6. Least squares regression

- curve fitting and statistics
- straight line fit to data
- linearization of nonlinear equations
- fitting a polynomial to data
- multiple linear regression
- general linear least squares

Curve fitting: motivation

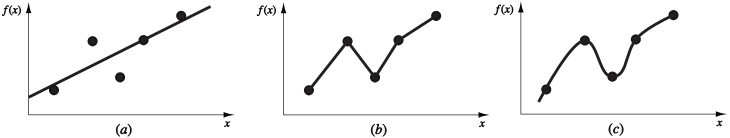
- data are often available only at discrete points along a continuum
- we may need estimates at points between known values
- we can use simple function to approximate complicated data
- this is called **curve fitting**

Regression

- data contain significant error or noise
- derive curve representing general trend
- curve does not necessarily pass through all points
- example: least squares regression

Interpolation

- data are very accurate
- fit a curve (or piecewise curves) exactly
- estimate values between points
- example: interpolation



(a) (b) (c)

Engineering practice and curve fitting

- common engineering need: estimating intermediate values
- many properties are not tabulated $\longrightarrow$ must fit your own data
- two main applications: *trend analysis* and *hypothesis testing*

Trend analysis

- use data patterns for prediction
  - interpolation: within the range of available data
  - extrapolation: outside the available range
- applications appear in all fields of engineering

Hypothesis testing

- compare existing mathematical model with observed data
- two cases:
  1. model coefficients unknown $\rightarrow$ determine best-fit values
  2. model coefficients known $\rightarrow$ check adequacy of predictions
- multiple models may be tested, best selected empirically

Other uses of curve fitting

- derive simpler functions to approximate complicated ones
- essential tool in numerical methods:
  - numerical integration
  - solution of differential equations
- provides efficiency and insight into underlying physical systems

Statistics for experimental data

- engineering measurements often provide limited raw information
- example:

  24 readings of coefficient of thermal expansion of structural steel [$\times 10^{-6}$ in/(in$\cdot$ $^\circ$F)]

6.495	6.595	6.615	6.635	6.485	6.555
6.665	6.505	6.435	6.625	6.715	6.655
6.755	6.625	6.715	6.575	6.655	6.605
6.565	6.515	6.555	6.395	6.775	6.685

  range: 6.395 to 6.775 $\times 10^{-6}$

- more insight is obtained by computing *descriptive statistics:*

  1. mean: location of the center of the data
  2. standard deviation and variance: spread of the data

Mean and standard deviation

given data points $y_1, \ldots, y_n$

Arithmetic mean

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

Standard deviation

$$s_y = \sqrt{\frac{S_t}{n-1}}, \quad S_t = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- measures the spread of data about mean
- if measurements are spread out widely around the mean, S_t (and s_y) will be large
- if they are grouped tightly, the standard deviation will be small
- the **variance** is the square of standard deviation:

$$s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} = \frac{\sum y_i^2 - (\sum y_i)^2/n}{n-1}$$

Coefficient of variation

Coefficient of variation

$$\text{c.v.} = \frac{s_y}{\bar{y}} \times 100\%$$

- provides a normalized measure of spread
- similar in spirit to relative error

Remark: S_t and s_y are based on $n-1$ degrees of freedom

- this nomenclature arises because $(\bar{y} - y_1) + (\bar{y} - y_2) + \cdots + (\bar{y} - y_n) = 0$
- if $\bar{y}$ is known and $n-1$ of the values are specified, the remaining value is fixed
- hence only $n-1$ of the values are freely determined
- another justification: there is no spread of a single data point
- however, it is also common to be defined by dividing by n instead of $n-1$

Example

6.495	6.595	6.615	6.635	6.485	6.555
6.665	6.505	6.435	6.625	6.715	6.655
6.755	6.625	6.715	6.575	6.655	6.605
6.565	6.515	6.555	6.395	6.775	6.685

