CE 513: Statistical methods in civil engineering

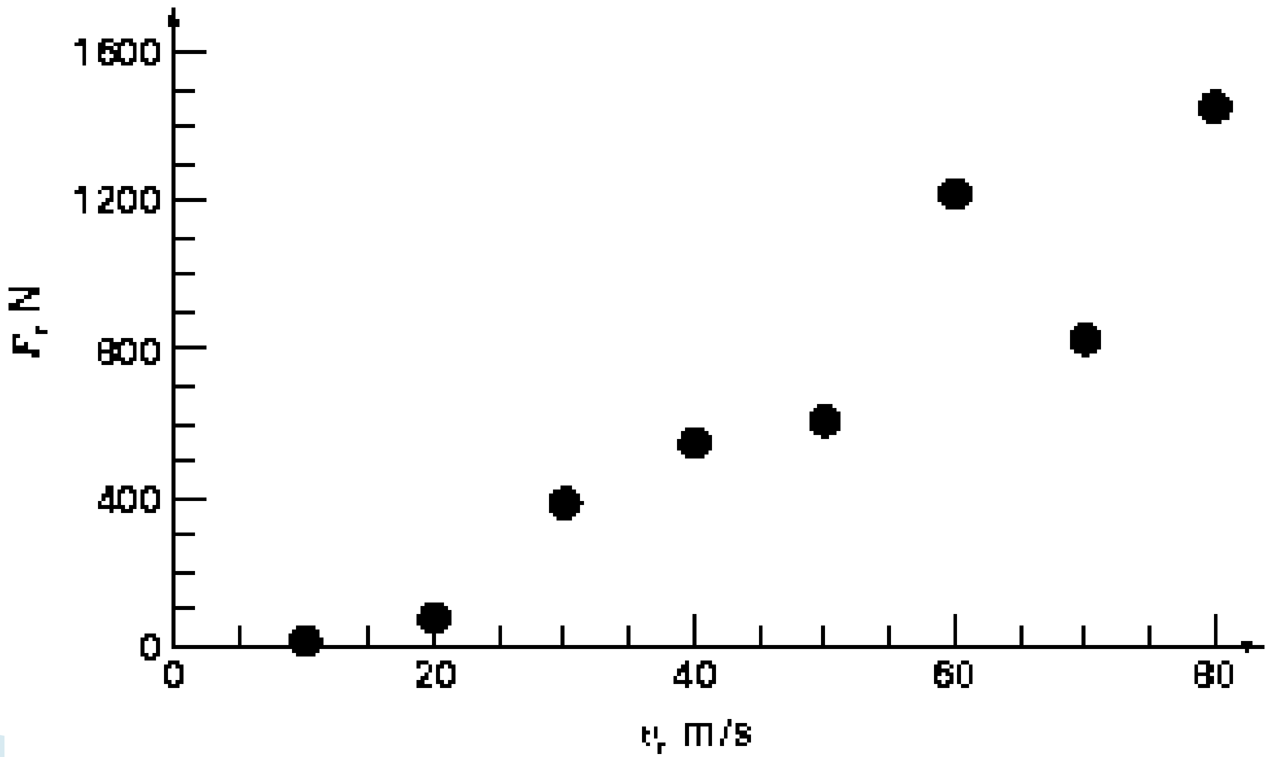
LECTURE : Regression

Dr. Budhaditya Hazra
Department of Civil engineering

