ME 209
Numerical Methods

7. Regression I

Linear Regression

Asst. Prof. Dr. Nurettin Furkan DOĞAN

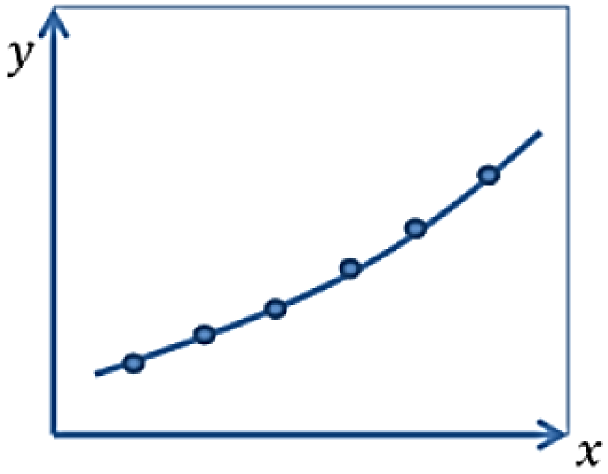
Mechanical Engineering Department

Gaziantep University

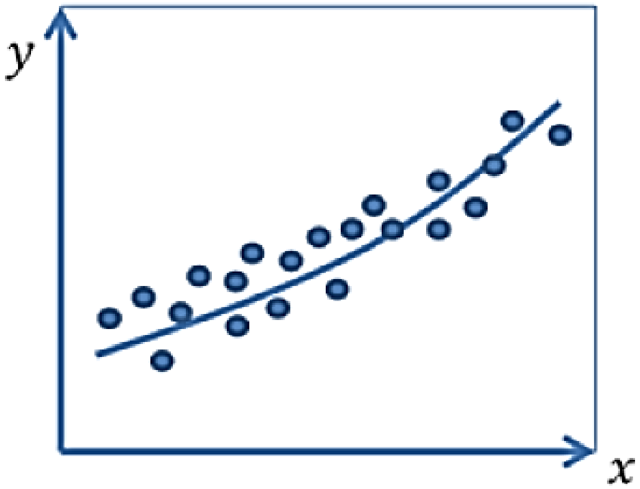
2024

WHAT IS REGRESSION?

- Rather than finding an equation that fits all the points, regression analysis in statistics refers to *identifying an equation that represents the trend of the points*.



Interpolation



Regression

- For example, if 100 points are found as a result of an experiment and all of them are written in the Lagrangian polynomial, a 99th degree polynomial is obtained, which is time-consuming and not useful to do mathematically.
- More importantly, incorrect values that occur due to measurement during experiments and observations will be added to the polynomial.
- Instead, it would be easier and more realistic to express these points with a continuous function passing as close as possible to the given points.
- The process of finding such a function is called **curve fitting (regression)**.

- Given n data points $(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)$ best fit $y = f(x)$ to the data.

