L22: Linear Regression

Part 1: $y = ax$ and $y = ax + b$

Lucas A. J. Bastien

E7 Spring 2017, University of California at Berkeley

March 13, 2017

Version: release

Announcements

Lab 08 is due on March 17 at 12 pm (noon)

Today:

- Linear regression
  - Introduction
  - Specific cases:
    - $y = ax$
    - $y = ax + b$
    - "Transform" $y = bx^a$ into a linear form

Wednesday:

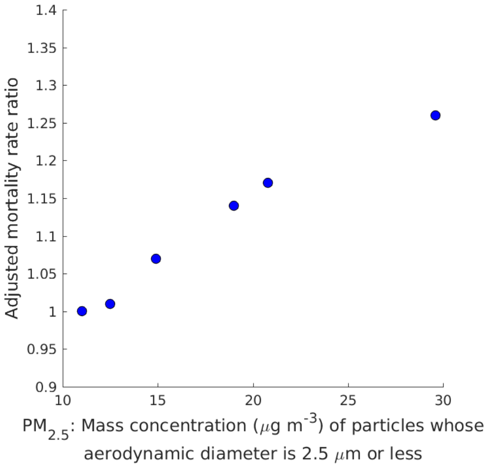
- Linear regression, continued
  - General approach

Wednesday:

- Linear regression: discussion and applications

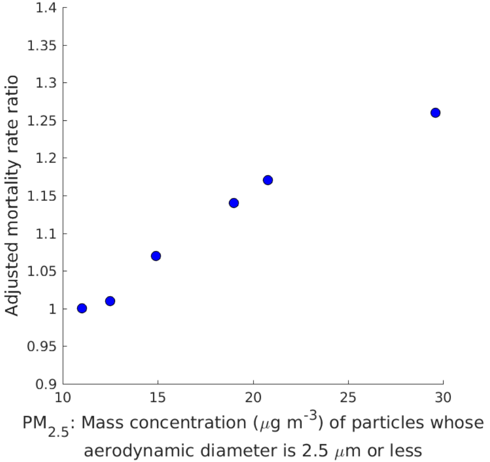
Introduction to linear regression

Harvard six cities study:



Data from: Dockery et al. (1993). An Association Between Air Pollution and Mortality in Six U.S. Cities. *The New England Journal of Medicine*, 329 (24) 1753–1759. Each data point corresponds to a different city.

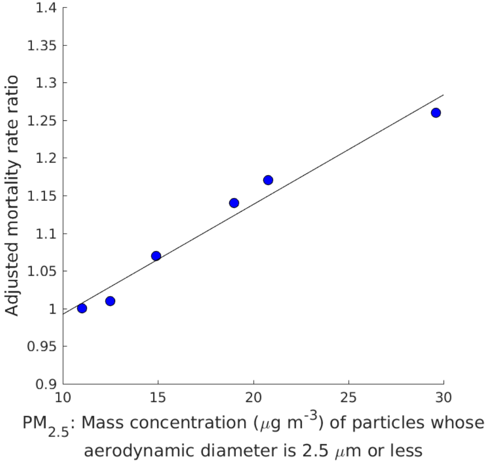
Harvard six cities study: It appears that there is a correlation between exposure to particulate matter pollution and premature mortality.



Data from: Dockery et al. (1993). An Association Between Air Pollution and Mortality in Six U.S. Cities. *The New England Journal of Medicine*, 329 (24) 1753–1759. Each data point corresponds to a different city.

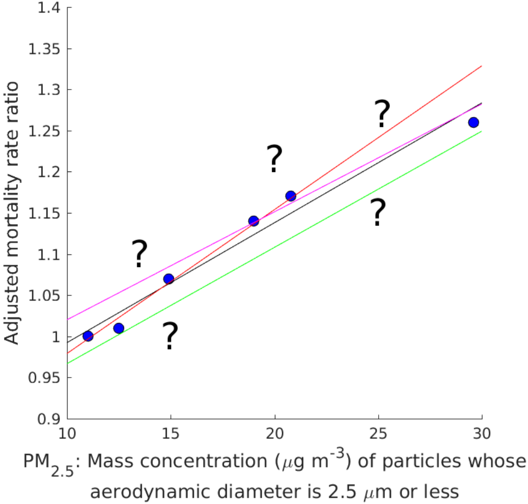
Introduction to linear regression

Harvard six cities study: It appears that there is a correlation between exposure to particulate matter pollution and premature mortality. We can fit a straight line to the data reasonably well



Data from: Dockery et al. (1993). An Association Between Air Pollution and Mortality in Six U.S. Cities. *The New England Journal of Medicine*, 329 (24) 1753–1759. Each data point corresponds to a different city.