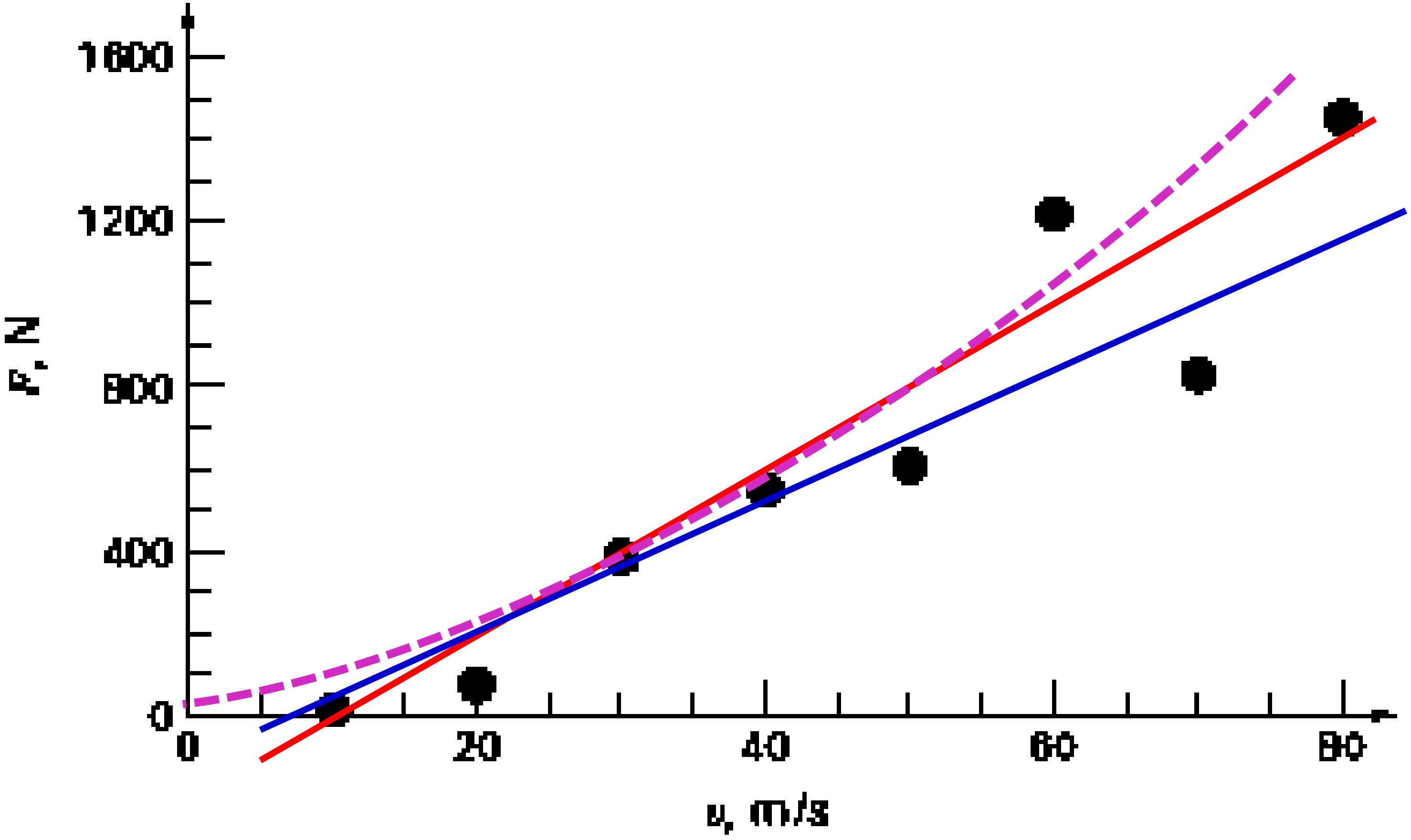
REGRESSION

TABLE Experimental data for force (N) and velocity (m/s) from a wind tunnel experiment.

