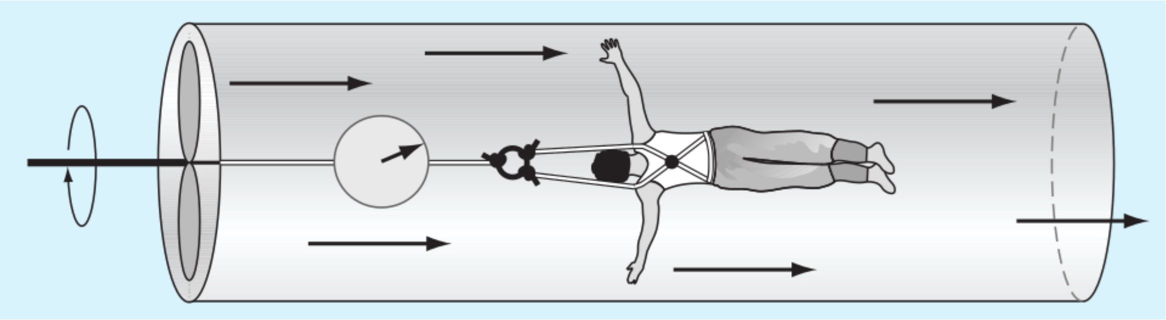
7 Regression

- Regression is a method to find "best fit" curves to experimental data.

- If we take measurements of an experiment we can plot discrete data points of our observations.

Wind tunnel measuring force on a body vs. wind velocity

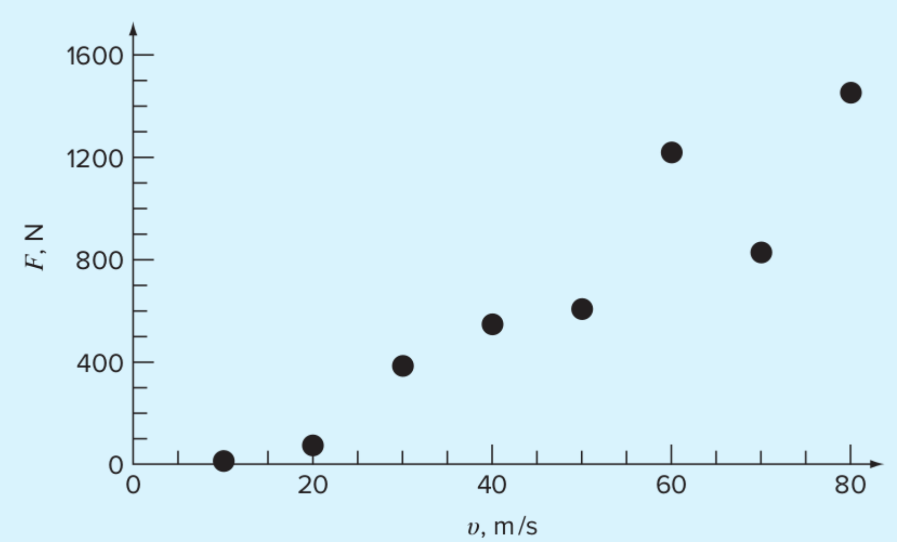


v, m/s	10	20	30	40	50	60	70	80
F, N	25	70	380	550	610	1220	830	1450

- Since we cannot take measurements for every value of wind velocity but we would like to know what force will be exerted for any wind velocity we encounter, we formulate a relationship between Force and Wind Velocity.

$$F_U = c_d v^2$$

Drag coefficient



- The model equation above comes from fitting experimental data to the "best fitting" curve so that we can approximate future values of force for any velocity.

7.1 Linear Least-Squares Regression

- If we suspect that the variables have a linear relationship we can try to fit a straight line to the data:

$$y = a_0 + a_1 x + e$$

y-intercept (a_0), **Slope** (a_1), **Error (residual)** (e)

- The error is the difference between the straight line approximation and the true data point.

$$e = y - a_0 - a_1 x$$

- Each data point has an associated error. If we try to minimise the total error it can lead to a line which is not a best fit:

For n data points

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

Total error

Or go to www.pollev.com/jsands601

Which data point has the largest error for the fitted equation?