- Residual (error) in each point is

$$E_i = y_i - f(x_i)$$

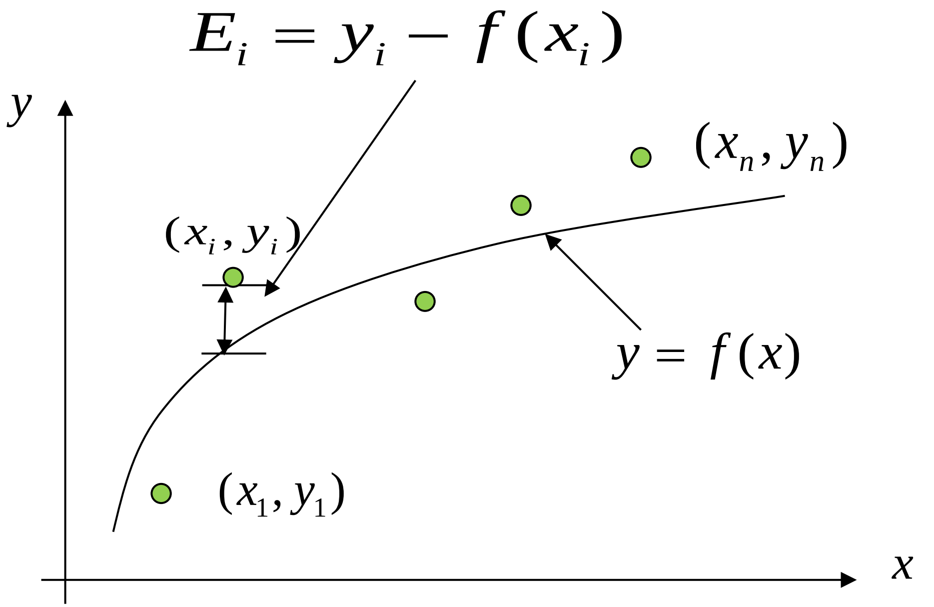
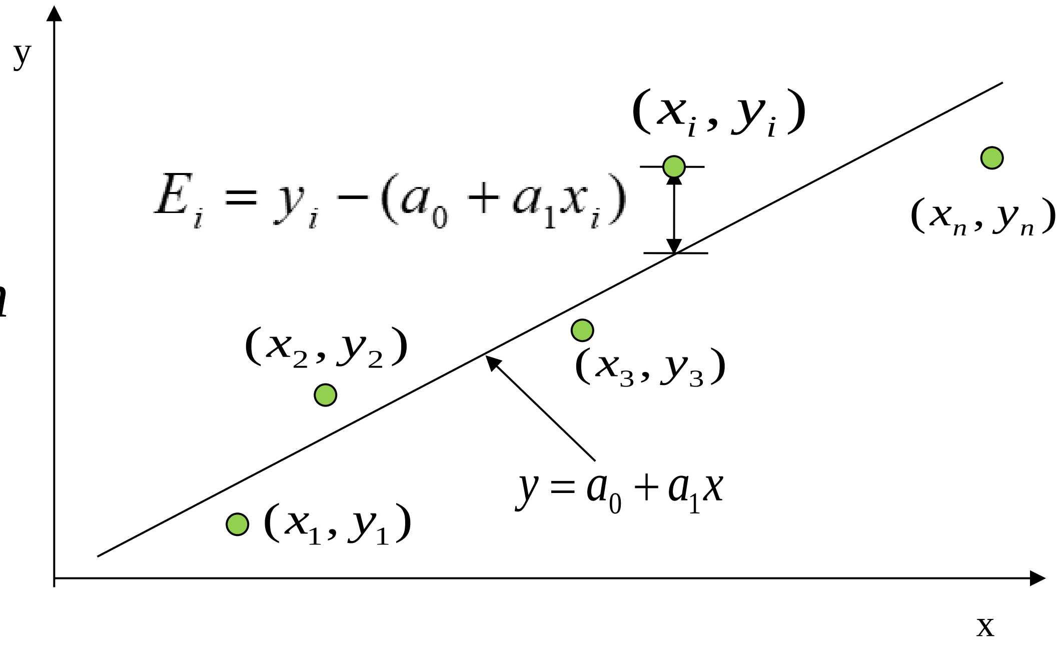


Figure. Basic model for regression

7.1 Linear Regression



Linear regression is the most popular regression model. In this model, we wish to predict response to n data points $(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)$ by a regression model given y

$$y = a_0 + a_1 x \tag{1}$$

where a_0 and a_1 are the constants of the regression model.