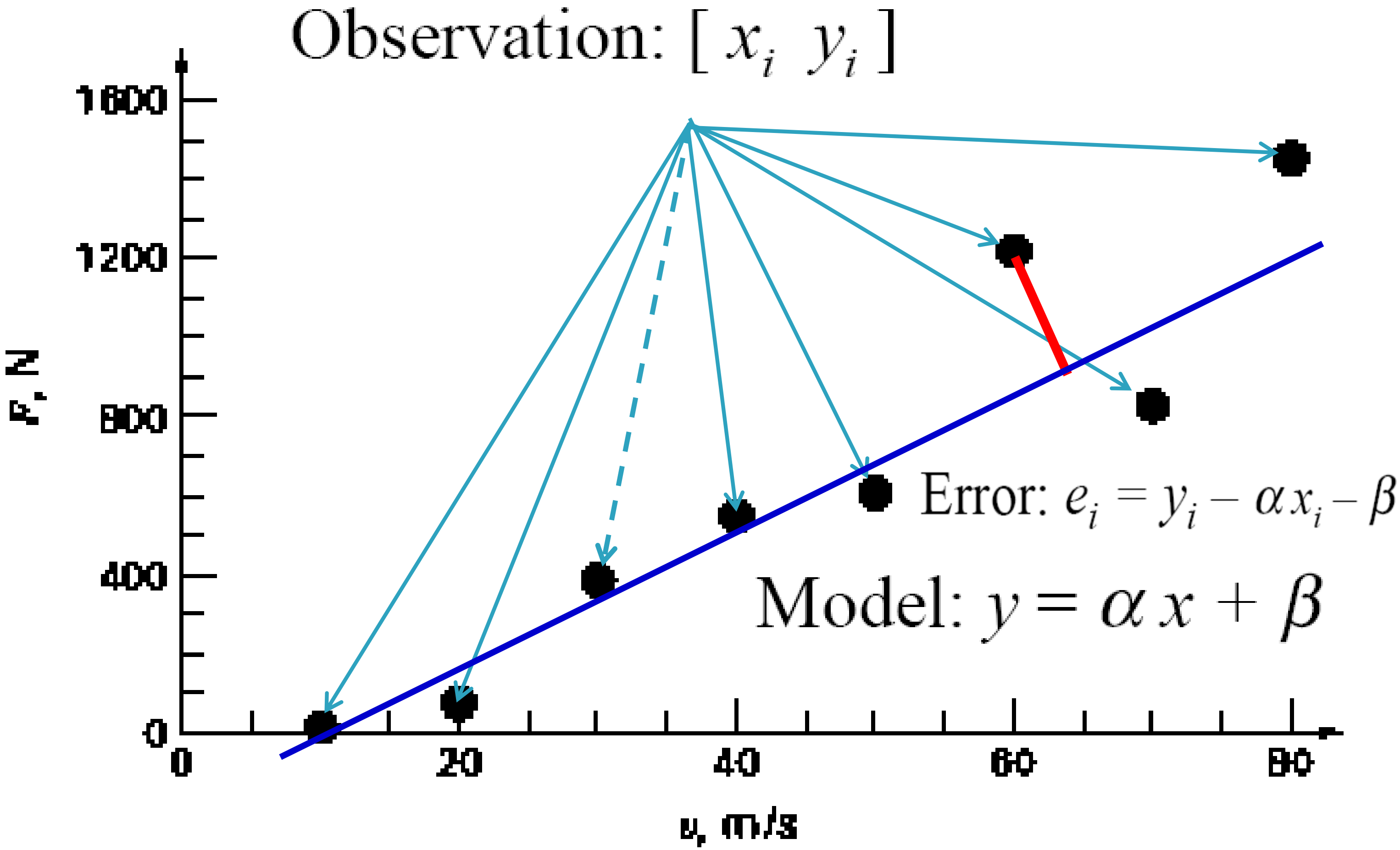
v, m/s	10	20	30	40	50	60	70	80
F, N	25	70	380	550	610	1220	830	1450



REGRESSION



LINEAR REGRESSION



CRITERIA FOR BEST FIT

Minimize sum of the square of the errors

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \beta - \alpha x_i\right)^2$$

Differentiate with respect to each coefficient:

$$\frac{\partial S_r}{\partial \beta} = -2 \sum (y_i - \beta - \alpha x_i)$$

$$\frac{\partial S_r}{\partial \alpha} = -2 \sum [(y_i - \beta - \alpha x_i) x_i]$$

BEST FIT LINE

$$\alpha = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2}$$

$$\beta = \overline{y} - \alpha\overline{x}$$

$$\hat{y} = \alpha x + \beta$$

MULTI LINEAR REGRESSION

2-variable case

$$y = c_0 + c_1 x_1 + c_2 x_2$$

Sum of squares of the residual: $S_r = \sum (y_i - c_0 - c_1 x_{1i} - c_2 x_{2i})^2$

Differentiate with respect to unknowns:

$$\frac{\partial S_r}{\partial c_0} = -2 \sum (y_i - c_0 - c_1 x_{1i} - c_2 x_{2i})$$

$$\frac{\partial S_r}{\partial c_1} = -2 \sum x_{1i} (y_i - c_0 - c_1 x_{1i} - c_2 x_{2i})$$

$$\frac{\partial S_r}{\partial c_2} = -2 \sum x_{2i} (y_i - c_0 - c_1 x_{1i} - c_2 x_{2i})$$

MULTI LINEAR REGRESSION

Setting the partial derivatives to 0

$$\begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 \end{bmatrix} \begin{Bmatrix} c_0 \\ c_1 \\ c_2 \end{Bmatrix} = \begin{Bmatrix} \sum y_i \\ \sum x_{1i}y_i \\ \sum x_{2i}y_i \end{Bmatrix}$$

