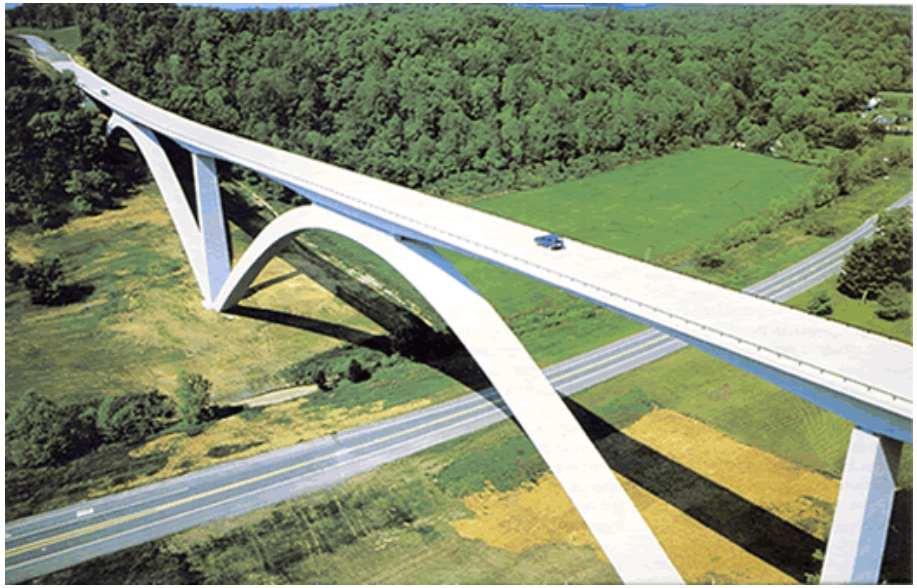
# Quantitative Risk Analysis Using Correlation and Simple Linear Regression

COORDINATOR:
Dr Lihai Zhang
Infrastructure Engineering Department (Room B307)
lihzhang@unimelb.edu.au



# Quantitative Analysis – Environmental Risks

An environmental engineer is studying the rate of absorption of a certain chemical into skin. She obtains a series of experimental results as follows

| Volume (mL) | Time (h) | Percent Absorbed |
|---|---|---|
| 0.05 | 2 | 48.3 |
| 0.05 | 2 | 51.0 |
| 0.05 | 2 | 54.7 |
| 2.00 | 10 | 63.2 |
| 2.00 | 10 | 67.8 |
| 2.00 | 10 | 66.2 |
| 5.00 | 24 | 83.6 |
| 5.00 | 24 | 85.1 |
| 5.00 | 24 | 87.8 |

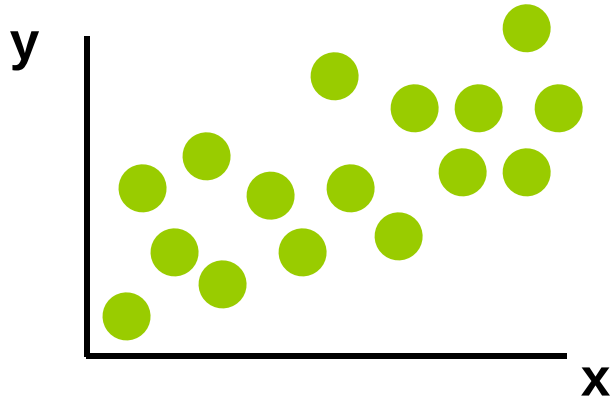

***Is any correlation between time and absorption?***

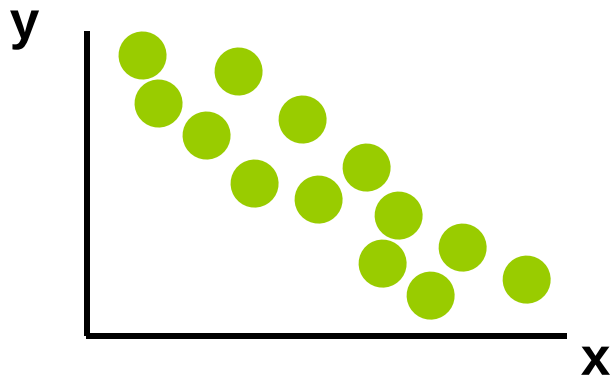
***Is any correlation between volume and absorption?***

# Scatter Plots

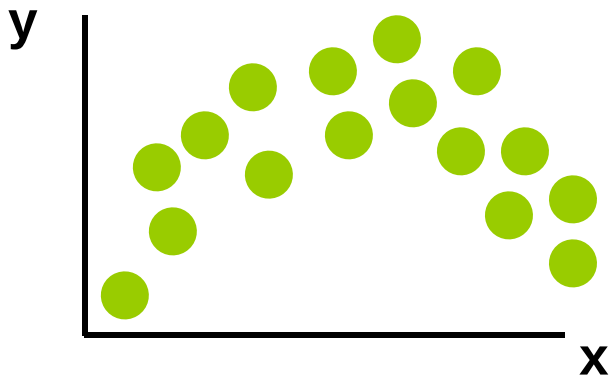
**Scatter Plots are used to show the relationship between two variables.**