- $n = 24$ measurements of coefficient of thermal expansion
- average (mean):

$$\sum y_i = 158.4, \quad \bar{y} = \frac{158.4}{24} = 6.6$$

- standard deviation and variance:

$$\sum (y_i - \bar{y})^2 = 0.217, \quad s_y = \sqrt{\frac{0.217}{24-1}} = 0.097133, \quad s_y^2 = 0.009435$$

- coefficient of variation

$$\text{c.v.} = \frac{0.097133}{6.6} \times 100\% = 1.47\%$$

indicates that the data are tightly clustered around the mean

Outline

- curve fitting and statistics
- **straight line fit to data**
- linearization of nonlinear equations
- fitting a polynomial to data
- multiple linear regression
- general linear least squares

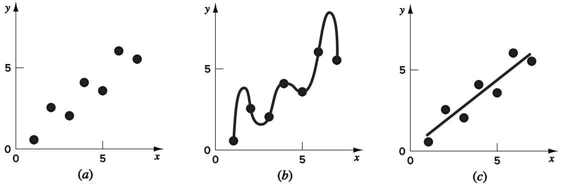
Straight line data fitting

simplest example of least squares: fitting a straight line to observations

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

Line model: $y = a_0 + a_1 x + e$ where a_0, a_1 are to determined based on data

- a_0 is intercept
- a_1 is slope
- $e = y - a_0 - a_1 x$ is *error* or *residual*
- residual is discrepancy between true value of y and approximate value $a_0 + a_1 x$



Least squares fit of straight line

minimize the sum of squared residuals over data:

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

- called *linear regression*
- to find a_0 and a_1 that minimize S_r, we set partial derivatives w.r.t. a_0, a_1 to zero:

$$\frac{\partial S_r}{\partial a_0} = -2 \sum_i (y_i - a_0 - a_1 x_i) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum_i (y_i - a_0 - a_1 x_i) x_i = 0$$

- yields a unique line for a given data set

Solution

- rewriting previous equation as

$$
\begin{aligned}
-\sum_i y_i + na_0 + a_1 \sum_i x_i &= 0 \\
-\sum_i (y_i x_i) + a_0 \sum_i x_i + a_1 \sum_i x_i^2 &= 0
\end{aligned}
$$

- which can be written as:

$$
\begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix}
$$

these are called the *normal equations*

- solving the normal equations:

$$
a_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}, \quad a_0 = \bar{y} - a_1 \bar{x}
$$

where $\bar{x}$ and $\bar{y}$ are the sample means of x and y

Example

fit a straight line to the x and y values in the table

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

- compute the following quantities:

$$n = 7, \quad \sum x_i = 28, \quad \bar{x} = \frac{28}{7} = 4$$
$$\sum y_i = 24, \quad \bar{y} = \frac{24}{7} = 3.428571$$
$$\sum x_i y_i = 119.5, \quad \sum x_i^2 = 140$$

- thus

$$a_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{7(119.5) - (28)(24)}{7(140) - (28)^2} = 0.8392857$$
$$a_0 = \bar{y} - a_1\bar{x} = 3.428571 - (0.8392857)(4) = 0.07142857$$

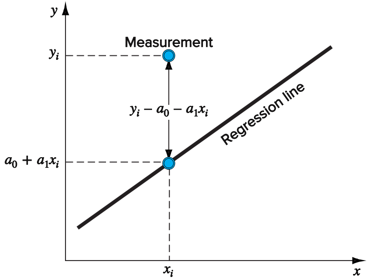
and

$$y = 0.07142857 + 0.8392857x$$

Residuals and error analysis

Error for the linear fit

x_i	y_i	$(y_i - \bar{y})^2$	$(y_i - a_0 - a_1 x_i)^2$
1	0.5	8.5765	0.1687
2	2.5	0.8622	0.5625
3	2.0	2.0408	0.3473
4	4.0	0.3265	0.3265
5	3.5	0.0051	0.5896
6	6.0	6.6122	0.7972
7	5.5	4.2908	0.1993
Σ	24.0	22.7143	2.9911



sum of squared residuals:

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 = 2.9911$$

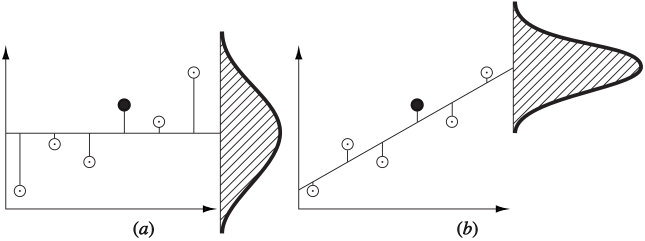
- least squares line is unique: any other line gives a larger S_r
- residuals quantify the vertical discrepancies between observed y_i and the line

Standard error of the estimate

a "standard deviation" for the regression line can be defined as

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

- $s_{y/x}$ is called the *standard error of the estimate*
- we divide by $n-2$ since two estimates (a_0 and a_1) were used to compute S_r
  - there is no such thing as the "spread of data" around a straight line connecting two points
- $s_{y/x}$ quantifies spread of data around the *regression line*



Coefficient of determination

- S_t: total sum of squares around the mean (before regression)
- S_r: sum of squares of residuals around regression line (unexplained error)
- $S_t - S_r$: improvement of straight line fit compared with average value

Normalized improvement

$$r^2 = \frac{S_t - S_r}{S_t} \implies r = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

- r^2: *coefficient of determination*
- r: *correlation coefficient*
- $r^2 = 1$: perfect fit ($S_r = 0$)
- $r^2 = 0$: no improvement ($S_r = S_t$)

Example

compute total standard deviation, standard error of estimate, and correlation coefficient for data in last example

- standard deviation:

$$s_y = \sqrt{\frac{22.7143}{7-1}} = 1.9457$$

- standard error of the estimate:

$$s_{y/x} = \sqrt{\frac{2.9911}{7-2}} = 0.7735$$

  since $s_{y/x} < s_y$, the linear regression model has merit

- extent of improvement is quantified by

$$r^2 = \frac{22.7143 - 2.9911}{22.7143} = 0.868 \quad \text{or} \quad r = \sqrt{0.868} = 0.932$$

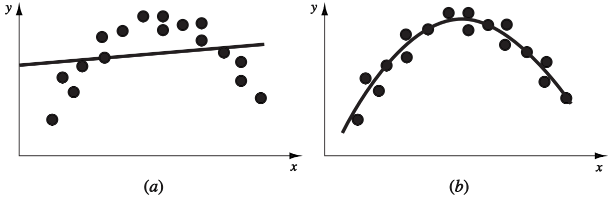
- interpretation: 86.8% of original uncertainty has been explained by linear model
  - caution: high r does not always imply a good fit

Outline

- curve fitting and statistics
- straight line fit to data
- **linearization of nonlinear equations**
- fitting a polynomial to data
- multiple linear regression
- general linear least squares

Linear transformation

- line fitting assumes linear relation between dep. and indep. variables
- always begin regression analysis by *plotting the data*
- for nonlinear data, other approaches are required such as polynomial regression



- nonlinear models can sometime be **transformed** into linear form
  - linear regression can then be applied to estimate coefficients
  - results must be transformed back for predictive use

Exponential model

$$y = \alpha_1 e^{\beta_1 x}$$

- α_1, β_1 are constants
- models growth or decay (population, radioactive decay)
- nonlinear for $\beta_1 \neq 0$

Linearization: take natural log:

$$\ln y = \ln \alpha_1 + \beta_1 x$$

- plot $\ln y$ vs x
- slope = β_1, intercept = $\ln \alpha_1$

Power model

$$y = \alpha_2 x^{\beta_2}$$

- α_2, β_2 are constants
- widely used in engineering (*e.g.*, scaling laws)