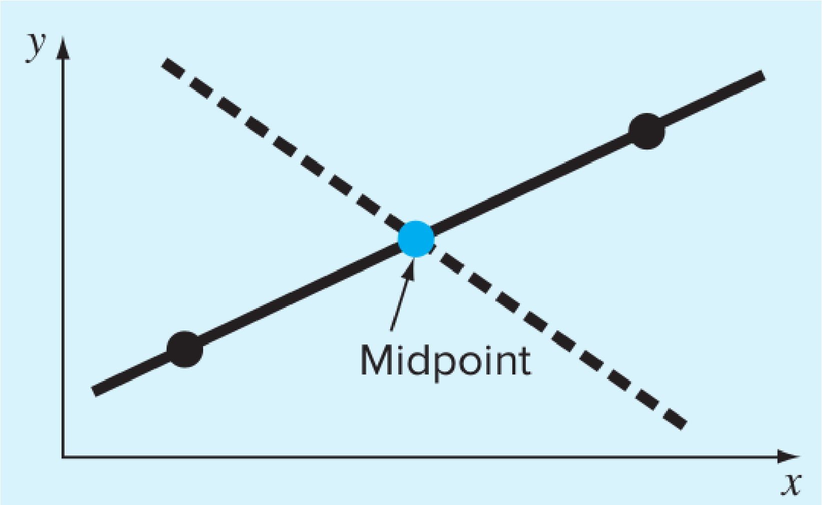
$$y = 3x$$

x	y	e
-1	-3	
0	1	
1	4	
2	4	
3	10	

- $x = -1$
- $x = 0$
- $x = 1$
- $x = 2$
- $x = 3$

T(0

- For the data points on the right hand side, the best fit is the straight line joining both points.



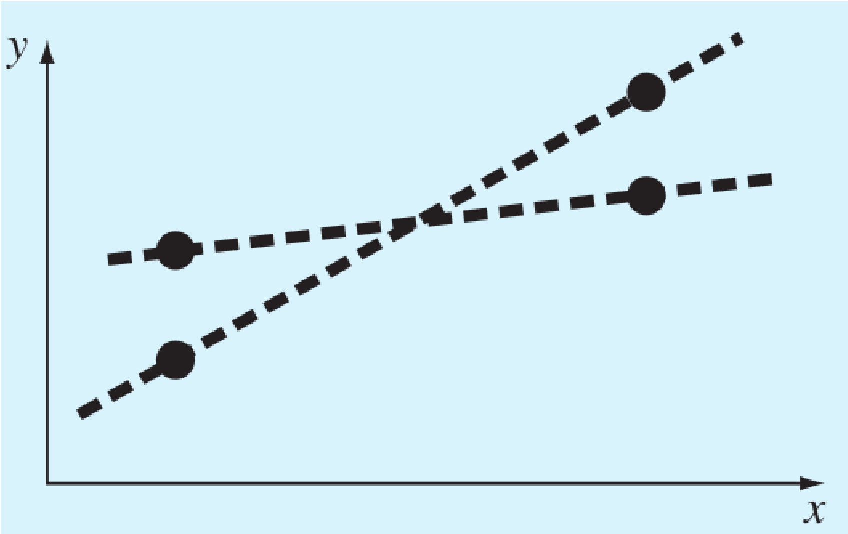
- However the dashed line also produces a total error of 0 since positive and negative error cancel out.

- Therefore using total error does not produce a unique solution to the best fit problem.