A measure of goodness of fit, which indicates how well the equation $a_0 + a_1 x$ predicts the response variable y, is based on the magnitude of the residuals ε_i at each of the n data points.

$$E_i = y_i - (a_0 + a_1 x_i) \tag{2}$$

7.1.1 Least Squares Methods

- Ideally, if all the residuals ε_i are zero, an equation may have found in which all the points lie on the model.
- Thus, minimization of the residual is an objective of obtaining regression coefficients.
- The most popular method to minimize the residual is the ***least squares methods***, where the estimates of the constants in the model are chosen such that the sum of the squared residuals is minimized, that is minimize

$$\sum_{i=1}^{n} E_i^{\,2}$$

- The least squares criterion minimizes the sum of the square of the residuals in the model, and also produces a unique line.
- Sum of the squares of residual:

$$S_r = \sum_{i=1}^{n} E_i^{\,2} = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right)^2$$

- The least squares criterion minimizes the sum of the square of the residuals in the model, and also produces a unique line.
- Sum of the square of the residuals:

$$S_r = \sum_{i=1}^{n} E_i^{\,2} = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right)^2$$

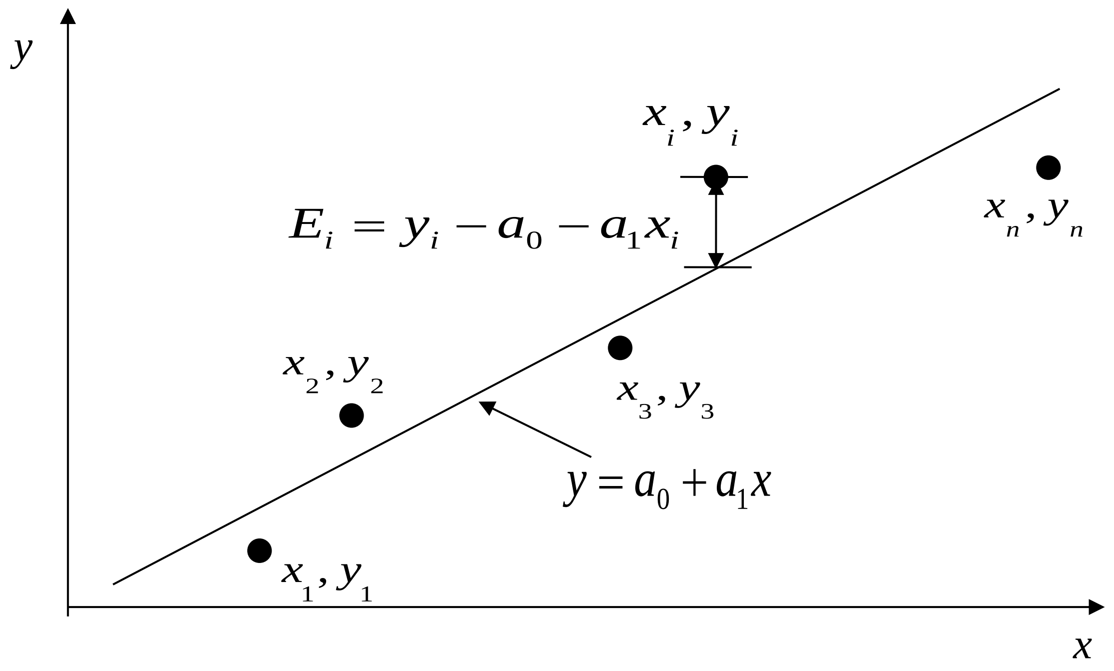
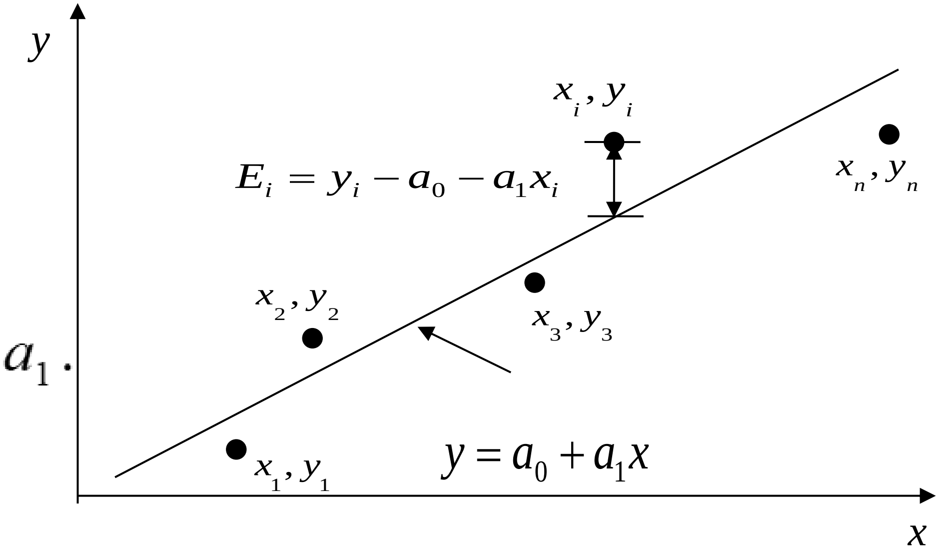