How to choose the line that “fits best”?



Data from: Dockery et al. (1993). An Association Between Air Pollution and Mortality in Six U.S. Cities. *The New England Journal of Medicine*, 329 (24) 1753–1759. Each data point corresponds to a different city.

Linear regression: goals and motivation

Given two sets of data (x- and y-values), **can we model one set as a function of the other**, using a "simple function" (*e.g.*, straight line, polynomial)?

Linear regression: goals and motivation

Given two sets of data (x- and y-values), **can we model one set as a function of the other**, using a "simple function" (*e.g.*, straight line, polynomial)?

For a given function shape (*e.g.*, straight line, polynomial), **how to obtain the "best fit"?** For example:

- What is the best straight line that we can fit to the data?
- What is the best polynomial that we can fit to the data?

Linear regression: goals and motivation

Given two sets of data (x- and y-values), **can we model one set as a function of the other**, using a "simple function" (*e.g.*, straight line, polynomial)?

For a given function shape (*e.g.*, straight line, polynomial), **how to obtain the "best fit"?** For example:

- What is the best straight line that we can fit to the data?
- What is the best polynomial that we can fit to the data?

In the context of linear regression, **what do "simple function" and "best fit"** mean?

Linear regression: goals and motivation

Given two sets of data (x- and y-values), **can we model one set as a function of the other**, using a "simple function" (*e.g.*, straight line, polynomial)?

For a given function shape (*e.g.*, straight line, polynomial), **how to obtain the "best fit"?** For example:

- What is the best straight line that we can fit to the data?
- What is the best polynomial that we can fit to the data?

In the context of linear regression, **what do "simple function" and "best fit"** mean?

Linear regression does not necessarily match all points exactly (it does when it can, though)

Linear regression: goals and motivation

Given two sets of data (x- and y-values), **can we model one set as a function of the other**, using a "simple function" (*e.g.*, straight line, polynomial)?

For a given function shape (*e.g.*, straight line, polynomial), **how to obtain the "best fit"?** For example:

- What is the best straight line that we can fit to the data?
- What is the best polynomial that we can fit to the data?

In the context of linear regression, **what do "simple function" and "best fit"** mean?

Linear regression does not necessarily match all points exactly (it does when it can, though)

How can we **measure how good or bad a fit** is?

What type of "simple function" can be fitted?

Answer: any function of the form:

$$y = a_1 f_1(x) + a_2 f_2(x) + \ldots a_n f_n(x)$$

where:

- The a_i's are the coefficients to be determined using linear regression
- The f_i's are real-valued functions

(*i.e.* any **linear** combination of known functions)

What type of "simple function" can be fitted?

Answer: any function of the form:

$$y = a_1 f_1(x) + a_2 f_2(x) + \ldots a_n f_n(x)$$

where:

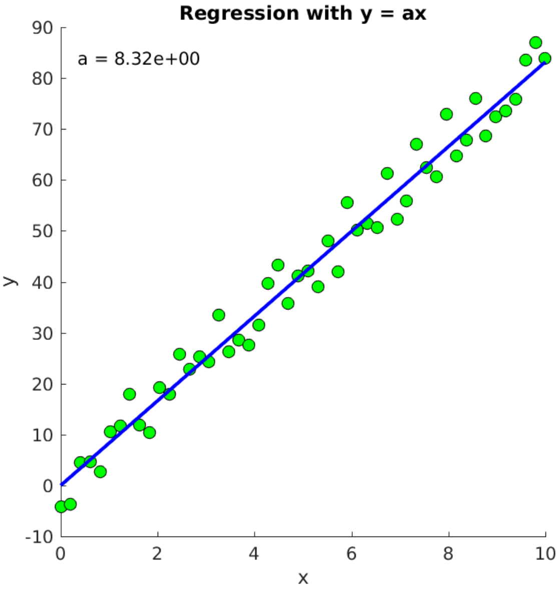
- The a_i's are the coefficients to be determined using linear regression
- The f_i's are real-valued functions

(*i.e.* any **linear** combination of known functions)