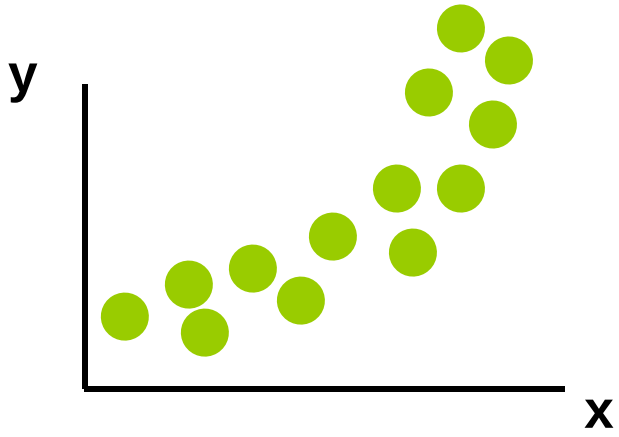
**Linear relationships**




**Nonlinear relationships**

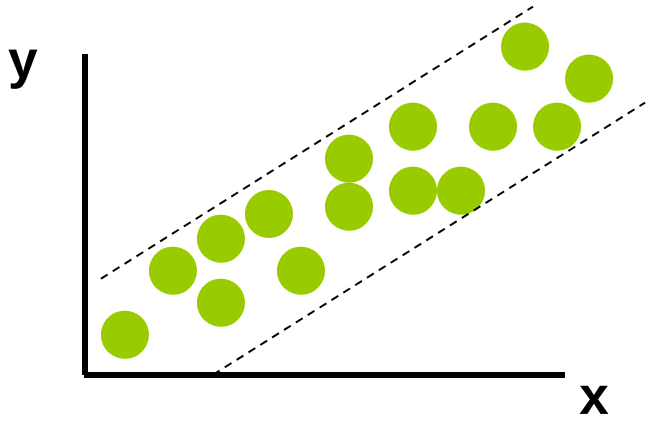


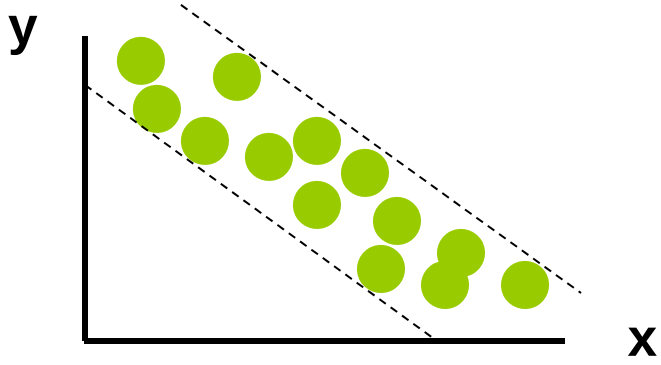



# Correlation

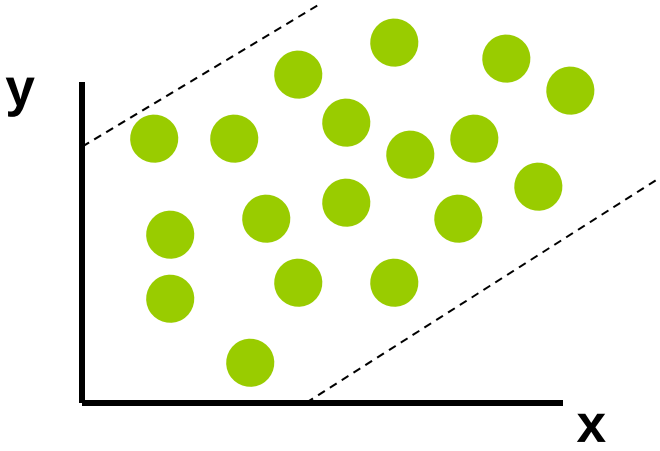
**Correlation analysis is used to measure strength of the association (linear relationship) between two variables**

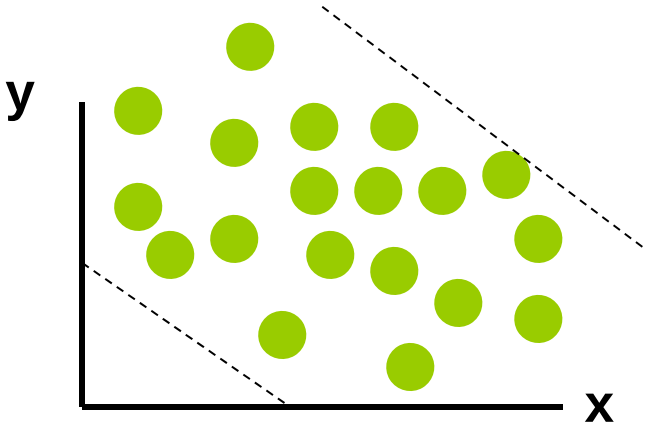
**Strong relationships**




**Weak relationships**

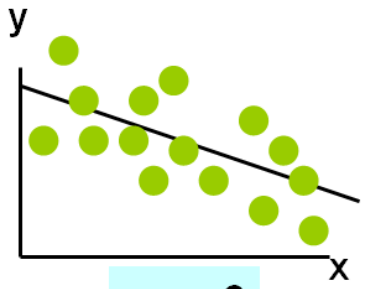


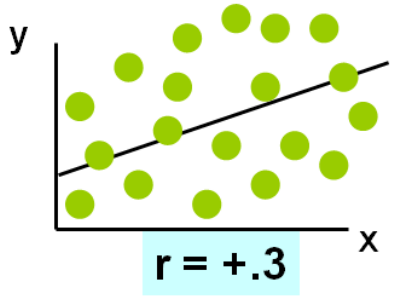



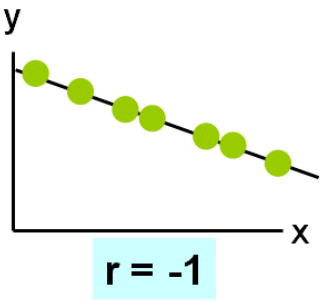
# Correlation

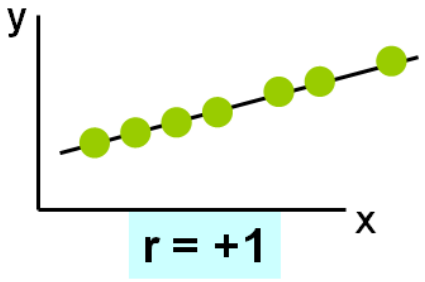
**Correlation coefficient** – A numerical measure of the strength of the linear relationship between two variables.