Figure. Linear regression of y vs x data showing residuals at a typical point, x_i .

$$S_r = \sum_{i=1}^{n} E_i^{\,2} = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right)^2$$



To find a_0 and a_1, we minimize S_r with respect to a_0 and a_1.

$$\frac{\partial S_r}{\partial a_0} = 2\sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right)\left(-1\right) = 0$$

$$\frac{\partial S_r}{\partial a_1} = 2\sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right)\left(-x_i\right) = 0$$

giving

$$-\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} a_0 + \sum_{i=1}^{n} a_1 x_i = 0$$

$$-\sum_{i=1}^{n} y_i x_i + \sum_{i=1}^{n} a_0 x_i + \sum_{i=1}^{n} a_1 x_i^2 = 0$$

Noting that $\sum_{i=1}^{n} a_0 = a_0 + a_0 + . . . + a_0 = na_0$

$$na_0 + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$na_0 + a_1\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$a_0\sum_{i=1}^{n} x_i + a_1\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Solving the equations above for a_0 and a_1:

$$\boxed{a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}}$$

$$\boxed{a_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}}$$

$$\boxed{a_0 = \bar{y} - a_1\bar{x}}$$

where $\bar{x}$ and $\bar{y}$ are the average values:

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

Example: The torque, T needed to turn the torsion spring of a mousetrap through an angle, is given below. Find the constants for the linear model given by $T = k_1 + k_2\theta$

Table. Torque vs Angle for a torsional spring

Angle, θ	**Torque, T**
Radians	*N.m*
0.698132	0.188224
0.959931	0.209138
1.134464	0.230052
1.570796	0.250965
1.919862	0.313707