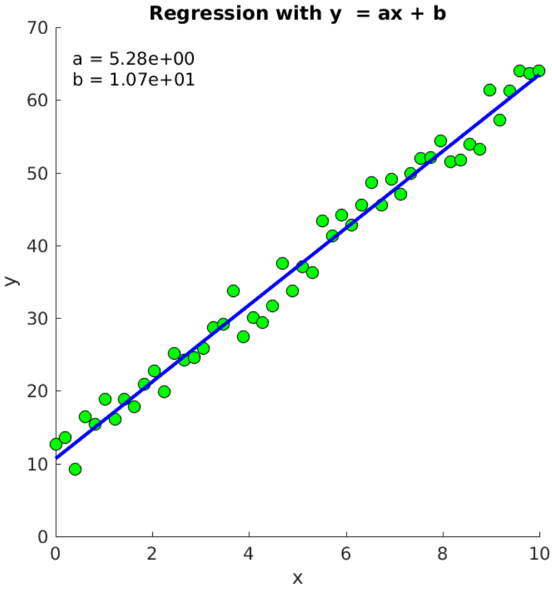
Examples:

- $y = ax$ (coefficient: a)
- $y = ax + b$ (coefficients: a and b)
- $y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$ (coefficients: a_0, a_1, a_2, and a_3)
- $y = a\cos(x) + b\sin(x) + c$ (coefficients: a, b, and c)

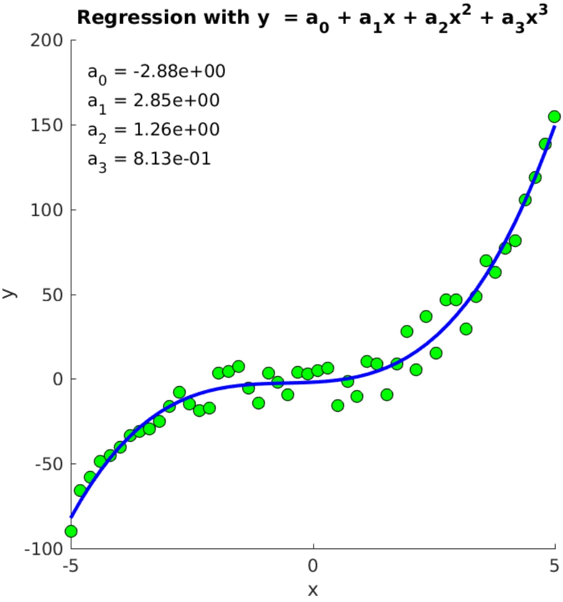
Example of linear regression: $y = ax$



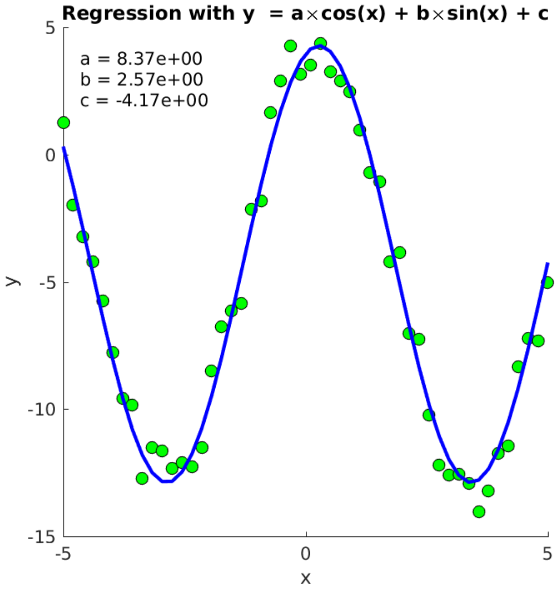
Example of linear regression: $y = ax + b$



Example of linear regression: $y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$



Example of linear regression: $a \times \cos(x) + b \times \sin(x) + c$



Today and Wednesday

Today:

- We fit lines of equation $y = ax$ and $y = ax + b$
- We derive the formulas for a and b by hand
- We calculate a and b "manually"

Today and Wednesday

Today:

- We fit lines of equation $y = ax$ and $y = ax + b$
- We derive the formulas for a and b by hand
- We calculate a and b "manually"

Next lecture (L23, Wednesday March 15):

- We fit any line of equation $y = a_1 f_1(x) + a_2 f_2(x) + \dots a_n f_n(x)$
- We let Matlab do most of the work

Definition of “best fit” for linear regression

Consider that we have:

- A set of x- and y-data (n data points)
- A set of predicted y values

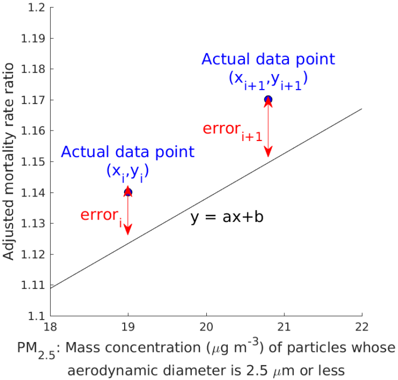
How good is the fit?