- We can try to get around this problem by taking the absolute value of the total error:

$$\sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |y_i - a_0 - a_1 x_i|$$

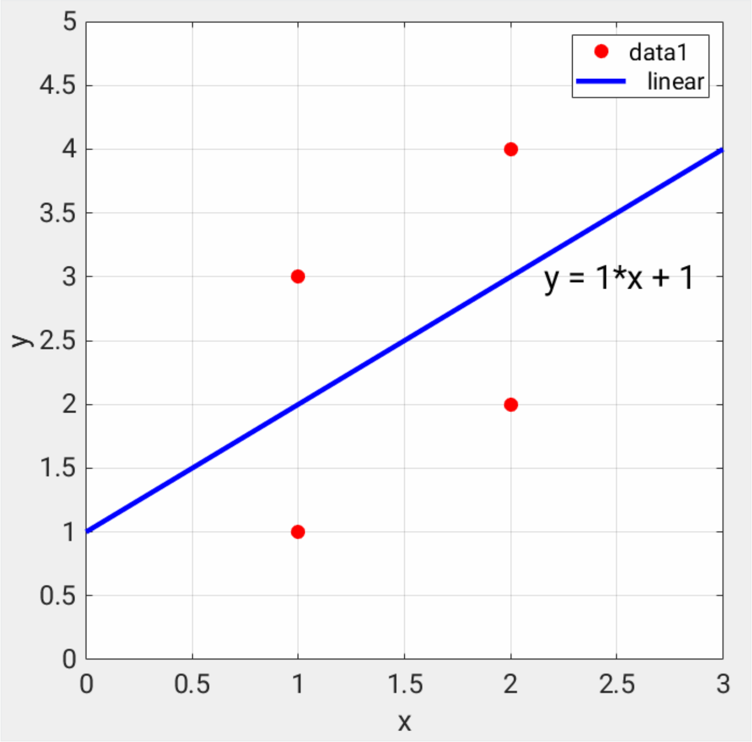
- However this also results in a solution that is not unique. For the data below any straight line in between the dashed lines will minimise the absolute value of the error.



EXAMPLE 1 The best fitting linear line for the following data is $y = x + 1$. However, using total absolute error produces non-unique results.

$$y_1 = x + 1 \qquad y_2 = x + 2$$

x	y	$\lvert y - y_1 \rvert$	$\lvert y - y_2 \rvert$
1	1	1	2
2	2	1	2
1	3	1	0
2	4	1	0
	TOTAL	4	4



Same (minimum) error
Not a unique solution

- To get a unique solution we can use the square of the error as our criterion for minimising the error:

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

- For this reason we call it **least squares** regression.

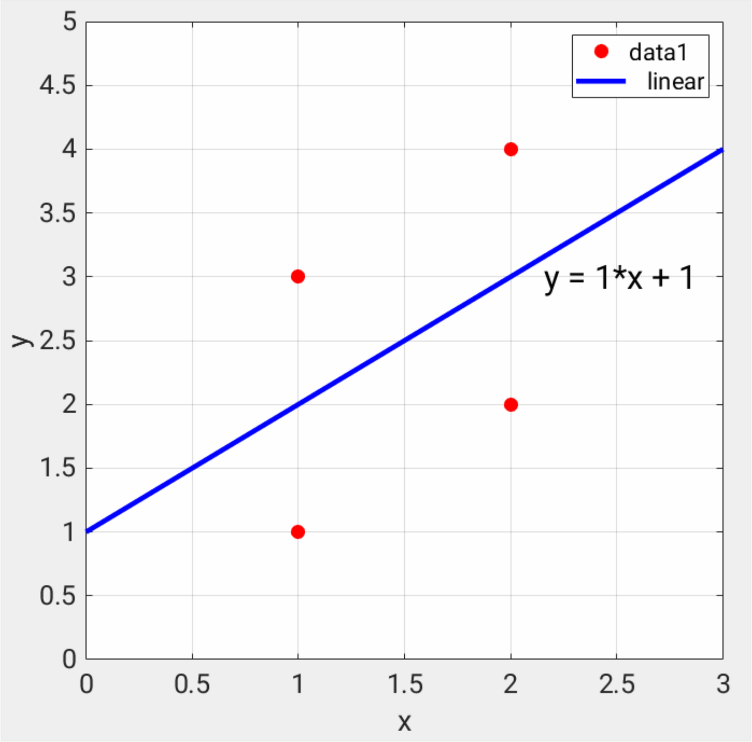
- In particular this is for linear graphs so it is in fact **linear least squares**.