- Range between -1 and 1
- The closer to -1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker the linear relationship

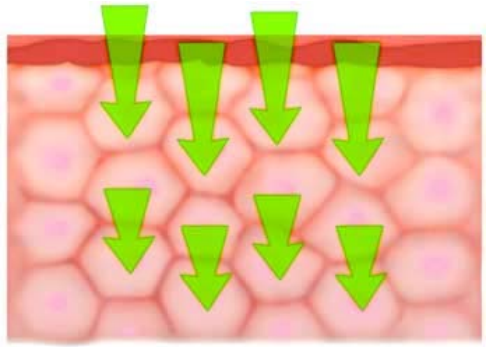


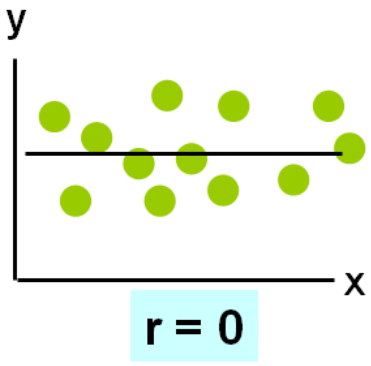









# Correlation Coefficient

## Sample correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}}$$

where: $r$ = Sample correlation coefficient
$n$ = Sample size
$x_i$ = Value of the independent variable
$y_i$ = Value of the dependent variable

# Quantitative Analysis – Environmental Risks

An environmental engineer is studying the rate of absorption of a certain chemical into skin. She obtains a series of experimental results as follows

| Volume (mL) | Time (h) | Percent Absorbed |
|---|---|---|
| 0.05 | 2 | 48.3 |
| 0.05 | 2 | 51.0 |
| 0.05 | 2 | 54.7 |
| 2.00 | 10 | 63.2 |
| 2.00 | 10 | 67.8 |
| 2.00 | 10 | 66.2 |
| 5.00 | 24 | 83.6 |
| 5.00 | 24 | 85.1 |
| 5.00 | 24 | 87.8 |


***(a) Is any correlation between time and absorption?***

***(b) Is any correlation between volume and absorption?***

**Solution:**

## Solution:

***(a) Is any correlation between time and absorption?***

| Time (h) x | Percent Absorbed y |
|---|---|
| 2 | 48.3 |
| 2 | 51.0 |
| 2 | 54.7 |
| 10 | 63.2 |
| 10 | 67.8 |
| 10 | 66.2 |
| 24 | 83.6 |
| 24 | 85.1 |
| 24 | 87.8 |
| $\overline{x} = 12$ | $\overline{y} = 67.5$ |

$$\overline{x} = \frac{1}{n_x = 9}\sum_{i=1}^{n} x_i = 12$$

$$\overline{y} = \frac{1}{n_y = 9}\sum_{i=1}^{n} y_i = 67.5$$

## Solution (continued):

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\left[\sum (x-\bar{x})^2\right]\left[\sum (y-\bar{y})^2\right]}}$$

$$= \frac{(2-12)(48.3-67.5)+(2-12)(51.0-67.5)+\ldots(24-12)(87.8-67.5)}{\sqrt{\left[(2-12)^2+\ldots(24-12)^2\right]\left[(48.3-67.5)^2+\ldots(87.8-67.5)^2\right]}}$$

$$= \frac{1143.6}{1159.2} = 0.987$$

***As r is close to 1, there is a stronger positive correlation between time and absorption.***

# Quantitative Analysis – Environmental Risks

```
1     % Environmental Risk
2
3 -   clear all;
4 -   close all;
5 -   clc;
6
7 -   volume = [0.05 0.05 0.05 2.00 2.00 2.00 5.00 5.00 5.00];
8 -   time = [2 2 2 10 10 10 24 24 24];
9 -   absorbtion = [48.3 51.0 54.7 63.2 67.8 66.2 83.6 85.1 87.8];
10
11    % correlation between time and absorbtion
12
13 -  r_tmp=corrcoef(time,absorbtion);
14 -  r = r_tmp(1,2);
15
16 -  display('Correlation coefficient for Absorbtion vs Time:');
17 -  display(r);
18 -  display('suggests a strong positive correlation relationship between time and absorption.');
19 -  fprintf('\n')
20
```

**Command Window**

```
Correlation coefficient for Absorbtion vs Time:

r =

    0.9866

suggests a strong positive correlation relationship between time and absorption.
```

## Solution (continued):

***(b) Is any correlation between volume and absorption?***

| Volume (mL) x | Percent Absorbed y |
|---|---|
| 0.05 | 48.3 |
| 0.05 | 51.0 |
| 0.05 | 54.7 |
| 2.00 | 63.2 |
| 2.00 | 67.8 |
| 2.00 | 66.2 |
| 5.00 | 83.6 |
| 5.00 | 85.1 |
| 5.00 | 87.8 |
| $\overline{x}$ = 2.35 | $\overline{y}$ = 67.5 |

$$\overline{x} = \frac{1}{n_x = 9}\sum_{i=1}^{n} x_i = 2.35$$

$$\overline{y} = \frac{1}{n_y = 9}\sum_{i=1}^{n} y_i = 67.5$$

## Solution (continued):

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\left[\sum (x-\bar{x})^2\right]\left[\sum (y-\bar{y})^2\right]}}$$