Definition of "best fit" for linear regression

Consider that we have:

- A set of *x*- and *y*-data (*n* data points)
- A set of predicted *y* values

How good is the fit?



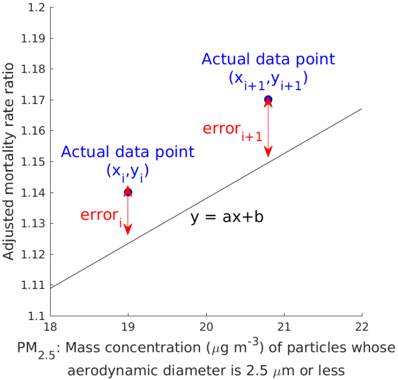
Data from: Dockery et al. (1993). An Association Between Air Pollution and Mortality in Six U.S. Cities. *The New England Journal of Medicine*, 329 (24) 1753–1759. Each data point corresponds to a different city.

Definition of "best fit" for linear regression

Consider that we have:

- A set of x- and y-data (n data points)
- A set of predicted y values

How good is the fit?



Total squared error E_2:
("square error" for short)

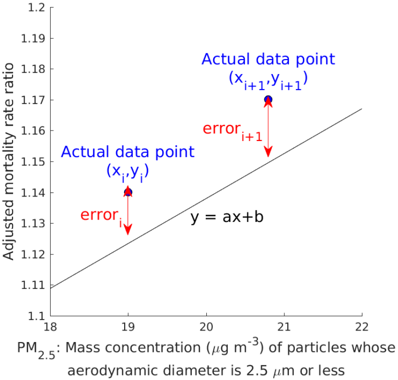
$$E_2 = \sum_{i=1}^{n} (\text{error}_i)^2$$
$$= \sum_{i=1}^{n} (y_{\text{i,predicted}} - y_{\text{i}})^2$$

Data from: Dockery et al. (1993). An Association Between Air Pollution and Mortality in Six U.S. Cities. *The New England Journal of Medicine*, 329 (24) 1753–1759. Each data point corresponds to a different city.

Consider that we have:

- A set of x- and y-data (n data points)
- A set of predicted y values

How good is the fit?



Total squared error E_2:
("square error" for short)

$$E_2 = \sum_{i=1}^{n} (\text{error}_i)^2$$
$$= \sum_{i=1}^{n} (y_{\text{i,predicted}} - y_{\text{i}})^2$$

In this example:

$$E_2 = \sum_{i=1}^{n} (ax_{\text{i}} + b - y_{\text{i}})^2$$

Data from: Dockery et al. (1993). An Association Between Air Pollution and Mortality in Six U.S. Cities. *The New England Journal of Medicine*, 329 (24) 1753–1759. Each data point corresponds to a different city.

Definition of “best fit” for linear regression

Total squared error E_2:
(“square error” for short)

$$E_2 = \sum_{i=1}^{n} (\mathrm{error}_i)^2$$
$$= \sum_{i=1}^{n} (y_{\mathrm{i,predicted}} - y_{\mathrm{i}})^2$$

Definition of “best fit” for linear regression

Total squared error E_2:
(“square error” for short)

$$E_2 = \sum_{i=1}^{n} (\mathrm{error}_i)^2$$
$$= \sum_{i=1}^{n} (y_{\mathrm{i,predicted}} - y_{\mathrm{i}})^2$$

Best fit:

(i.e. the best choice of parameters)

The fit that minimizes the total squared error

Fitting a line of equation $y = ax$: manual derivation

Objective: **given a set of x- and y-data, what is the value of a such that the line of equation $y = ax$ is the best possible fit to the data?**

Fitting a line of equation $y = ax$: manual derivation

Objective: **given a set of x- and y-data, what is the value of a such that the line of equation $y = ax$ is the best possible fit to the data?**

The total square error E_2 is a function of a:

$$
\begin{aligned}
E_2(a) &= \sum_{i=1}^{n} (y_{\mathrm{i,predicted}} - y_{\mathrm{i}})^2 \\
&= \sum_{i=1}^{n} (ax_{\mathrm{i}} - y_{\mathrm{i}})^2
\end{aligned}
$$

Fitting a line of equation $y = ax$: manual derivation

Objective: given a set of x- and y-data, what is the value of a such that the line of equation $y = ax$ is the best possible fit to the data?