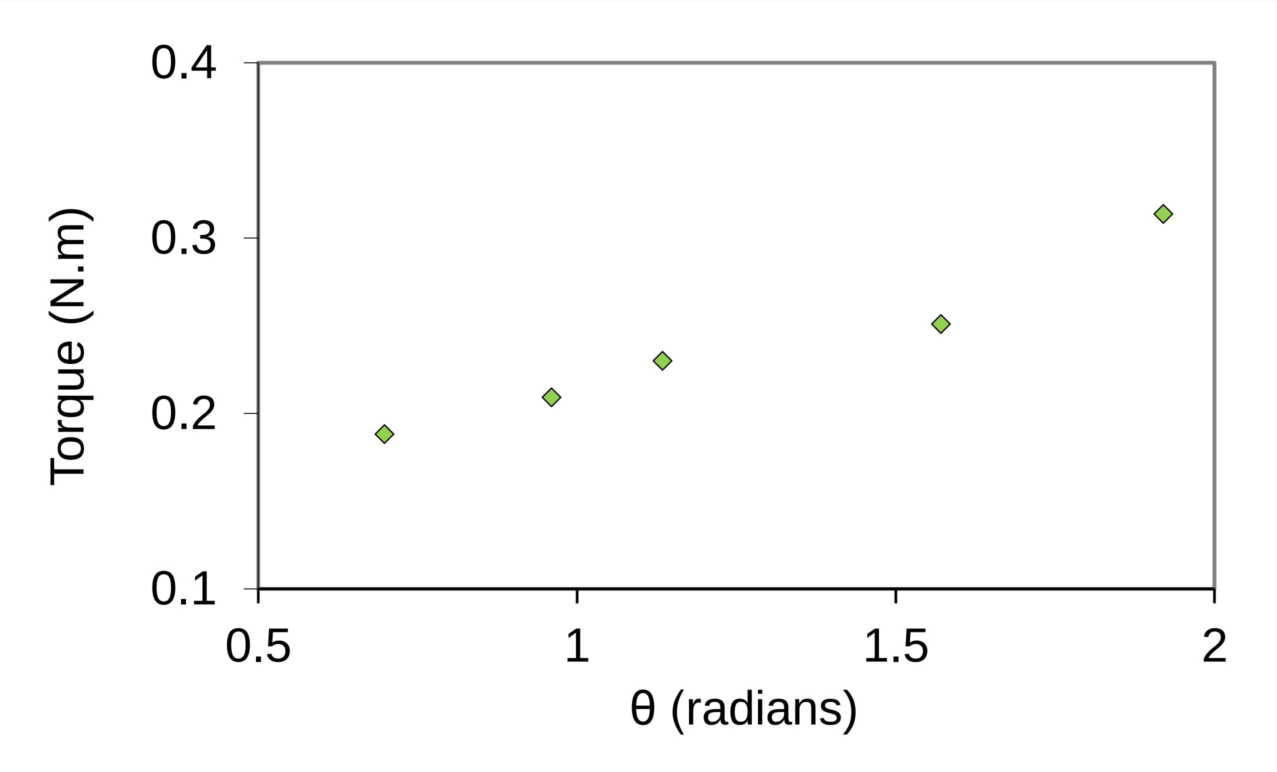


Figure. Data points for Torque vs Angle data

Table. Tabulation of data for calculation of important summations

θ	T	θ^2	$T\theta$
Radians	*N.m*	*Radians*^{2}	*N.m-Radians*
0.698132	0.188224	0.487388	0.131405
0.959931	0.209138	0.921468	0.200758
1.134464	0.230052	1.2870	0.260986
1.570796	0.250965	2.4674	0.394215
1.919862	0.313707	3.6859	0.602274
$\sum_{i=1}^{5} =$ 6.2831	1.1921	8.8491	1.5896

Using equations described for a_0 and a_1 with $n = 5$

$$k_2 = \frac{n\sum_{i=1}^{5}\theta_i T_i - \sum_{i=1}^{5}\theta_i \sum_{i=1}^{5} T_i}{n\sum_{i=1}^{5}\theta_i^2 - \left(\sum_{i=1}^{5}\theta_i\right)^2}$$

$$= \frac{5(1.5896) - (6.2831)(1.1921)}{5(8.8491) - (6.2831)^2}$$

$$= 9.6091 \times 10^{-2} \ \textit{N.m/rad}$$

Table. Tabulation of data for calculation of important summations

θ	T	θ^2	$T\theta$
Radians	*N.m*	*Radians*^{2}	*N.m-Radians*
0.698132	0.188224	0.487388	0.131405
0.959931	0.209138	0.921468	0.200758
1.134464	0.230052	1.2870	0.260986
1.570796	0.250965	2.4674	0.394215
1.919862	0.313707	3.6859	0.602274
$\sum_{i=1}^{5} =$ 6.2831	1.1921	8.8491	1.5896