$$= \frac{(0.05-2.35)(48.3-67.5)+(0.05-2.35)(51.0-67.5)+\ldots(5-2.35)(87.8-67.5)}{\sqrt{\left[(0.05-2.35)^2+\ldots(5-2.35)^2\right]\left[(48.3-67.5)^2+\ldots(87.8-67.5)^2\right]}}$$

$$= \frac{256.5}{259.6} = 0.988$$

***As r is close to 1, there is also a stronger positive correlation between volume and absorption.***

# Quantitative Analysis – Environmental Risks

```
1     % Environmental Risk
2
3 -   clear all;
4 -   close all;
5 -   clc;
6
7 -   volume = [0.05 0.05 0.05 2.00 2.00 2.00 5.00 5.00 5.00];
8 -   time = [2 2 2 10 10 10 24 24 24];
9 -   absorbtion = [48.3 51.0 54.7 63.2 67.8 66.2 83.6 85.1 87.8];
10
11    % correlation between time and absorbtion
12
13 -  r_tmp=corrcoef(volume,absorbtion);
14 -  r = r_tmp(1,2);
15
16 -  display('Correlation coefficient for Volume vs Absorption:');
17 -  display(r);
18 -  display('suggests a strong positive correlation relationship between volume and absorption.');
19 -  fprintf('\n')
```

**Command Window**

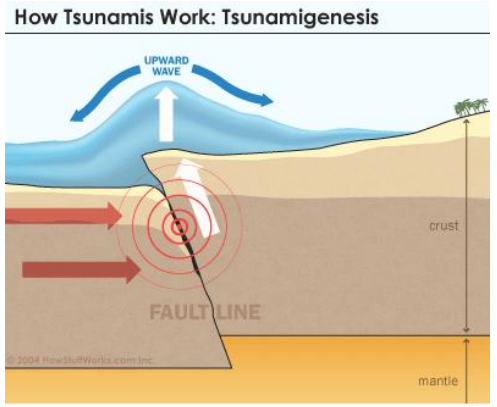
```
Correlation coefficient for Volume vs Absorption:

r =

    0.9882

suggests a strong positive correlation relationship between volume and absorption.
```

## Example (Earthquakes)

In a study of ground motion caused by earthquakes, the peak velocity (in m/s) were recorded for five earthquakes. The results are presented in the following table.

| Velocity | 1.54 | 1.60 | 0.95 | 1.30 | 2.92 |
|---|---|---|---|---|---|
| Acceleration | 7.64 | 8.04 | 8.04 | 6.37 | 5.00 |

Compute the correlation coefficient between peak velocity and peak acceleration. Is any correlation between them?



# Correlation Coefficient

***Solution:***

## Solution:

$$\bar{x} = \frac{1}{n_x = 5}\sum_{i=1}^{n} x_i = 1.66 \qquad \bar{y} = \frac{1}{n_y = 5}\sum_{i=1}^{n} y_i = 7.00$$

| Velocity | 1.54 | 1.60 | 0.95 | 1.30 | 2.92 | $\bar{x}$ = 1.66 |
|---|---|---|---|---|---|---|
| Acceleration | 7.64 | 8.04 | 8.04 | 6.37 | 5.00 | $\bar{y}$ = 7.0 |

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\left[\sum(x-\bar{x})^2\right]\left[\sum(y-\bar{y})^2\right]}} = \frac{-3.17}{3.95} = -0.80$$

***There is reasonably stronger negative correlation between peak velocity and peak acceleration.***

# Quantitative Analysis – Environmental Risks

```
1     % Earthquakes
2
3 -   clear all;
4 -   close all;
5 -   clc;
6
7 -   velocity = [1.54 1.60 0.95 1.30 2.92];
8 -   acceleration = [7.64 8.04 8.04 6.37 5.00];
9
10    % correlation
11
12 -  r_tmp=corrcoef(velocity,acceleration);
13 -  r = r_tmp(1,2);
14
15 -  display('Correlation coefficient for Peak Acceleration vs Peak Velocity:');
16 -  display(r);
17 -  display('suggests a strong negative correlation between peak velocity and peak acceleration.');
18 -  fprintf('\n')
```

**Command Window**

```
Correlation coefficient for Peak Acceleration vs Peak Velocity:

r =

   -0.8028

suggests a strong negative correlation between peak velocity and peak acceleration.
```

# Significance Test for Correlation

- Hypotheses

$H_0$: $\rho = 0$ (no correlation)

$H_1$: $\rho \neq 0$ (correlation exists)

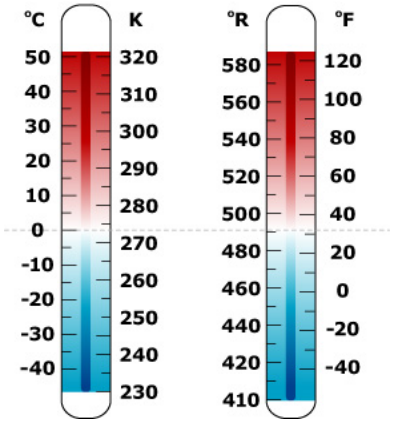
- Test statistic
  - $$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$ (with n – 2 degrees of freedom)