The total square error E_2 is a function of a:

$$E_2(a) = \sum_{i=1}^{n} (y_{\mathrm{i,predicted}} - y_{\mathrm{i}})^2$$
$$= \sum_{i=1}^{n} (ax_{\mathrm{i}} - y_{\mathrm{i}})^2$$

If E_2 is minimum for $a = a_{\mathrm{min}}$, then $E_2'(a_{\mathrm{min}}) = 0$

$$E_2'(a) = \sum_{i=1}^{n} 2x_{\mathrm{i}}(ax_{\mathrm{i}} - y_{\mathrm{i}})$$
$$= \sum_{i=1}^{n} 2ax_{\mathrm{i}}^2 - 2x_{\mathrm{i}}y_{\mathrm{i}}$$

Fitting a line of equation $y = ax$: manual derivation

$$E_2'(a) = \sum_{i=1}^{n} 2ax_{\mathrm{i}}^2 - 2x_{\mathrm{i}}y_{\mathrm{i}}$$

Fitting a line of equation $y = ax$: manual derivation

$$E_2'(a) = \sum_{i=1}^{n} 2ax_{\mathrm{i}}^2 - 2x_{\mathrm{i}}y_{\mathrm{i}}$$

Assuming that at least one of the x_i is non-zero *i.e.* :

$$\sum_{i=1}^{n} x_{\mathrm{i}}^2 \neq 0$$

we have:

$$E_2'(a) = 0 \iff a = \frac{\sum_{i=1}^{n} x_{\mathrm{i}}y_{\mathrm{i}}}{\sum_{i=1}^{n} x_{\mathrm{i}}^2}$$

Fitting a line of equation $y = ax$: manual derivation

$$E_2'(a) = \sum_{i=1}^{n} 2ax_{\mathrm{i}}^2 - 2x_{\mathrm{i}}y_{\mathrm{i}}$$

Assuming that at least one of the x_i is non-zero *i.e.* :

$$\sum_{i=1}^{n} x_{\mathrm{i}}^2 \neq 0$$

we have:

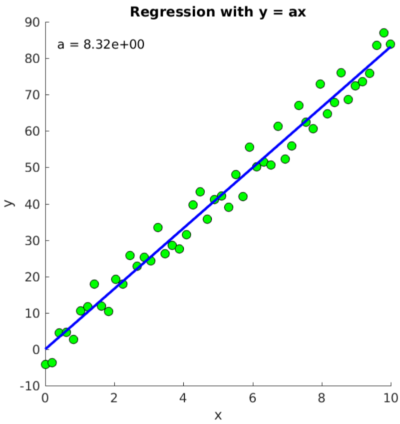
$$E_2'(a) = 0 \iff a = \frac{\sum_{i=1}^{n} x_{\mathrm{i}}y_{\mathrm{i}}}{\sum_{i=1}^{n} x_{\mathrm{i}}^2}$$

The value of a that yields the best fit for the line $y = ax$ is:

$$a = \frac{\sum_{i=1}^{n} x_{\mathrm{i}}y_{\mathrm{i}}}{\sum_{i=1}^{n} x_{\mathrm{i}}^2}$$

Fitting a line of equation $y = ax$: example

$$a = \frac{\sum_{i=1}^{n} x_\mathrm{i} y_\mathrm{i}}{\sum_{i=1}^{n} x_\mathrm{i}^2}$$




```
% Determine and plot the linear regression line
a = sum(x.*y) / sum(x.*x);
plot(x, a*x, 'b', 'LineWidth', 2)
```

Fitting a line of equation $y = ax + b$: manual derivation

Objective: **given a set of x- and y-data, what are the values of a and b such that the line of equation $y = ax + b$ is the best possible fit to the data?**

Fitting a line of equation $y = ax + b$: manual derivation

Objective: **given a set of x- and y-data, what are the values of a and b such that the line of equation $y = ax + b$ is the best possible fit to the data?**

The total square error E_2 is a function of a and b:

$$
\begin{aligned}
E_2(a, b) &= \sum_{i=1}^{n} (y_{\mathrm{i,predicted}} - y_{\mathrm{i}})^2 \\
&= \sum_{i=1}^{n} (ax_{\mathrm{i}} + b - y_{\mathrm{i}})^2
\end{aligned}
$$

Fitting a line of equation $y = ax + b$: manual derivation

Objective: given a set of x- and y-data, what are the values of a and b such that the line of equation $y = ax + b$ is the best possible fit to the data?