Example

x_1	x_2	y
0	0	5
2	1	10
2.5	2	9
1	3	0
4	6	3
7	2	27

$$\begin{bmatrix} 6 & 16.5 & 14 \\ 16.5 & 76.25 & 48 \\ 14 & 48 & 54 \end{bmatrix} \begin{Bmatrix} c_0 \\ c_1 \\ c_2 \end{Bmatrix} = \begin{Bmatrix} 54 \\ 243.5 \\ 100 \end{Bmatrix}$$

$c_0 = 5$

$c_1 = 4$

$c_2 = -3$

GENERAL CASE

$$S(\mathbf{p}) = \sum_{k=1}^{m} \left[z^{(k)} - f(\mathbf{x}^{(k)}, \mathbf{p}) \right]^2; \quad \mathbf{p}^* = \operatorname{argmin} S(\mathbf{p})$$

$$\frac{\partial S}{\partial p_j} = 0; \quad j = 1 \ldots \nu$$

$$\sum_{k=1}^{m} \left\{ g_j(\mathbf{x}^{(k)}) \left[z^{(k)} - \sum_{i=1}^{\nu} p_i g_i(\mathbf{x}^{(k)}) \right] \right\} = 0; \quad j = 1 \ldots \nu$$

$$\mathrm{Q}\mathbf{p} = \mathbf{q}$$

$$Q_{ij} = \sum_{b-1}^{m} g_i(\mathbf{x}^{(k)}) g_j(\mathbf{x}^{(k)}); \quad q_j = \sum_{b-1}^{m} z^{(k)} g_j(\mathbf{x}^{(k)}); \quad i, j = 1 \ldots \nu$$

ERRORS

Define

$$S_{xy} = \sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i,$$

$$S_{xx} = \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2;$$

$$S_{yy} = \sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2$$

Sum of the square of the errors :

$$S_r = \frac{S_{xx}S_{yy} - S_{xy}^2}{S_{xx}}$$

Standard error of estimate :

$$S_{y/x} = \sqrt{\frac{S_r}{n-2}}$$

Standard deviation:

$$S_y = \sqrt{\frac{S_{yy}}{n-2}}$$

Coefficient of determination :

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

Example: error analysis of the linear fit

	x_i	y_i	x_i^2	$x_i y_i$	y_i^2
	1	0.5	1	0.5	0.25
	2	2.5	4	5.0	6.25
	3	2.0	9	6.0	4
	4	4.0	16	16.0	16
	5	3.5	25	17.5	12.25
	6	6.0	36	36.0	36
	7	5.5	49	38.5	30.25
Σ	**28**	**24**	**140**	**119.5**	**105**

$$S_{xx} = 140 - 28^2 / 7 = 28$$

$$S_{yy} = 105 - 24^2 / 7 = 22.7$$

$$S_{xy} = 119.5 - 28 \times 24 / 7 = 23.5$$

$$S_r = (28 \times 22.7 - 23.5^2) / 28$$

$$= 2.977$$

Since $s_{y/x} < s_y$, linear regression has merit.