- The reason why this produces a unique solution and the others do not become apparent once we derive the formulas for the coefficients a_0 and a_1.

EXAMPLE 2 Using the square of the error produces a unique solution.

$$y_1 = x + 1 \qquad y_2 = x + 2$$

x	y	$(y - y_1)^2$	$(y - y_2)^2$
1	1	1	4
2	2	1	4
1	3	1	0
2	4	1	0
	TOTAL	4	8



Unique solution for minimum error

- To derive the coefficients we minimise the total least squares error function using calculus techniques.

- The variables are a_0 and a_1 so to minimise we must take partial derivatives with respect to these variables and find the critical points.

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

$$\frac{\partial S_r}{\partial a_1} = -2 \sum [(y_i - a_0 - a_1 x_i) x_i]$$

$$0 = \sum y_i - \sum a_0 - \sum a_1 x_i$$

$$0 = \sum x_i y_i - \sum a_0 x_i - \sum a_1 x_i^2$$

$$\longrightarrow \quad n a_0 + \left(\sum x_i\right) a_1 = \sum y_i$$

$$\longrightarrow \quad \left(\sum x_i\right) a_0 + \left(\sum x_i^2\right) a_1 = \sum x_i y_i$$

- Solving these equations simultaneously gives:

$$\boxed{\begin{aligned} a_1 &= \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \\ a_0 &= \bar{y} - a_1 \bar{x} \end{aligned}}$$

Mean of *y* **Mean of *x***

- Note that the critical point is unique, and that it must be a minimum since there is no maximum error.

EXAMPLE 3 Fit a straight line to the following data for drag coefficients.