The total square error E_2 is a function of a and b:

$$E_2(a, b) = \sum_{i=1}^{n} (y_{\mathrm{i,predicted}} - y_{\mathrm{i}})^2$$
$$= \sum_{i=1}^{n} (ax_{\mathrm{i}} + b - y_{\mathrm{i}})^2$$

If E_2 is minimum for $a = a_{\min}$ and $b = b_{\min}$, then

$$\left.\frac{\partial E_2}{\partial a}\right|_{(a_{\min}, b_{\min})} = 0 \quad \text{and} \quad \left.\frac{\partial E_2}{\partial b}\right|_{(a_{\min}, b_{\min})} = 0$$

Fitting a line of equation $y = ax + b$: manual derivation

$$\frac{\partial E_2}{\partial a} = \sum_{i=1}^{n} 2x_\mathrm{i}(ax_\mathrm{i} + b - y_\mathrm{i})$$

$$\frac{\partial E_2}{\partial b} = \sum_{i=1}^{n} 2(ax_\mathrm{i} + b - y_\mathrm{i})$$

Fitting a line of equation $y = ax + b$: manual derivation

$$\frac{\partial E_2}{\partial a} = \sum_{i=1}^{n} 2x_{\mathrm{i}}(ax_{\mathrm{i}} + b - y_{\mathrm{i}})$$

$$\frac{\partial E_2}{\partial b} = \sum_{i=1}^{n} 2(ax_{\mathrm{i}} + b - y_{\mathrm{i}})$$

Solve the following system of two equations and two unknowns (a and b):

$$\frac{\partial E_2}{\partial a} = 0 \quad \text{and} \quad \frac{\partial E_2}{\partial b} = 0$$

Fitting a line of equation $y = ax + b$: manual derivation

$$\frac{\partial E_2}{\partial a} = \sum_{i=1}^{n} 2x_{\mathrm{i}}(ax_{\mathrm{i}} + b - y_{\mathrm{i}})$$

$$\frac{\partial E_2}{\partial b} = \sum_{i=1}^{n} 2(ax_{\mathrm{i}} + b - y_{\mathrm{i}})$$

Solve the following system of two equations and two unknowns (a and b):

$$\frac{\partial E_2}{\partial a} = 0 \quad \text{and} \quad \frac{\partial E_2}{\partial b} = 0$$

to obtain the values of a and b that yield the best fit for the line $y = ax + b$:

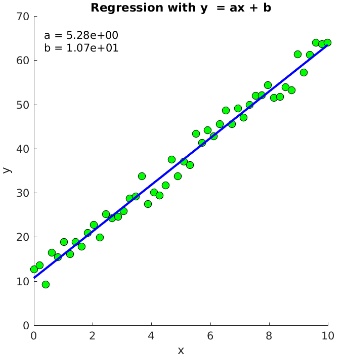
$$a = \frac{\sum_{i=1}^{n}(x_{\mathrm{i}} - \bar{x})(y_{\mathrm{i}} - \bar{y})}{\sum_{i=1}^{n}(x_{\mathrm{i}} - \bar{x})^2} \quad \text{and} \quad b = \bar{y} - a\bar{x}$$

where $\bar{x}$ and $\bar{y}$ are the mean values of the x- and y-data, respectively

Fitting a line of equation $y = ax + b$: example

$$a = \frac{\sum_{i=1}^{n}(x_{\mathrm{i}} - \bar{x})(y_{\mathrm{i}} - \bar{y})}{\sum_{i=1}^{n}(x_{\mathrm{i}} - \bar{x})^2}$$

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




```
% Determine and plot the linear regression line
x_mean = mean(x);
y_mean = mean(y);
a = sum((x-x_mean).*(y-y_mean)) / sum((x-x_mean).^2);
b = y_mean - a*x_mean;
plot(x, a*x+b, 'b', 'LineWidth', 2)
```