$$s_y = \sqrt{\frac{22.7}{7-2}} = 2.131$$

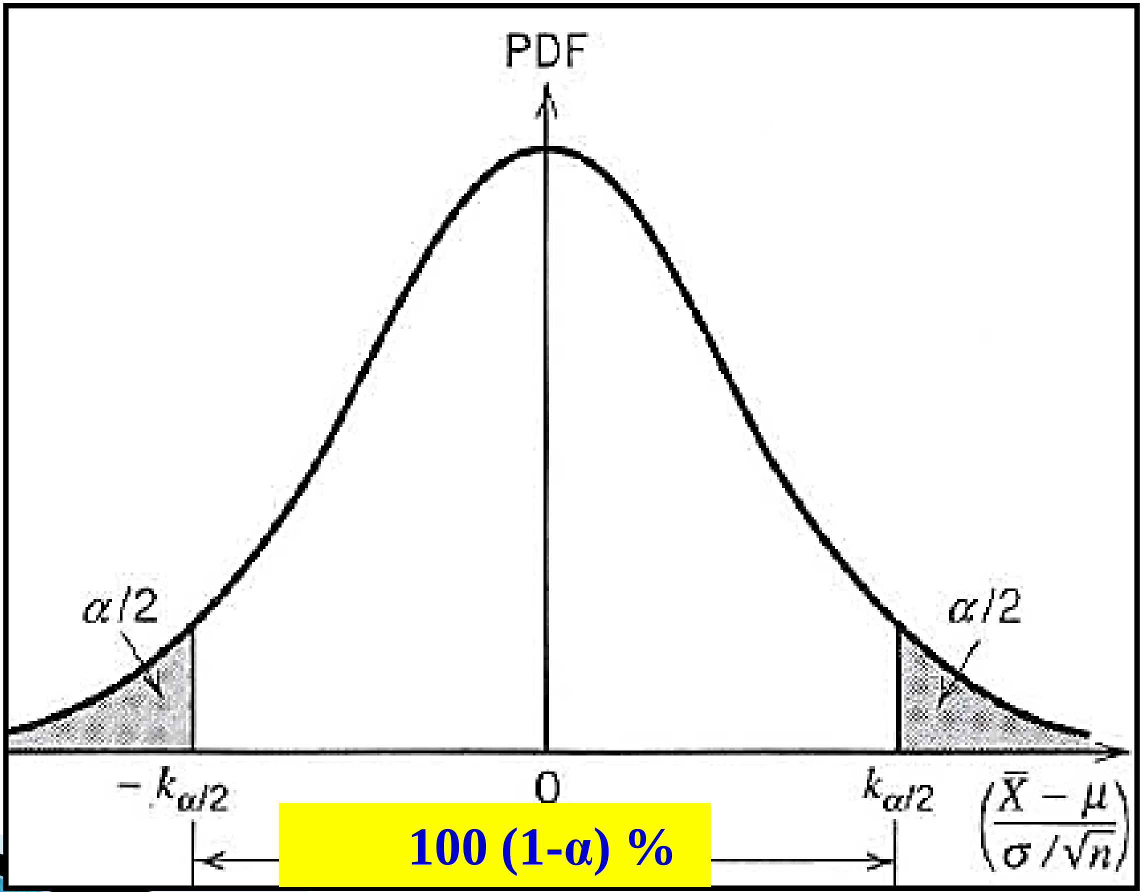
$$r^2 = \frac{23.5^2}{28 \times 22.7} = 0.869$$

$$s_{y/x} = \sqrt{\frac{2.977}{7-2}} = 0.772$$

Linear model explains 86.9% of original uncertainty.

CONFIDENCE INTERVAL

For mean μ with known variance



CONFIDENCE INTERVAL

$$P\left(-k_{\alpha/2} \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq k_{\alpha/2}\right) = 1-\alpha$$

$$<\mu>_{1-\alpha} = \left[\bar{x} - k_{\alpha/2}\frac{\sigma}{\sqrt{n}};\quad \bar{x} + k_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

CI- example

Consider the 41 observations of the Young's modulus given in Table (next slide). The sample mean and standard deviation are 29,576 ksi and 1,507 ksi, respectively.

Assume further that the Young's modulus is known to have a population standard deviation of 1,507 ksi.

Determine:

(a) the 95% confidence interval for the mean

(b) the 99% confidence interval for the mean.

CI- example

m	E (ksi)	m	E (ksi)
1	25,900	21	29,400
2	27,400	22	29,400
3	27,400	23	29,500
4	27,500	24	29,600
5	27,600	25	29,600
6	28,100	26	29,900
7	28,300	27	30,200
8	28,300	28	30,200
9	28,400	29	30,200
10	28,400	30	30,300
11	28,700	31	30,500
12	28,800	32	30,500
13	28,900	33	30,600
14	29,000	34	31,100
15	29,200	35	31,200
16	29,300	36	31,300
17	29,300	37	31,300
18	29,300	38	31,300
19	29,300	39	32,000
20	29,300	40	32,700
		41	33,400

CI- example

Step 1

$$1-\alpha = 0.95, \text{ or } \alpha = 1-0.95 = 0.05$$

$$\alpha/2 = 0.05/2 = 0.025, \text{ and } 1-\alpha/2 = 1-0.025 = 0.975.$$

Step 2

Using the standard normal table

$$k_{\alpha/2} = k_{0.025} = \Phi^{-1}(0.975) = 1.96.$$

Step 3

$$\frac{\sigma}{\sqrt{n}} k_{\alpha/2} = \frac{1,507}{\sqrt{41}} 1.96 = 461.$$

Thus, 95% CI is given by

$$<\mu>_{0.95} = (29,576 - 461;\ \ 29,576 + 461) = (29,115;\ \ 30,037) \text{ ksi.}$$

