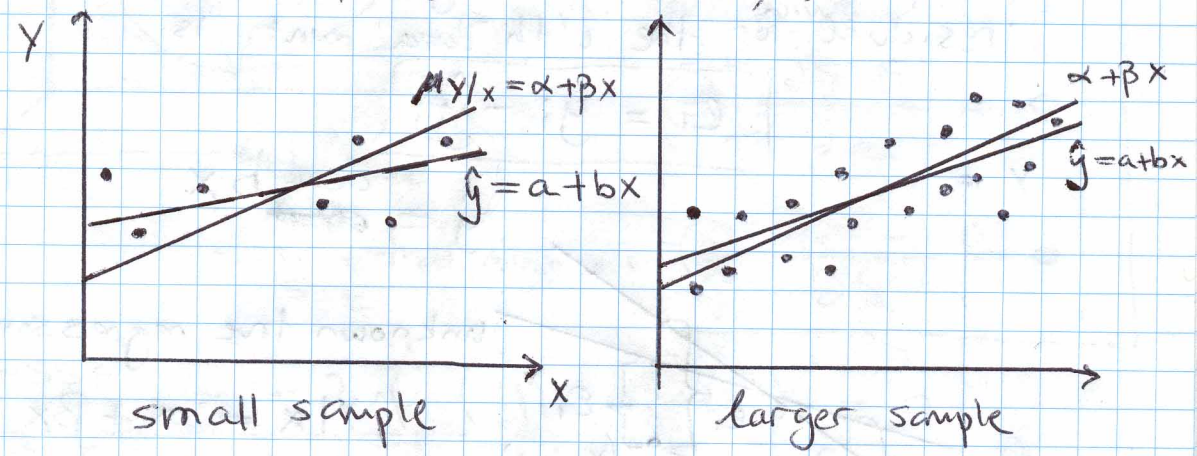
$$Y = \alpha + \beta x + \gamma x^2 + \varepsilon$$

This is a quadratic relationship.

Finding the parameters α and β from data

Example: In the example with Y signal strength and X distance to transmitter, we are given the pairs of observations (sample) (x_i, y_i) $i=1$ to n and we want to estimate α, β.

Let a be our estimate of the true population parameter α. Let b be our estimate of the true population parameter β. Just like the sample mean $\bar{x}$, a and b depend on the particular sample (they are random!) The larger the sample size the closer a should be to α and b to β (just like $\bar{x}$ to μ)



$\mu_{Y|X} = \alpha + \beta X$: expected value of Y given X (true regression) $Y = \alpha + \beta X$

$\hat{y} = a + bx$: fitted regression

↑ predicted (fitted) value

Conceptual model errors ε_i

$$Y = \alpha + \beta x + \varepsilon$$

$$E[Y] = E[\underbrace{\alpha + \beta x}_{\text{these are not random}} + \varepsilon]$$

$$= \alpha + \beta x + \cancel{E[\varepsilon]}$$

0 assumption about ε

$$\mu_{Y|x} = E[Y] = \alpha + \beta x$$

Let x_i be the data points for x

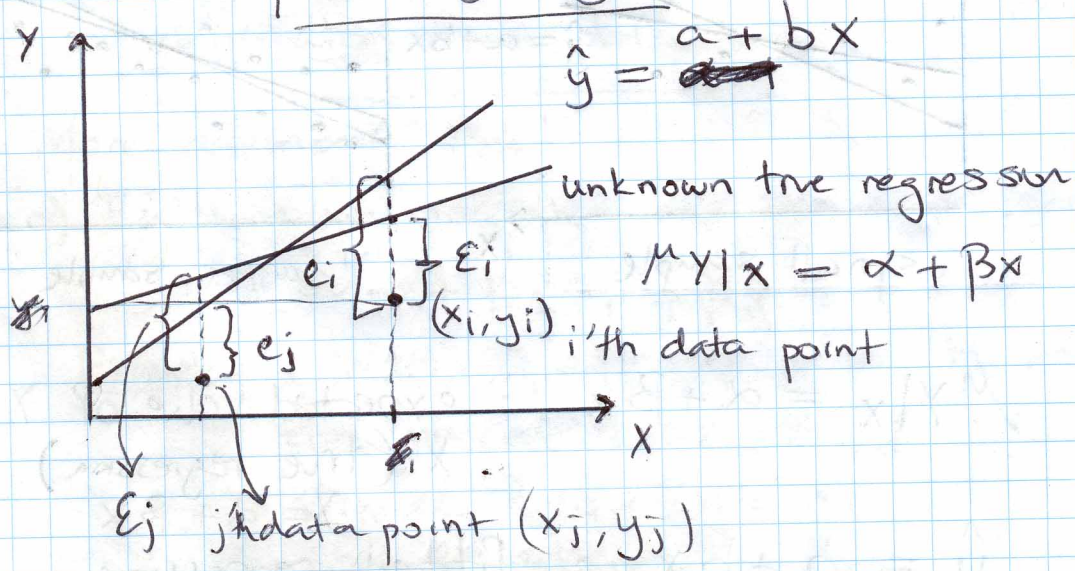
$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\varepsilon_i = Y_i - (\alpha + \beta x_i)$$ are the conceptual errors (realizations of the random var ε)

