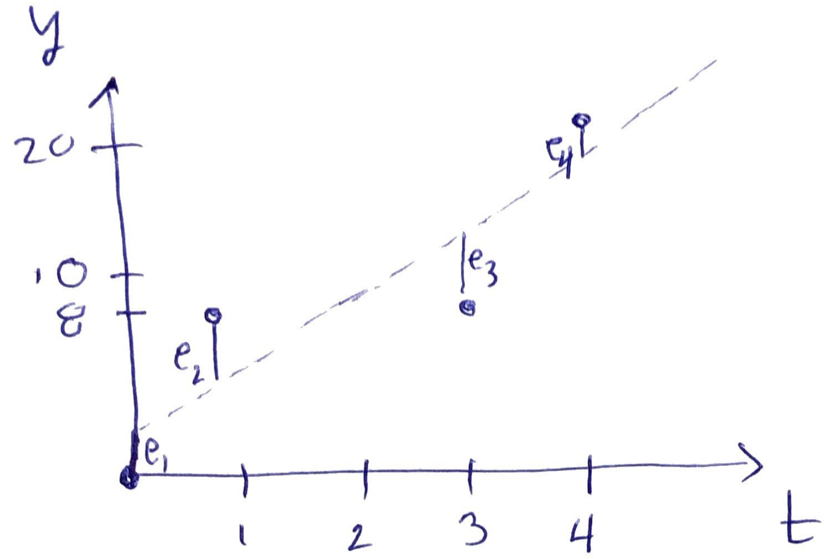
Data: $(0,0), (1,8), (3,8), (4,20)$

Section 12.1 In textbook



Line of best fit

$$y = C + Dt$$

is the one in which C and D are chosen such that

$$e_1^2 + e_2^2 + e_3^2 + e_4^2$$

is minimized.

We had
$$\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix} = \begin{bmatrix} 0-(C+D\cdot 0) \\ 8-(C+D\cdot 1) \\ 8-(C+3D) \\ 20-(C+4D) \end{bmatrix} = \begin{bmatrix} -C \\ 8-C-D \\ 8-C-3D \\ 20-C-4D \end{bmatrix}$$

So we can solve the problem by regarding ~~converting~~ it as a calculus problem:

$$f(C,D) = C^2 + (8-C-D)^2 + (8-C-3D)^2 + (20-C-4D)^2$$

We can minimize f by finding critical points, eg when $\nabla f = \begin{bmatrix} \frac{\partial f}{\partial C} \\ \frac{\partial f}{\partial D} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

Because the equat f is a sum of squares, it's concave up and so any critical point is a minimum.

A change of perspective

Suppose now that we assume that there is a 'ground truth' which is that there is a linear relationship between independent and dependent variables. This is the simplest mathematical relationship

$$Y = C + Dt + \text{error}$$

For example, assume that

+ 10 ppm CO_2

increases average global temp. by $0.1°C$.

Any deviations $\vec{e} = Y - (C + Dt)$ are due to random measurement noise.

Usually it is assumed that $e_1, \ldots, e_n$ are independent ~~normal~~ random variables, representing noise.

Then the obeservations are $Y_i = C + Dt_i + \varepsilon_i$

or

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\vec{Y} = A\vec{x} + \vec{\varepsilon} \quad (*)$$

$$= \vec{C} + D\vec{t} + \vec{\varepsilon}, \text{ where } \vec{C} = \begin{bmatrix} C \\ \vdots \\ C \end{bmatrix}, \ \vec{t} = \begin{bmatrix} t_1 \\ \vdots \\ t_n \end{bmatrix}$$

Usually, $\varepsilon_1, \dots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$.

This eqn (*) is called the <u>simple linear regression model</u>. The expected value of Y is a linear function of $\bar{x}$.

Eg if $\sigma^2 = 1$, what if we sampled a point at $t = 2$? What is ~~Pr(...)~~ prob that the measurement is ~~2 away from the ...~~ between 7 and 11?

$$P[7 \le Y_2 \le 11] = P[7 \le 1 + 4 \cdot 2 + \varepsilon_2 \le 11] = P[-2 \le \varepsilon_2 \le 2] \approx .98$$

<u>Example</u> Suppose we have a relationship between

x = applied stress (explanatory or predictor variable)

y = time-to-failure (response variable)

Suppose that the observed relationship is

$$y = 65 - 1.2x$$

with ~~[illegible]~~ deviations ~~having $N(0, \sigma^2)$ distribution~~ from this being random ~~wit~~ Gaussian with $\sigma = 8$.