Similarly, 99% CI can be found out as

$$<\mu>_{0.99} = (29,576 - 607;\ \ 29,576 + 607) = (28,969;\ \ 30,183) \text{ ksi.}$$

For mean μ with unknown variance

$$f_T(t) = \frac{\Gamma[(f+1)/2]}{\sqrt{\pi f}\,\Gamma(f/2)}\left(1+\frac{t^2}{f}\right)^{-(f+1)/2}, \quad -\infty < t < \infty$$

$$P\left(-t_{\alpha/2,n-1} \le \frac{\overline{X}-\mu}{S/\sqrt{n}} \le t_{\alpha/2,n-1}\right) = 1-\alpha.$$

$$<\mu>_{1-\alpha} = \left[\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}};\quad \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right]$$

Consider again the previous example

Assume that the variance is unknown

Since n = 41, use student-t distribution with (n-1) = 40 degrees of freedom

$$t_{\alpha/2,n-1} = t_{0.025,40} = t_{0.975,40} = 2.021$$

$$< \mu >_{0.95} = \left[29,576 - 2.021 \frac{1,507}{\sqrt{41}};\ 29,576 + 2.021 \frac{1,507}{\sqrt{41}} \right]$$

$$= [29,100;\ 30,052] \text{ ksi.}$$

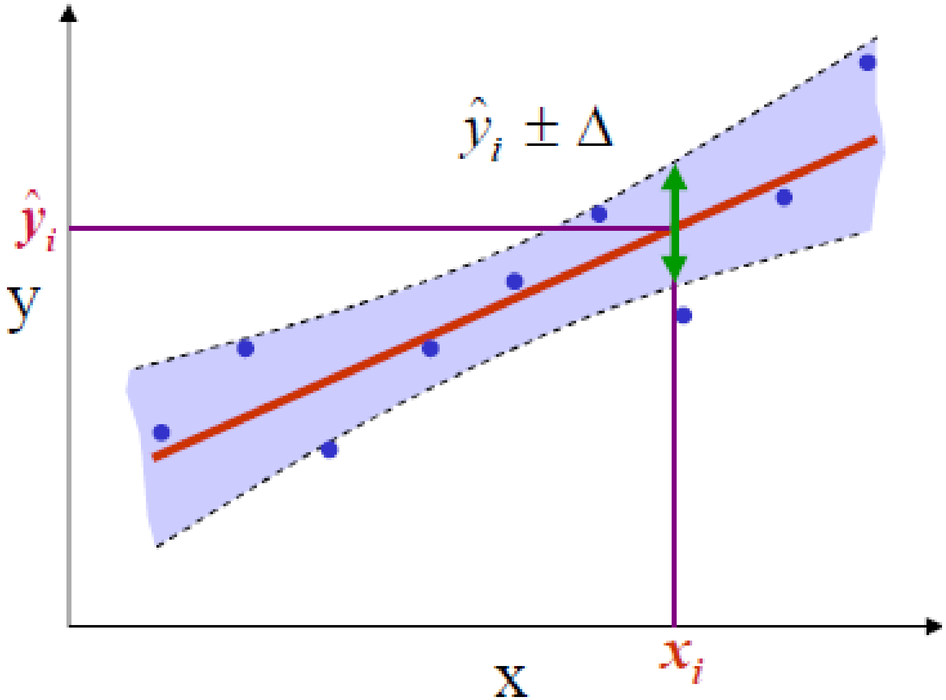
Consider again the previous example with a minor twist

Assume that the variance is unknown and only 10 samples of data are available

$$< \mu >_{0.95} = \left[29,576 - 2.262 \frac{1,507}{\sqrt{10}};\ \ 29,576 + 2.262 \frac{1,507}{\sqrt{10}} \right] \text{ksi}$$

$$= [28,498;\ \ 30,654] \text{ ksi.}$$

Linear Regression and CI



For CI 95%, you can be 95% confident that the two curved confidence bands enclose the true best-fit linear regression line, leaving a 5% chance that the true line is outside those boundaries.