Linearization: take base-10 log:

$$\log y = \beta_2 \log x + \log \alpha_2$$

- plot $\log y$ vs $\log x$
- slope = β_2, intercept = $\log \alpha_2$

Saturation-growth-rate model

$$y = \frac{\alpha_3 x}{\beta_3 + x}$$

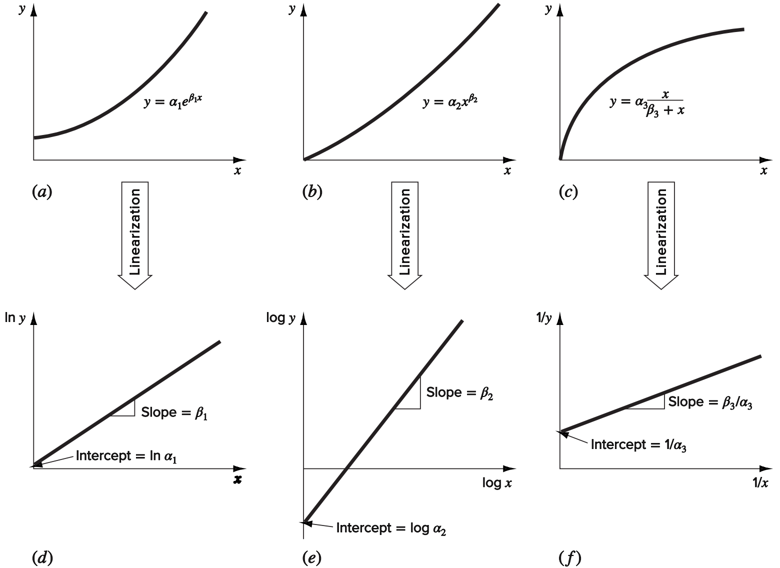
- used for population growth under limiting conditions
- levels off (saturates) as x increases

Linearization: invert the equation:

$$\frac{1}{y} = \frac{\beta_3}{\alpha_3}\frac{1}{x} + \frac{1}{\alpha_3}$$

- plot $1/y$ vs $1/x$
- slope = β_3/α_3, intercept = $1/\alpha_3$

Summary


y
$y = \alpha_1 e^{\beta_1 x}$
x
(a)
Linearization
y
$y = \alpha_2 x^{\beta_2}$
x
(b)
Linearization
y
$y = \alpha_3 \frac{x}{\beta_3 + x}$
x
(c)
Linearization
ln y
Slope = β_1
Intercept = ln α_1
x
(d)
log y
Slope = β_2
log x
Intercept = log α_2
(e)
1/y
Slope = β_3/α_3
Intercept = $1/\alpha_3$
1/x
(f)

Example

we fit data to the model $y = \alpha_2 x^{\beta_2}$

x	y	$\log x$	$\log y$
1	0.5	0.000	-0.301
2	1.7	0.301	0.226
3	3.4	0.477	0.534
4	5.7	0.602	0.753
5	8.4	0.699	0.922

- take logarithm:

$$\log y = \beta_2 \log x + \log \alpha_2$$

- this is a linear equation in $\log x$ and $\log y$
- apply linear regression to the transformed data to find β_2 and $\log \alpha_2$

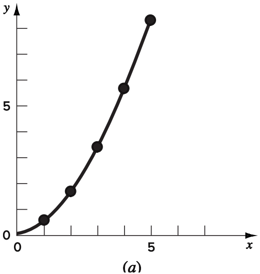
Example

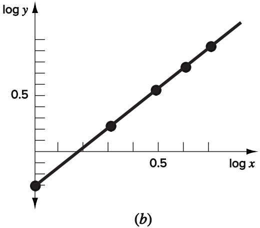
linear regression of the log-transformed data yields:

$$\log y = 1.75\ \log x - 0.300$$

- slope: $\beta_2 = 1.75$
- intercept: $\log \alpha_2 = -0.300 \Longrightarrow \alpha_2 = 10^{-0.300} \approx 0.501$
- final model:

$$y = 0.501\, x^{1.75}$$





Outline

- curve fitting and statistics
- straight line fit to data
- linearization of nonlinear equations
- **fitting a polynomial to data**
- multiple linear regression
- general linear least squares