***With consideration of sample size and standard deviation***

# Significance Test for Correlation Example

***Is any correlation between temperature and number of bushfires?***

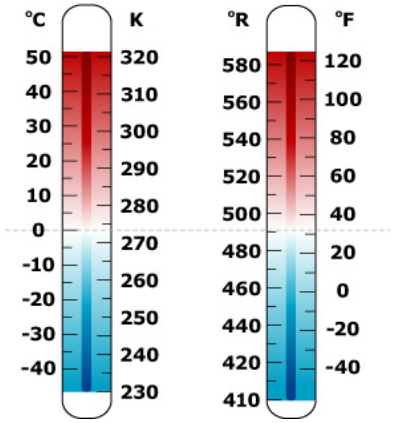
| Temperature | Number of bushfires |
|---|---|
| y | x |
| 35ºC | 8 |
| 49ºC | 9 |
| 27ºC | 7 |
| 33ºC | 6 |
| 60ºC | 13 |
| 21ºC | 7 |
| 45ºC | 11 |
| 51ºC | 12 |
| **Σ=321**ºC | **Σ=73** |




$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\left[\sum (x-\bar{x})^2\right]\left[\sum (y-\bar{y})^2\right]}} = 0.886$$

# Significance Test for Correlation Example

**Is there evidence of a linear relationship between temperature and number of bushfires at the 0.05 level of significance?**



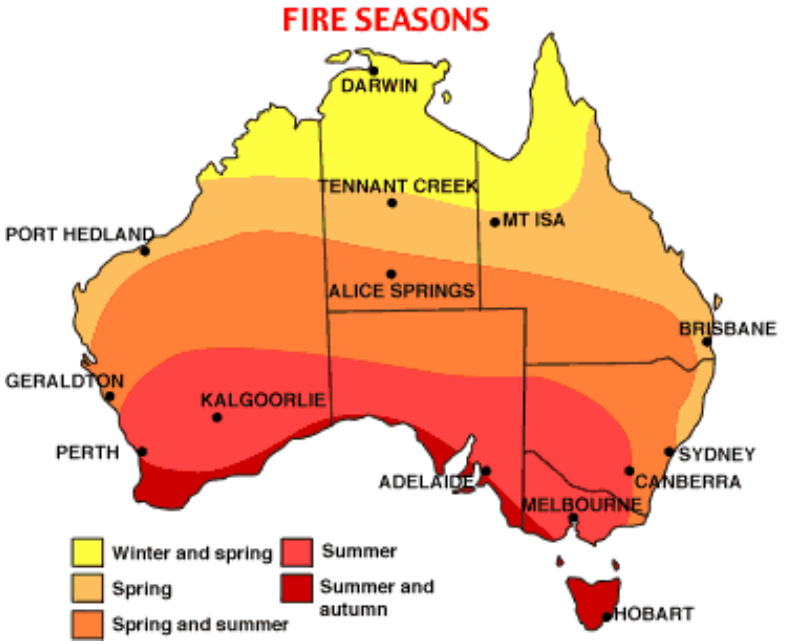

# Significance Test for Correlation Example

$H_0$: $\rho = 0$ (No correlation)

$H_1$: $\rho \neq 0$ (correlation exists)

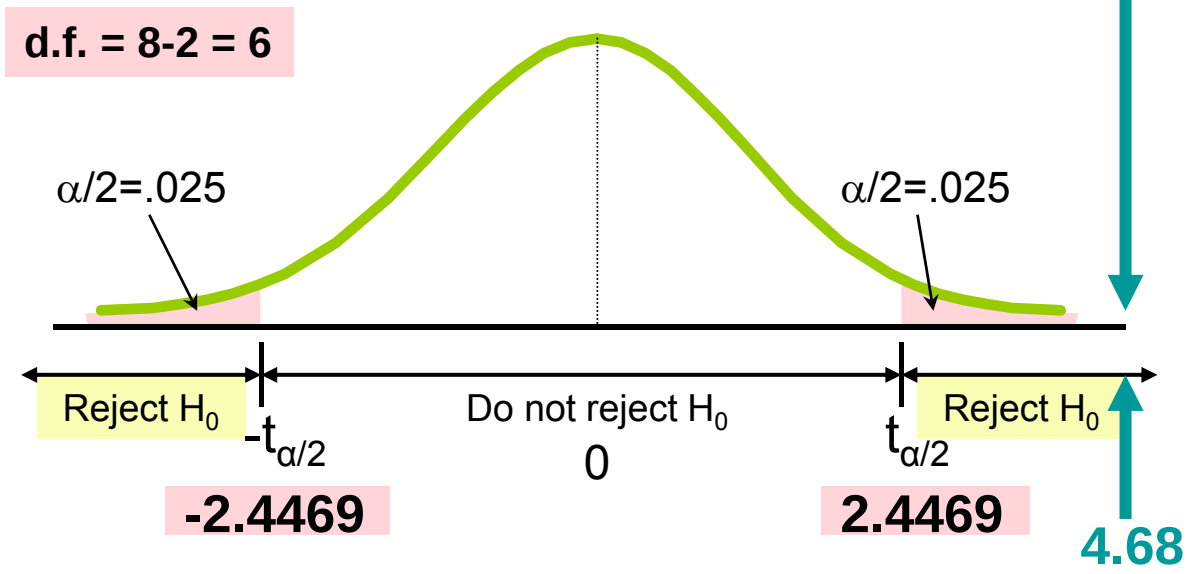
$\alpha = .05$ , df = 8 - 2 = 6

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.886}{\sqrt{\frac{1-.886^2}{8-2}}} = 4.68$$



# Significance Test for Correlation Example

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.886}{\sqrt{\frac{1-.886^2}{8-2}}} = 4.68$$