Coefficient of determination

The coefficient of determination is often written r^2 or R^2 and pronounced “R squared”. **It is another metric (in addition to the total square error E_2) that measures the goodness of fit**

Coefficient of determination

The coefficient of determination is often written r^2 or R^2 and pronounced "R squared". **It is another metric (in addition to the total square error E_2) that measures the goodness of fit**

$$
\begin{aligned}
r^2 &= 1 - \text{fraction of variance unexplained by the regression model} \\
&= 1 - \frac{\text{variance in } y \text{ not explained by the model}}{\text{variance in } y \text{ in the data}} \\
&= 1 - \frac{\sum_{i=1}^{n}(y_i - y_{\text{predicted,i}})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}
\end{aligned}
$$

$\bar{y}$ is the mean of the y-data

Coefficient of determination

The coefficient of determination is often written r^2 or R^2 and pronounced "R squared". **It is another metric (in addition to the total square error E_2) that measures the goodness of fit**

$$
\begin{aligned}
r^2 &= 1 - \text{fraction of variance unexplained by the regression model} \\
&= 1 - \frac{\text{variance in } y \text{ not explained by the model}}{\text{variance in } y \text{ in the data}} \\
&= 1 - \frac{\sum_{i=1}^{n}(y_i - y_{\text{predicted,i}})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}
\end{aligned}
$$

Property: $r^2 \leqslant 1$

The closer r^2 is to 1, the better the fit

$\bar{y}$ is the mean of the y-data

Making problems linear: example with the power law

Sometimes one wants to fit a line whose equation is not a linear combination of known functions

Making problems linear: example with the power law

Sometimes one wants to fit a line whose equation is not a linear combination of known functions

In some cases, it is possible to make the problem linear through mathematical manipulations

Making problems linear: example with the power law

Sometimes one wants to fit a line whose equation is not a linear combination of known functions

In some cases, it is possible to make the problem linear through mathematical manipulations

For example: fit the line of equation $y = bx^a$ (power law)

Making problems linear: example with the power law

Sometimes one wants to fit a line whose equation is not a linear combination of known functions

In some cases, it is possible to make the problem linear through mathematical manipulations

For example: fit the line of equation $y = bx^a$ (power law)

$$\begin{aligned} y &= bx^a \\ \ln(y) &= \ln(b) + a\ln(x) \\ Y &= c + aX \qquad\qquad \text{with } X = \ln(x) \\ & \qquad\qquad\qquad\qquad\quad Y = \ln(y) \\ & \qquad\qquad\qquad\qquad\quad c = \ln(b) \end{aligned}$$

Making problems linear: example with the power law

Sometimes one wants to fit a line whose equation is not a linear combination of known functions

In some cases, it is possible to make the problem linear through mathematical manipulations

For example: fit the line of equation $y = bx^a$ (power law)

$$y = bx^a$$

$$\ln(y) = \ln(b) + a\ln(x)$$

$$Y = c + aX \qquad \text{with} \quad X = \ln(x), \quad Y = \ln(y), \quad c = \ln(b)$$

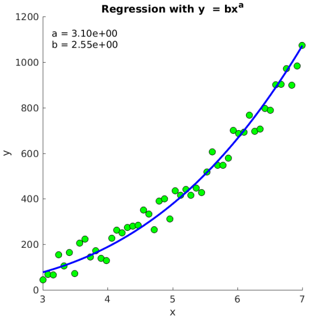
1. Calculate X and Y from x and y
2. Use linear regression on the X- and Y- data to determine the coefficients a and c
3. Calculate $b = \exp(c)$

Making problems linear: example with the power law

$$a = \frac{\sum_{i=1}^{n}(X_\mathrm{i} - \bar{X})(Y_\mathrm{i} - \bar{Y})}{\sum_{i=1}^{n}(X_\mathrm{i} - \bar{X})^2}$$

$$c = \bar{Y} - a\bar{X}$$

$$b = \exp(c)$$




```
% Determine and plot the linear regression line
logx = log(x);
logy = log(y);
logx_mean = mean(logx);
logy_mean = mean(logy);
a = sum((logx-logx_mean).*(logy-logy_mean)) / ...
    sum((logx-logx_mean).^2);
c = logy_mean - a*logx_mean;
b = exp(c);
plot(x, b*x.^a, 'b', 'LineWidth', 2)
```