Quadratic model and least squares objective

suppose data are related by a quadratic model:

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

- (a_0, a_1, a_2) are model parameters to be determined
- given data $(x_1, y_1), \ldots, (x_n, y_n)$, the residual sum of squares is

$$S_r = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i - a_2 x_i^2\right)^2$$

- we minimize S_r by setting partial derivatives to zero

$$\frac{\partial S_r}{\partial a_0} = -2 \sum (y_i - a_0 - a_1 x_i - a_2 x_i^2) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum x_i (y_i - a_0 - a_1 x_i - a_2 x_i^2) = 0$$

$$\frac{\partial S_r}{\partial a_2} = -2 \sum x_i^2 (y_i - a_0 - a_1 x_i - a_2 x_i^2) = 0$$

Normal equations for the quadratic

- collecting terms yields a 3×3 linear system:

$$na_0 + \left(\sum x_i\right)a_1 + \left(\sum x_i^2\right)a_2 = \sum y_i$$

$$\left(\sum x_i\right)a_0 + \left(\sum x_i^2\right)a_1 + \left(\sum x_i^3\right)a_2 = \sum x_i y_i$$

$$\left(\sum x_i^2\right)a_0 + \left(\sum x_i^3\right)a_1 + \left(\sum x_i^4\right)a_2 = \sum x_i^2 y_i$$

- in matrix form:

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{bmatrix}$$

- solve for (a_0, a_1, a_2) with any linear solver

General mth-order polynomial regression

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m + e$$

- minimize $S_r = \sum_{i=1}^{n} (y_i - \sum_{k=0}^{m} a_k x_i^k)^2$ by setting partial derivatives to zero:

$$\begin{bmatrix} n & \sum x_i & \cdots & \sum x_i^m \\ \sum x_i & \sum x_i^2 & \cdots & \sum x_i^{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_i^m & \sum x_i^{m+1} & \cdots & \sum x_i^{2m} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \vdots \\ \sum x_i^m y_i \end{bmatrix}$$

- results in $m{+}1$ normal equations in $m{+}1$ unknowns
- *standard error of the estimate:*

$$s_{y/x} = \sqrt{\frac{S_r}{n - (m+1)}}$$

- *coefficient of determination:* $r^2 = \dfrac{S_t - S_r}{S_t}$, where $S_t = \sum_{i=1}^{n} (y_i - \bar{y})^2$

Example: fit a quadratic

fit quadratic $y = a_0 + a_1 x + a_2 x^2 + e$ model to data

x_i	0	1	2	3	4	5
y_i	2.1	7.7	13.6	27.2	40.9	61.1

- we have

$$n = 6, \quad \sum x_i = 15, \quad \sum x_i^2 = 55, \quad \sum x_i^3 = 225, \quad \sum x_i^4 = 979$$
$$\sum y_i = 152.6, \quad \sum x_i y_i = 585.6, \quad \sum x_i^2 y_i = 2488.8$$

- normal equations:

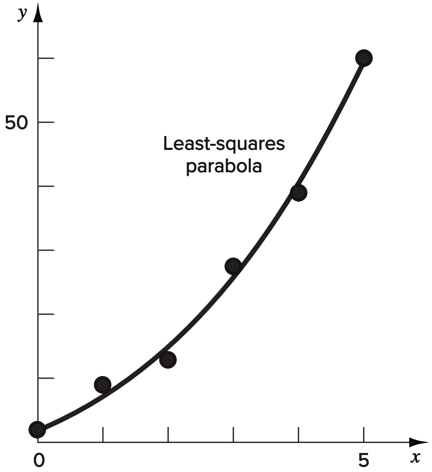
$$\begin{bmatrix} 6 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 152.6 \\ 585.6 \\ 2488.8 \end{bmatrix}$$