**Decision:**
Reject $H_0$

**Conclusion:**
There is evidence of a linear relationship at the 5% level of significance

# Significance Test for Correlation Example

```
% Temperature vs Bushfires
% H0: rho = 0 (no correlation)
% HA: rho ~= 0 (correlation exits)
% p < alpha, reject H0

clear all;
close all;
clc;

bushfires = [8 9 7 6 13 7 11 12];           %x values
temperature = [35 49 27 33 60 21 45 51];    %y values

alpha = 0.05;

% correlation
r_tmp=corrcoef(bushfires,temperature);
r = r_tmp(1,2);

display(r);

% Significance test
len = length(bushfires);
df = len-2;
display(df);

t = r/sqrt((1-r^2)/df);
display(t);

display('Two tailed test');
p = 2*tcdf(-abs(t),df);

if p > alpha
    display(p);
    display(alpha);
    display('Since p > alpha,  we do not rejectH_0.');
    display('There is no evidence of a linear relationship at the 5% level of significance.');
else
    display(p);
    display(alpha);
    display('Since p < alpha, we reject H_0.');
    display('There is evidence of a linear relationship at the 5% level of significance');
end
```

# Significance Test for Correlation Example

**Command Window**

```
r =

    0.8862

df =

     6

t =

    4.6861

Two tailed test

p =

    0.0034

alpha =

    0.0500

Since p < alpha, we reject H_0.
There is evidence of a linear relationship at the 5% level of significance
```

# The Least-Squares Line

- When two variables have a linear relationship, the scatterplot tends to be clustered around a line known as **the least-squares line**

## Example:

**An accelerated test, steel structures are operated under extreme conditions until failure**

***Two variables:***
**Lifetime vs Temperature**

| Temperature (°C) | Lifetime (hours) |
|---|---|
| 40 | 851 |
| 45 | 635 |
| 50 | 764 |
| 55 | 708 |
| 60 | 469 |
| 65 | 661 |
| 70 | 586 |
| 75 | 371 |
| 80 | 337 |
| 85 | 245 |
| 90 | 129 |
| 95 | 158 |

# The Least-Squares Line

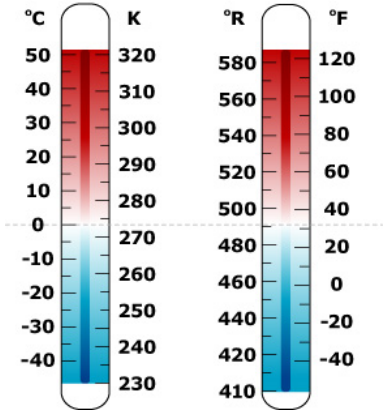
**Dependent variable:** the variable we wish to explain