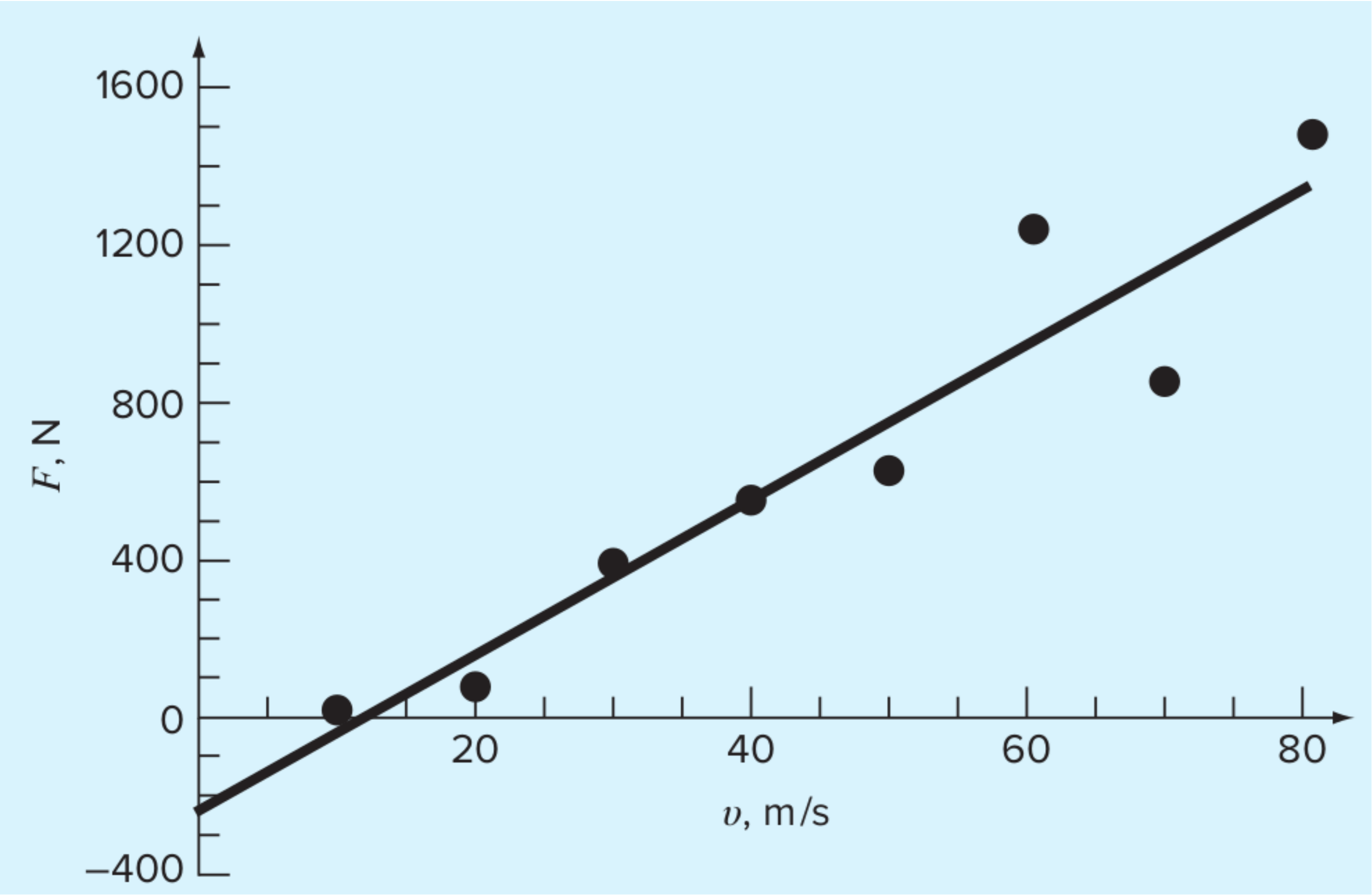
i	x_i	y_i	x_i^2	$x_i y_i$
1	10	25	100	250
2	20	70	400	1,400
3	30	380	900	11,400
4	40	550	1,600	22,000
5	50	610	2,500	30,500
6	60	1,220	3,600	73,200
7	70	830	4,900	58,100
8	80	1,450	6,400	116,000
Σ	360	5,135	20,400	312,850

$$\bar{x} = \frac{360}{8} = 45 \qquad \bar{y} = \frac{5{,}135}{8} = 641.875$$

$$a_1 = \frac{8(312{,}850) - 360(5{,}135)}{8(20{,}400) - (360)^2} = 19.47024$$

$$a_0 = 641.875 - 19.47024(45) = -234.2857$$

$$F = -234.2857 + 19.47024v$$



EXAMPLE 4 Fit a straight line to the data.