Using the average torque and average angle to calculate k_1

$$\bar{T} = \frac{\sum_{i=1}^{5} T_i}{n} = \frac{1.1921}{5} = 2.3842 \times 10^{-1}$$

$$\bar{\theta} = \frac{\sum_{i=1}^{5} \theta_i}{n} = \frac{6.2831}{5} = 1.2566$$

Using, $k_1 = \bar{T} - k_2\,\bar{\theta}$

$$= 2.3842 \times 10^{-1} - (9.6091 \times 10^{-2})(1.2566)$$

$$= 1.1767 \times 10^{-1} \quad N.m$$

Using linear regression, a trend line is found from the data

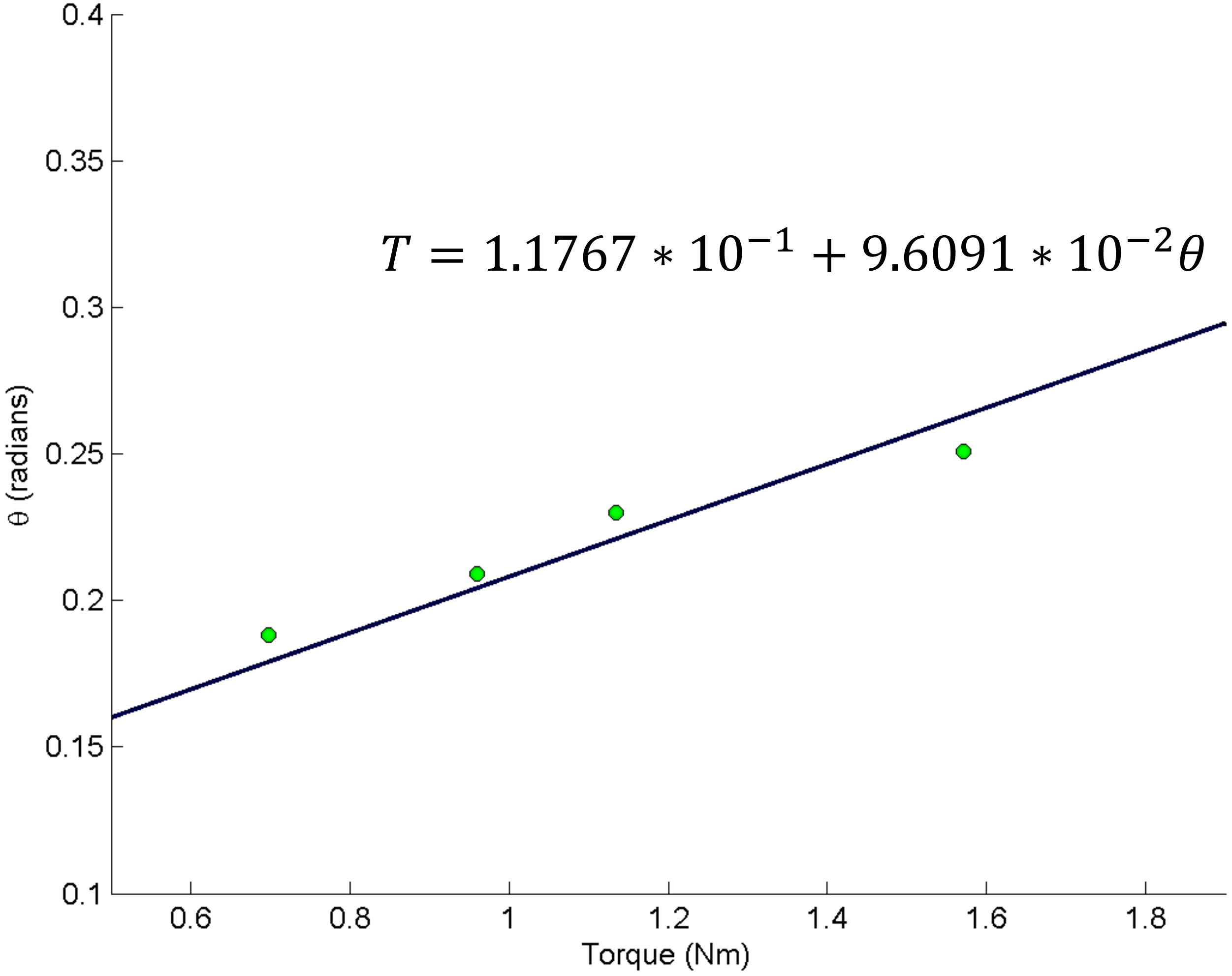


Figure. Linear regression of Torque versus Angle data

Example: To find the longitudinal modulus of composite, the following data is collected. Find the longitudinal modulus, E using the regression model $\sigma = E\varepsilon$ and the sum of the square of the residuals.

Table. Stress vs. Strain data

Strain	Stress
(%)	*(MPa)*
0	0
0.183	306
0.36	612
0.5324	917
0.702	1223
0.867	1529
1.0244	1835
1.1774	2140
1.329	2446
1.479	2752
1.5	2767
1.56	2896