A 100 (1 - α) % confidence interval for y_i is given by

Confidence interval 95% $\rightarrow$ $\alpha = 0.05$

$$\hat{y}_i \pm t_{\alpha/2}\, s_{y/x} \sqrt{\frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{xx}}}$$

Example: to estimate ***y*** when ***x*** is 3.4 using 95% confidence interval:

x_i	y_i
1	0.5
2	2.5
3	2.0
4	4.0
5	3.5
6	6.0
7	5.5

$$\hat{y} = \alpha x + \beta = 0.8363(3.4) + 0.0714 = 2.9148$$

$$95\%\text{ Confidence} \rightarrow \alpha = 0.05 \rightarrow t_{\alpha/2} = t_{0.025}(\text{df} = \text{n-2} = 5) = 2.571$$

Interval: $2.9148 \pm (2.571)(0.772)\sqrt{\frac{1}{7} + \frac{(3.4-4)^2}{28}}$

$$2.9148 \pm 0.7832$$

MATLAB functions

Polynomial fitting:

Second-order polynomial:

$$y = a_0 + a_1 x + a_2 x^2$$

Sum of the squares of the residuals:

$$S_r = \sum (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2$$

Fit a second-order polynomial to the data

	x_i	y_i
	0	2.1
	1	7.7
	2	13.6
	3	27.2
	4	40.9
	5	61.1
Σ	**15**	**152.6**

Solving by MATLAB `polyfit` Function

```
>> x = [0 1 2 3 4 5];
>> y = [2.1 7.7 13.6 27.2 40.9 61.1];
>> c = polyfit(x, y, 2)
>> [c, s] = polyfit(x, y, 2)
>> st = sum((y - mean(y)).^2)
>> sr = sum((y - polyval(c, x)).^2)
>> r = sqrt((st - sr) / st)
```

Evaluate polynomial at the points defined by the input vector

```
>> y = polyval(c, x)
```

where x = Input vector

y = Value of polynomial evaluated at x

c = vector of coefficient in descending order

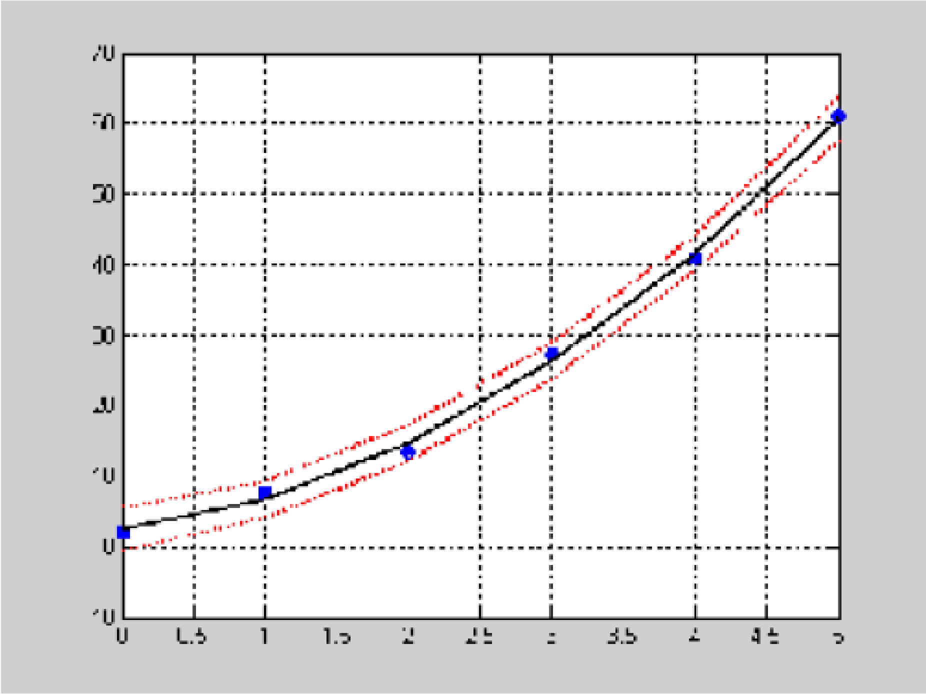
$$Y = c(1)*x^{n} + c(2)*x^{(n-1)} + \ldots + c(n)*x + c(n+1)$$

Example: $y = 1.86071x^2 + 2.35929x + 2.47857$

```
>> c = [1.86071 2.35929 2.47857]
```

Errors

By passing an optional second output parameter from **polyfit** as an input to **polyval**.



```
>> [c,s] = polyfit(x,y,2)

>> [y2,delta] = polyval(c,x,s)

>> plot(x,y,'o',x,y2,'g-',x,y2+2*delta,'r:',x,y2-2*delta,'r:')
```

Interval of $\pm 2\Delta$ = 95% confidence interval