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

x_i	y_i
-1	-3
0	1
1	4
2	4
3	10

Correlation Coefficient

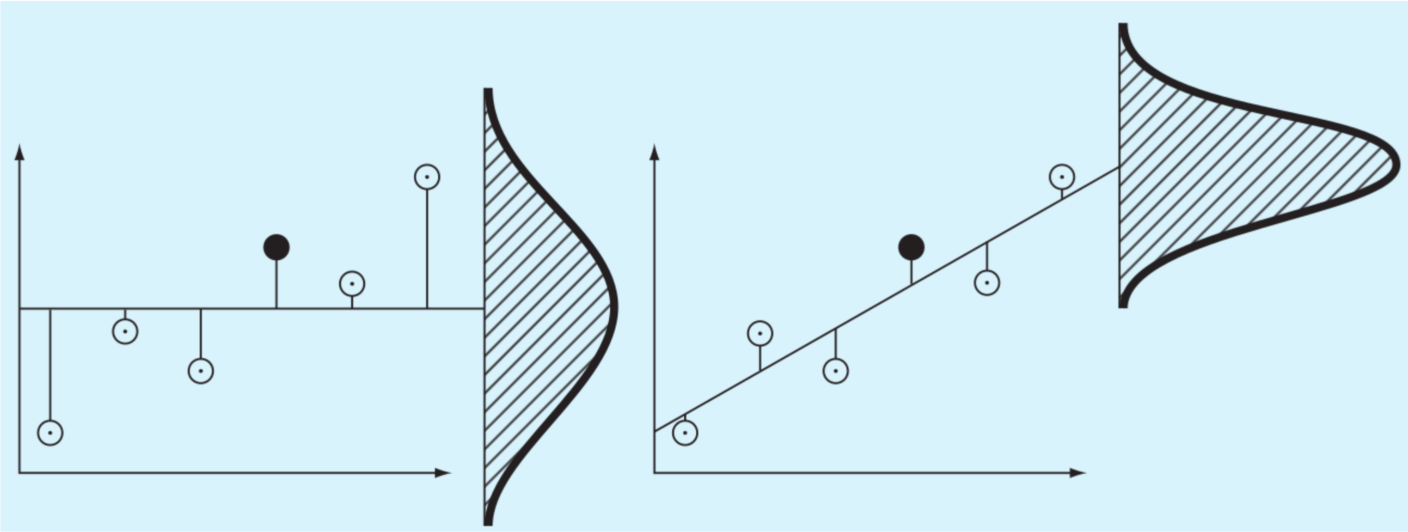
- We measure how good our linear regression fit is by comparing it with the spread about the mean value of the data:

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

Spread about mean