Then for any fixed value x^* of stress, time to failure has normal distribution with mean $\mu = 65 - 1.2x^*$ and $\sigma = 8$

What is the mean time to failure when applied stress is $x = 20$.

What is the probability that the time to failure is ≥ 50 when the applied stress is $x = 20$?

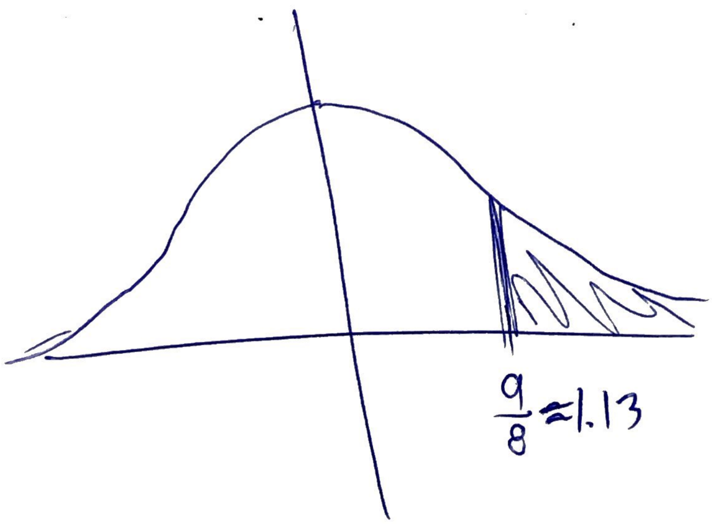
When $x = 2$ $Y \sim N(41, 8)$ so

$$\cancel{P[Y \geq 50 \mid x = 20]}$$

standardize $\frac{Y-\mu}{\sigma} \sim Z$

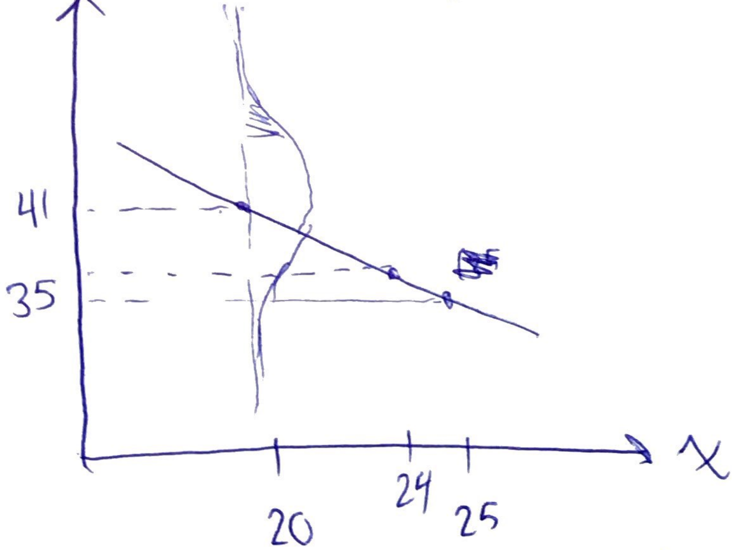
$$P[Y \geq 50] = P\left[\frac{Y-41}{8} \geq \frac{9}{8}\right]$$

$$= P\left[Z \geq \frac{9}{8}\right] = .13$$



$\frac{9}{8} \approx 1.13$

Section 12.1



~~What is~~ Let U and V be the time until failures when $x=25$ and $x=24$ ~~$x=20$ and $x=25$~~

$V \sim N(36.2, 8)$

$U \sim N(35, 8)$

~~What is $P[Y_1 \geq W]$?~~

What is $P[U \geq \cancel{W}]$?

$$P[\underbrace{U-\cancel{W}}_{} \geq 0] = P\left[\underbrace{\frac{(U-V)-(-1.2)}{\sqrt{128}}}_{Z} \geq \frac{-1.2}{\sqrt{128}}\right]$$

Normal with mean $36.2-35=-1.2$

$Var(U-V) = Var(U)+Var(V)$

$= 8^2+8^2 = 128$

So $\sigma = \sqrt{128} \approx 11.3$

$\approx P[Z \geq .11]$

$= .45$