Residuals : error in fit

Given a set of data (sample) (x_i, y_i) $i = 1$ to n and a fitted model $\hat{y}_i = a + b x_i$ then the residual for the i'th data point is

$$\boxed{e_i = y_i - \hat{y}_i}$$



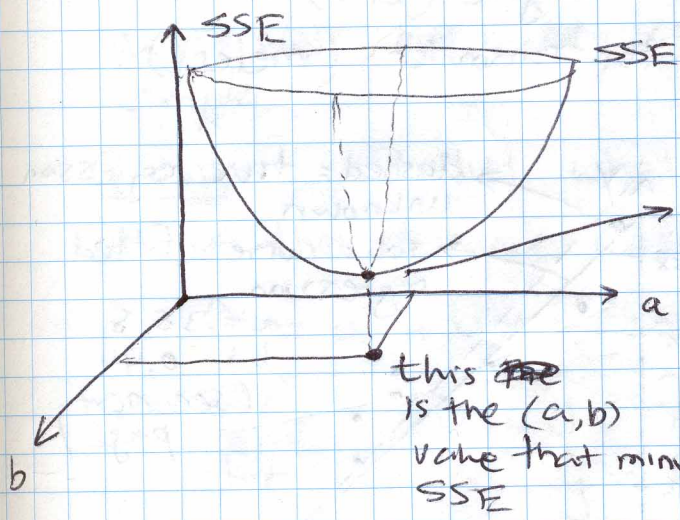
Least squares fitting method

One way to choose reasonable values for a and b is to choose them to minimize the sum of squared residuals.

$$SSE(\text{sum of squared errors}) = \sum_{i=1}^{n} e_i^2 \leftarrow$$ Note this is e_i not ε_i

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

To minimize SSE with respect to a and b, from calculus we know that the partial derivatives of SSE with respect to a and b must be 0.



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

at this point: *

$$\boxed{\begin{aligned} &① \ \frac{\partial SSE}{\partial a} = 0 \\ &② \ \frac{\partial SSE}{\partial b} = 0 \end{aligned}}$$

$$\frac{\partial SSE}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

$$\frac{\partial SSE}{\partial b} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) x_i = 0$$

Rearranging terms gives:

① $na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$

② $a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$

Solving ① and ② simultaneously gives the following formulas for a and b:

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

and $a = \bar{y} - b\bar{x}$

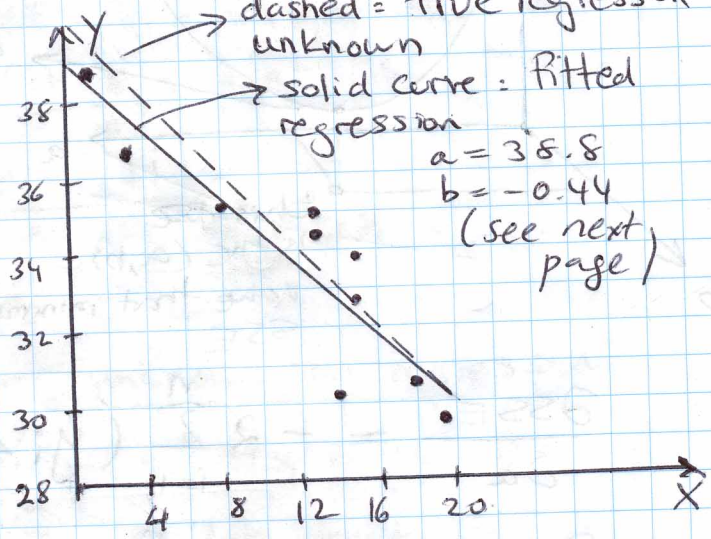
where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

<u>Example</u> Y : signal strength (dB)

X : distance to transmitter (meters)

10 measurements

i	x_i	y_i
1	13	34.4
2	1	38.4
3	17	30.4
4	19	29.7
5	14	30.1
6	15	33.9
7	15	32.8
8	8	35.2
9	13	34.9
10	3	36.8



Notice that for the same x value we have two different y values (non-deterministic) same for x = 13

Least squares fitting

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 11.8 \qquad \bar{y} = 33.66$$

$$b = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{-137.58}{315.6} = -0.4359 \approx -0.44$$

$$a = \bar{y} - b\bar{x} = 38.8 \Rightarrow \boxed{\hat{y} = 38.8 - 0.44x}$$

Normally, we wouldn't know the true values α, β, but in this case, I generated the data myself according to $Y = 40 - 0.5x + \varepsilon$ where ε was normally distributed with mean 0 and variance 1. So α was -0.5 and β was 40. Our estimates are quite close, but they could be better with a larger sample.

<u>Question</u>: Predict what the signal strength would be if the distance was 10 meters and 26 meters?

$$\hat{y} = a + bx = 38.8 - 0.44x$$

so for $x = 10$ $\quad \hat{y} = 34.4$ dB

for $x = 26$ $\quad \hat{y} = 27.36$ dB

<u>Note</u>: Sometimes we have control over for which x_i we make measurements. For instance, we could have measured signal strength y_i for x_i at regularly spaced intervals

2, 4, 6, 8, 10, 12, 14, 16, 18, 20 meters.