**Independent variable:** the variable used to explain the dependent variable

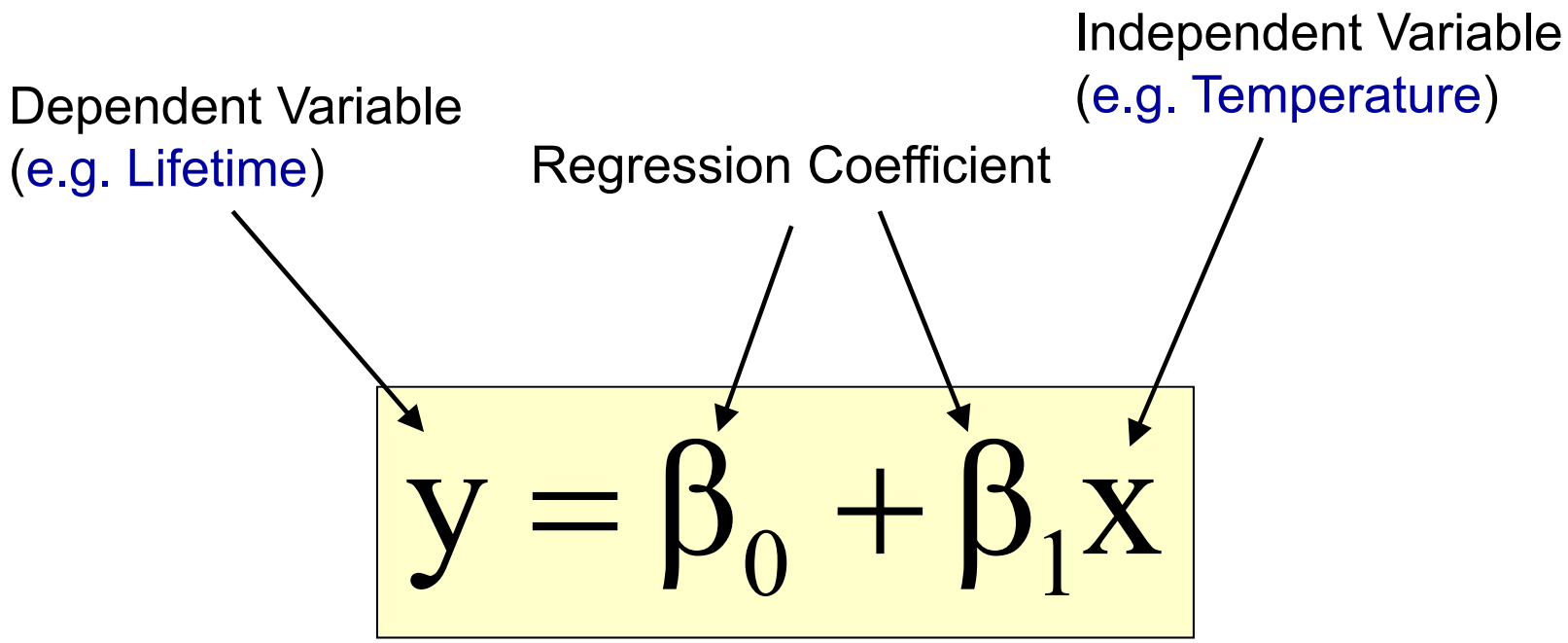
**Lifetime vs Temperature**

**Dependent variable**

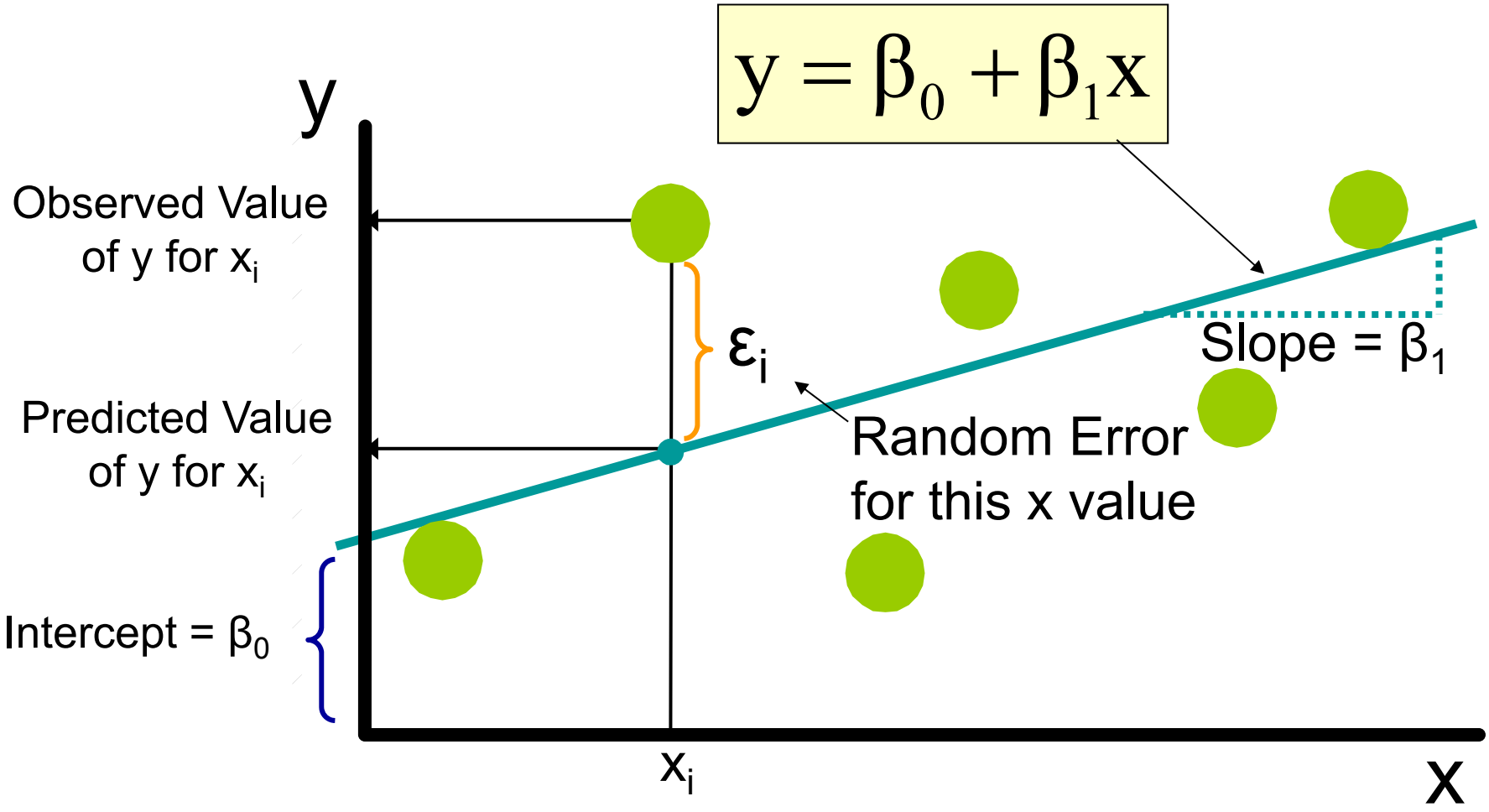
**Independent variable**



# The Least-Squares Line



# The Least-Squares Line



# The Least-Squares Line

$$y = \beta_0 + \beta_1 x$$

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

# The Least-Squares Line

**Example:**

**An accelerated test, steel structures are operated under extreme conditions until failure**

**(a) Compute the least-squares line for predicting life-time from temperature.**

**(b) Predict the lifetime for a temperature of 73ºC.**

| Temperature (ºC) | Lifetime (hours) |
|---|---|
| 40 | 851 |
| 45 | 635 |
| 50 | 764 |
| 55 | 708 |
| 60 | 469 |
| 65 | 661 |
| 70 | 586 |
| 75 | 371 |
| 80 | 337 |
| 85 | 245 |
| 90 | 129 |
| 95 | 158 |

# The Least-Squares Line

***Solution:***

# The Least-Squares Line

## Solution (a):

| Temperature (°C) | Lifetime (hours) |
|---|---|
| 40 | 851 |
| 45 | 635 |
| 50 | 764 |
| 55 | 708 |
| 60 | 469 |
| 65 | 661 |
| 70 | 586 |
| 75 | 371 |
| 80 | 337 |
| 85 | 245 |
| 90 | 129 |
| 95 | 158 |
| $\overline{X} = 70$ | $\overline{y} = 460.3$ |

$$\overline{x} = \frac{1}{n_x = 12}\sum_{i=1}^{n} x_i = 70 \qquad \overline{y} = \frac{1}{n_y = 12}\sum_{i=1}^{n} y_i = 460.3$$

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = -12.6$$

$$\beta_0 = \overline{y} - \beta_1 \overline{x} = 460.3 - (-12.6) \times 70 = 1343.7$$

# The Least-Squares Line

***Solution (a) continued:***

***The* least-squares line is:**

$$y = \beta_0 + \beta_1 x = -12.6x + 1343.7$$

***Solution (b):***

**x = 73**

**y = -12.6x73+1343.7=422.4**

The lifetime for a temperature of 73ºC is 422.4 hours

## *Using cftool:*

```
% The lease squares line
% Lifetime vs Temperature

clear all;
close all;
clc;

temperature = [40 45 50 55 60 65 70 75 80 85 90 95];
lifetime = [851 635 764 708 469 661 586 371 337 245 129 158];

% Least squares line

cftool;
```

# The Least-Squares Line

## *Using cftool:*

Least square line

Fit name: Least square line

X data: temperature

Y data: lifetime

Z data: (none)

Weights: (none)

Polynomial

Degree: 1

Robust: Off

Center and scale

Fit Options...