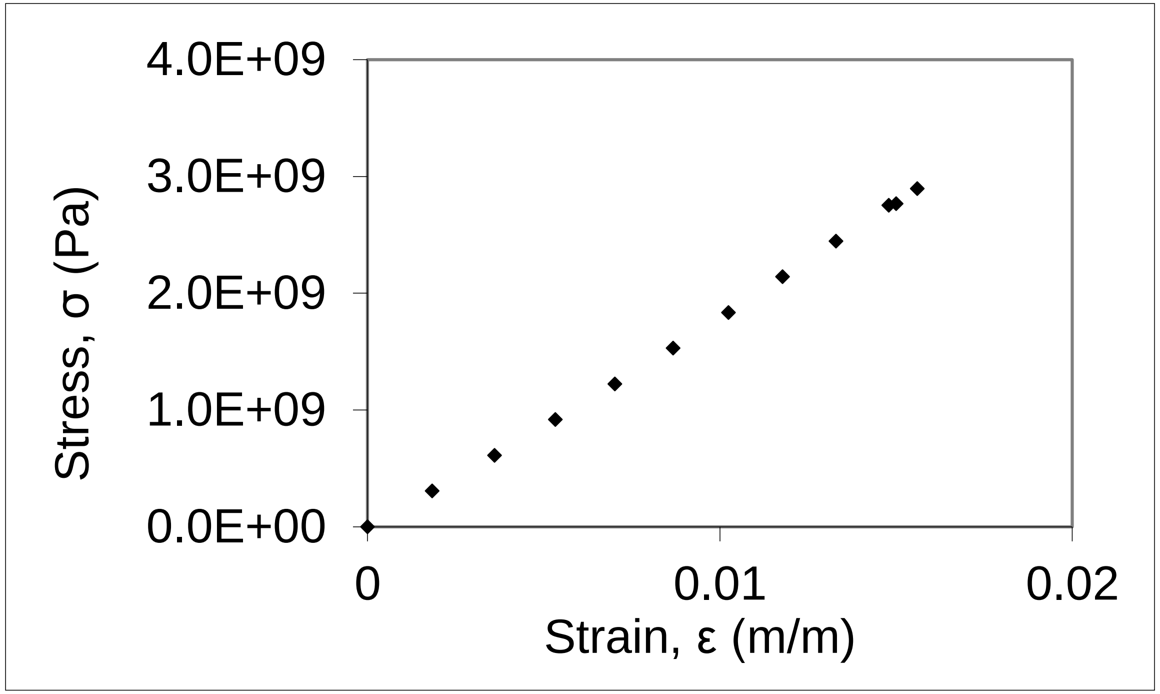


Figure. Data points for Stress vs. Strain data

Table. Summation data for regression model

i	ε	σ	ε ²	εσ
1	0.00	0.00E+00	0.00E+00	0.00E+00
2	0.00	3.06E+08	3.35E-06	5.60E+05
3	0.00	6.12E+08	1.30E-05	2.20E+06
4	0.01	9.17E+08	2.83E-05	4.88E+06
5	0.01	1.22E+09	4.93E-05	8.59E+06
6	0.01	1.53E+09	7.52E-05	1.33E+07
7	0.01	1.84E+09	1.05E-04	1.88E+07
8	0.01	2.14E+09	1.39E-04	2.52E+07
9	0.01	2.45E+09	1.77E-04	3.25E+07
10	0.01	2.75E+09	2.19E-04	4.07E+07
11	0.02	2.77E+09	2.25E-04	4.15E+07