- solution: $a_0 = 2.47857$, $a_1 = 2.35929$, $a_2 = 1.86071$

- quadratic fit:

$$y = 2.47857 + 2.35929\, x + 1.86071\, x^2$$

Example: fit a quadratic



Example: fit a quadratic

x_i	y_i	$(y_i - \bar{y})^2$	$\left(y_i - a_0 - a_1 x_i - a_2 x_i^2\right)^2$
0	2.1	544.44	0.14332
1	7.7	314.47	1.00286
2	13.6	140.03	1.08158
3	27.2	3.12	0.80491
4	40.9	239.22	0.61951
5	61.1	1272.11	0.09439
Σ	152.6	2513.39	3.74657

- from the residuals table: $S_r = 3.74657$, $S_t = 2513.39$
- standard error (quadratic, $m+1 = 3$ parameters):

$$s_{y/x} = \sqrt{\frac{S_r}{n-(m+1)}} = \sqrt{\frac{3.74657}{6-3}} = 1.12$$

- coefficient of determination:

$$r^2 = \frac{S_t - S_r}{S_t} = \frac{2513.39 - 3.74657}{2513.39} = 0.99851, \qquad r = 0.99925$$

so 99.851% of original variability is explained by quadratic model; fit is excellent

Outline

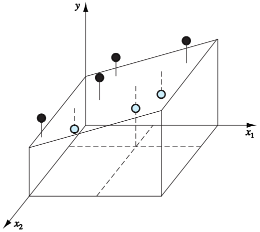
- curve fitting and statistics
- straight line fit to data
- linearization of nonlinear equations
- fitting a polynomial to data
- **multiple linear regression**
- general linear least squares

Multiple linear regression

linear model with multiple predictors:

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_m x_m + e$$

- in data-fitting, y is *outcome* and $x_1, \ldots, x_m$ are *features* or *regressors*
- for two predictors, the best-fit "line" becomes a plane in (x_1, x_2, y)



- choose coefficients $\{a_j\}_{j=0}^{m}$ that minimize the sum of squared residuals

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_{1i} - \cdots - a_m x_{mi})^2$$

over data (x_i, y_i) for $i = 1, \ldots, n$ where $x_i = (x_{1i}, \ldots, x_{mi})$ is an m-vector

Least squares plane fit

for $m = 2$

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})^2$$

take partial derivatives and set to zero:

$$\begin{aligned}
\frac{\partial S_r}{\partial a_0} &= -2 \sum (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i}) = 0 \\
\frac{\partial S_r}{\partial a_1} &= -2 \sum x_{1i} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i}) = 0 \\
\frac{\partial S_r}{\partial a_2} &= -2 \sum x_{2i} (y_i - a_0 - a_1 x_{1i} - a_2 x_{2i}) = 0
\end{aligned}$$

matrix (normal equations) form:

$$\begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i} x_{2i} \\ \sum x_{2i} & \sum x_{1i} x_{2i} & \sum x_{2i}^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_{1i} y_i \\ \sum x_{2i} y_i \end{bmatrix}$$

Example

find model $y = a_0 + a_1x_1 + a_2x_2$) that fits the data:

	y	x_1	x_2	x_1^2	x_2^2	x_1x_2	x_1y	x_2y
	5	0	0	0	0	0	0	0
	10	2	1	4	1	2	20	10
	9	2.5	2	6.25	4	5	22.5	18
	0	1	3	1	9	3	0	0
	3	4	6	16	36	24	12	18
	27	7	2	49	4	14	189	54
Σ	54	16.5	14	76.25	54	48	243.5	100

$$\sum y = 54, \quad \sum x_1 = 16.5, \quad \sum x_2 = 14$$

$$\sum x_1^2 = 76.25, \quad \sum x_2^2 = 54, \quad \sum x_1x_2 = 48$$

$$\sum x_1y = 243.5, \quad \sum x_2y = 100$$

normal equations:

$$\begin{bmatrix} 6 & 16.5 & 14 \\ 16.5 & 76.25 & 48 \\ 14 & 48 & 54 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 54 \\ 243.5 \\ 100 \end{bmatrix} \Rightarrow a_0 = 5,\ a_1 = 4,\ a_2 = -3$$