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

Spread about straight line fit



- Evaluating about the mean gives the largest spread. If our straight line approximation is better then the data will be less spread out.

- We therefore define the **correlation coefficient**, **r**, to be:

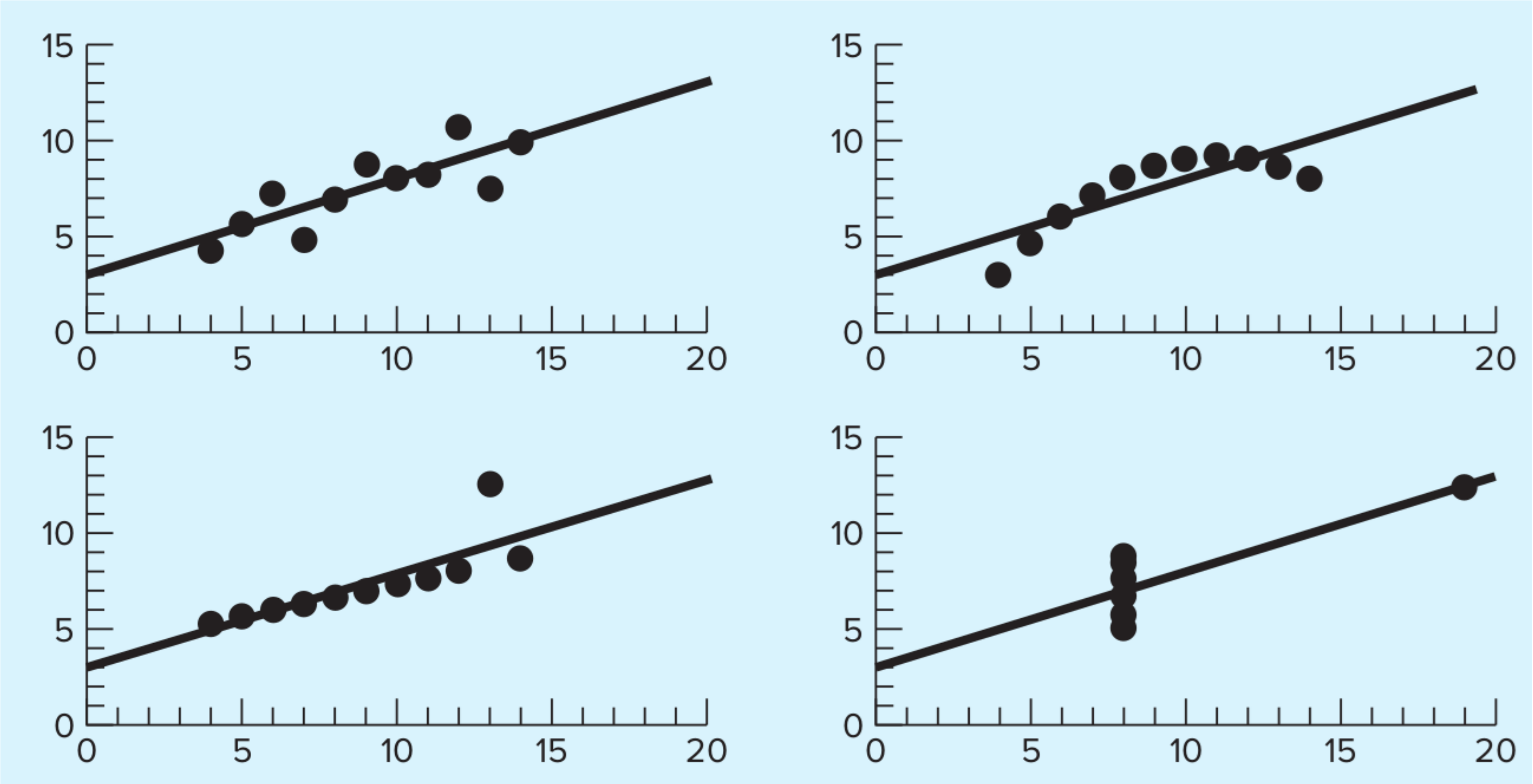
$$r^2 = \frac{S_t - S_r}{S_t}$$ **(Relative measure)**