12	0.02	2.90E+09	2.43E-04	4.52E+07
	1.0714E-01	1.9423E+10	1.2764E-03	2.3337E+08

$$E = \frac{\sum_{i=1}^{n} \sigma_i \varepsilon_i - \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \sum_{i=1}^{n} \sigma_i}{\sum_{i=1}^{n} {\varepsilon_i}^2 - \frac{1}{n} \left(\sum_{i=1}^{n} \varepsilon_i\right)^2}$$

$$= 1.8749E + 11\ Pa$$

$$= 187.49\ GPa$$

The equation $\sigma = 187.49\varepsilon$ describes the data.

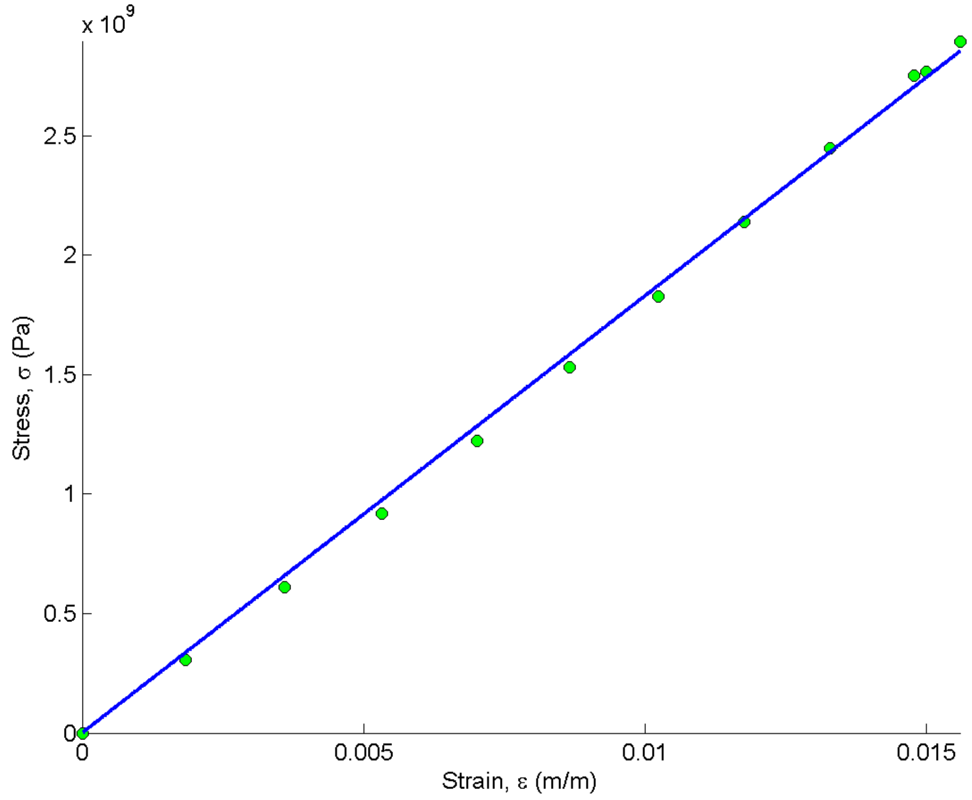


Figure. Linear regression for stress vs. strain data

NEXT LECTURE
Nonlinear Regression