Auto fit

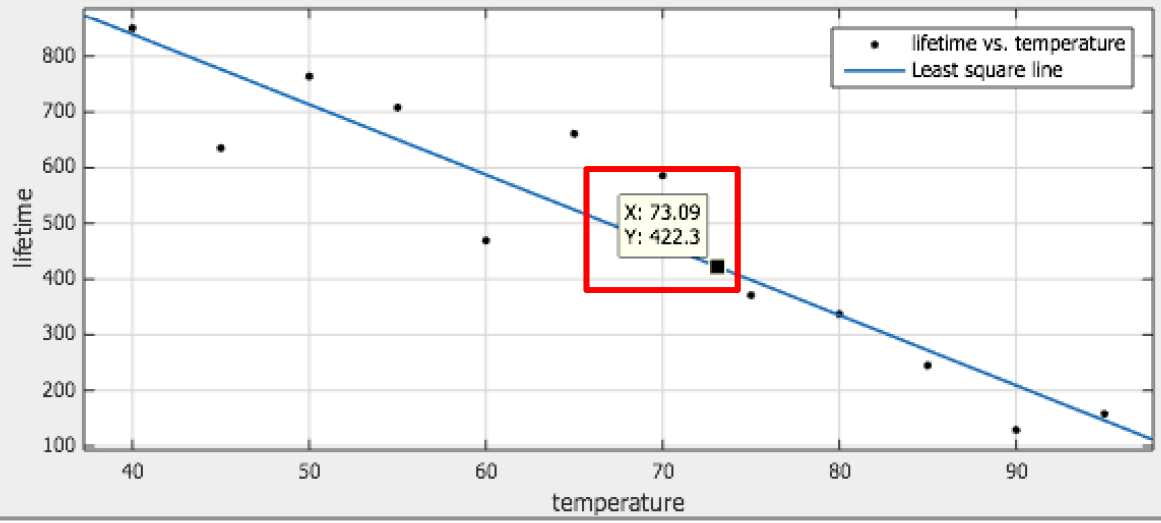
Fit

Stop

Results

Linear model Poly1:
f(x) = p1*x + p2
Coefficients (with 95% confidence b
p1 = -12.61 (-15.99, -9.23
p2 = 1344 (1109, 1580)

Goodness of fit:
SSE: 8.231e+04
R-square: 0.8735
Adjusted R-square: 0.8609
RMSE: 90.73



## Example (Fuel economy)

Inertial weight (in tons) and fuel economy (in mi/gal) were measured for a sample of seven diesel trucks. The results are presented in the following table

| Weight | Mileage |
|---|---|
| 8.00 | 7.69 |
| 24.50 | 4.97 |
| 27.00 | 4.56 |
| 14.50 | 6.49 |
| 28.50 | 4.34 |
| 12.75 | 6.24 |
| 21.25 | 4.45 |

(a) Compute the least-squares line for predicting mileage from weight.

(b) Predict the mileage for trucks with a weight of 15 tons.

# The Least-Squares Line

***Solution:***

# The Least-Squares Line

## Solution (a):

| Weight | Mileage |
|---|---|
| 8.00 | 7.69 |
| 24.50 | 4.97 |
| 27.00 | 4.56 |
| 14.50 | 6.49 |
| 28.50 | 4.34 |
| 12.75 | 6.24 |
| 21.25 | 4.45 |
| $\overline{X}$ = 19.5 | $\overline{y}$ = 5.53 |

$$\bar{x} = \frac{1}{n_x = 7}\sum_{i=1}^{n} x_i = 19.5$$

$$\bar{y} = \frac{1}{n_y = 7}\sum_{i=1}^{n} y_i = 5.53$$

$$\beta_1 = \frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\sum\left(x-\bar{x}\right)^2} = -0.155$$

$$\beta_0 = 5.53 - \left(-0.155\right)\left(19.5\right) = 8.559$$

**The least-squares line is:**

$$y = \beta_0 + \beta_1 x = -0.155x + 8.559$$

***Solution (b):***

**x = 15**

**y = -0.155x15+8.559=6.23**

the mileage for trucks with a weight of 15 tons is 6.23 mi/gal

## *Using cftool:*

```
% The lease squares line
% Fuel economy

clear all;
close all;
clc;

weight = [8 24.5 27 14.5 28.5 12.75 21.25];       %x values
mileage = [7.69 4.97 4.56 6.49 4.34 6.24 4.45];   %y values

% Least squares line

cftool;
```

# The Least-Squares Line

## *Using cftool:*

Least square line

Fit name: Least square line

X data: weight

Y data: mileage

Z data: (none)

Weights: (none)

Polynomial

Degree: 1

Robust: Off

Center and scale

Fit Options...

Auto fit

Fit

Stop

Results

Linear model Poly1:
f(x) = p1*x + p2
Coefficients (with 95% confidence b
p1 = -0.1551 (-0.2171, -0.(
p2 = 8.559 (7.27, 9.848)

Goodness of fit:
SSE: 1.069
R-square: 0.8923
Adjusted R-square: 0.8708
RMSE: 0.4624