Goodness of fit and uncertainty

- residual sum of squares

$$S_r = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \qquad \hat{y}_i = a_0 + a_1 x_{1i} + \cdots + a_m x_{mi}$$

- total sum of squares about the mean $\bar{y}$:

$$S_t = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- standard error of the estimate (multiple regression with m predictors)

$$s_{y/x} = \sqrt{\frac{S_r}{n - (m+1)}}$$

- coefficient of determination (explained variance fraction)

$$r^2 = \frac{S_t - S_r}{S_t}, \qquad 0 \le r^2 \le 1$$

Power-law via multiple linear regression

- many engineering relations are multiplicative:

$$y = a_0\, x_1^{a_1}\, x_2^{a_2} \cdots x_m^{a_m}$$

- take logarithms to linearize:

$$\log y = \log a_0 + a_1 \log x_1 + \cdots + a_m \log x_m$$

- perform multiple linear regression with response $\log y$ and predictors $\log x_k$
- recover coefficients via $a_0 = 10^{\text{intercept}}$, exponents a_k are the slopes

Outline

- curve fitting and statistics
- straight line fit to data
- linearization of nonlinear equations
- fitting a polynomial to data
- multiple linear regression
- **general linear least squares**

Linear-in-parameters model

model is *linear-in-parameter*

$$y = a_0 z_0 + a_1 z_1 + a_2 z_2 + \cdots + a_m z_m + e$$

- $z_0, \ldots, z_m$ are *basis functions*/*feature mapping* that we choose; e is residual
- the term "linear" refers only to linearity in the parameters a_j
- basis functions z_j may be nonlinear (*e.g.*, $z_j = \sin(\omega t)$)

Examples

- *simple linear regression (line model):* $z_0 = 1$, $z_1 = x$
- *polynomial regression:* $z_0 = 1$, $z_1 = x$, $z_2 = x^2, \ldots, z_m = x^m$
- *multiple linear regression:* $z_0 = 1$, $z_1 = x_1$, $z_2 = x_2, \ldots$
- $y = a_0 + a_1 \cos(\omega t) + a_2 \sin(\omega t)$, $z_0 = 1$, $z_1 = \cos(\omega t)$, $z_2 = \sin(\omega t)$

Normal equations

given data $(z_i, y_i)_{i=1}^n$ with $z_i = (z_{0i}, \ldots, z_{mi})$, the least squares criterion minimizes

$$S_r = \sum_{i=1}^{n} \Big(y_i - \sum_{j=0}^{m} z_{ji} a_j\Big)^2$$

- called *linear regression* or *least squares regression*
- differentiating w.r.t. each a_j and setting to zero yields the **normal equations**:

$$Z^T Z a = Z^T y$$

$$Z = \begin{bmatrix} z_{01} & z_{11} & \cdots & z_{m1} \\ z_{02} & z_{12} & \cdots & z_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ z_{0n} & z_{1n} & \cdots & z_{mn} \end{bmatrix} \in \mathbb{R}^{n \times m}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad a = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix}$$

  if $Z^T Z$ is invertible, then the solution is unique $a = (Z^T Z)^{-1} Z^T y$

  – in MATLAB: `a=Z\y`, which is called least squares *approximate* solution to $Za = y$
- this unifies linear, polynomial, and multiple regression under one framework

References and further readings

- S. C. Chapra and R. P. Canale. *Numerical Methods for Engineers* (8th edition). McGraw Hill, 2021. (Ch.17)

- S. C. Chapra. *Applied Numerical Methods with MATLAB for Engineers and Scientists* (5th edition). McGraw Hill, 2023. (Ch.14, 15)