$$\longrightarrow \quad r = \frac{n\sum(x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

- When $r = 0$ there is no improvement beyond the mean. When $r = 1$ the total S_r is 0 and we have a perfect fit.

- **Warning:** Even if r is close to 1 it does not necessarily mean the straight line is a good fit.

EXAMPLE 5 The following 4 data sets have $r = 0.67$ with the same best fit line of $y = 3 + 0.5x$.



- So remember to plot your data to check that it looks like a good fit.

- The correlation coefficient can then be used to determine quantitatively just how good that fit is, only after establishing that the line is the right shape in the first place.

7.2 Linearisation of Nonlinear Relationships

- When fitting experimental data sometimes our best option is to assume a certain form of nonlinear relationship and fit the parameters to that equation.

- Experiments will determine how reasonable our equations are for the parameters we have fitted.

Exponential Model

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

Power Equation

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

Saturation-Growth Equation

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

- The idea is to linearise the equations and perform linear regression analysis as before in order to obtain the unknown coefficients of the equations (α, β).

- The linearised versions are given below:

Exponential Model

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

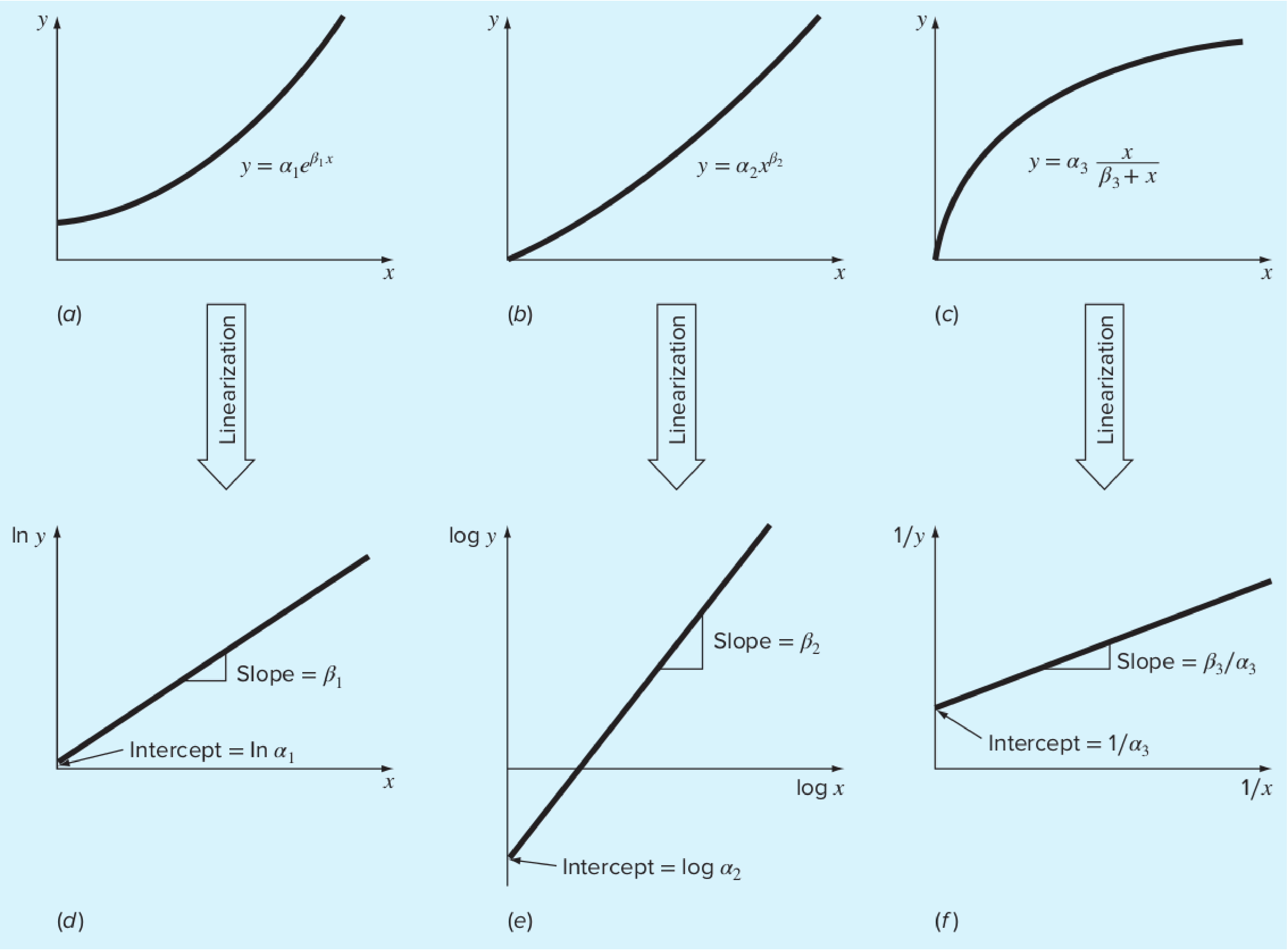
Power Equation

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

Can take any base but base-10 is standard convention

Saturation-Growth Equation

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



$y = \alpha_1 e^{\beta_1 x}$
(a)
Linearization
$y = \alpha_2 x^{\beta_2}$
(b)
Linearization
$y = \alpha_3 \frac{x}{\beta_3 + x}$
(c)
Linearization
ln y
Slope = β_1
Intercept = ln α_1
(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 6 Fit the data from **Example 3** to a power equation.

i	x_i	y_i	$\log x_i$	$\log y_i$	$(\log x_i)^2$	$\log x_i \log y_i$
1	10	25	1.000	1.398	1.000	1.398
2	20	70	1.301	1.845	1.693	2.401
3	30	380	1.477	2.580	2.182	3.811
4	40	550	1.602	2.740	2.567	4.390
5	50	610	1.699	2.785	2.886	4.732
6	60	1220	1.778	3.086	3.162	5.488
7	70	830	1.845	2.919	3.404	5.386
8	80	1450	1.903	3.161	3.622	6.016
Σ			12.606	20.515	20.516	33.622

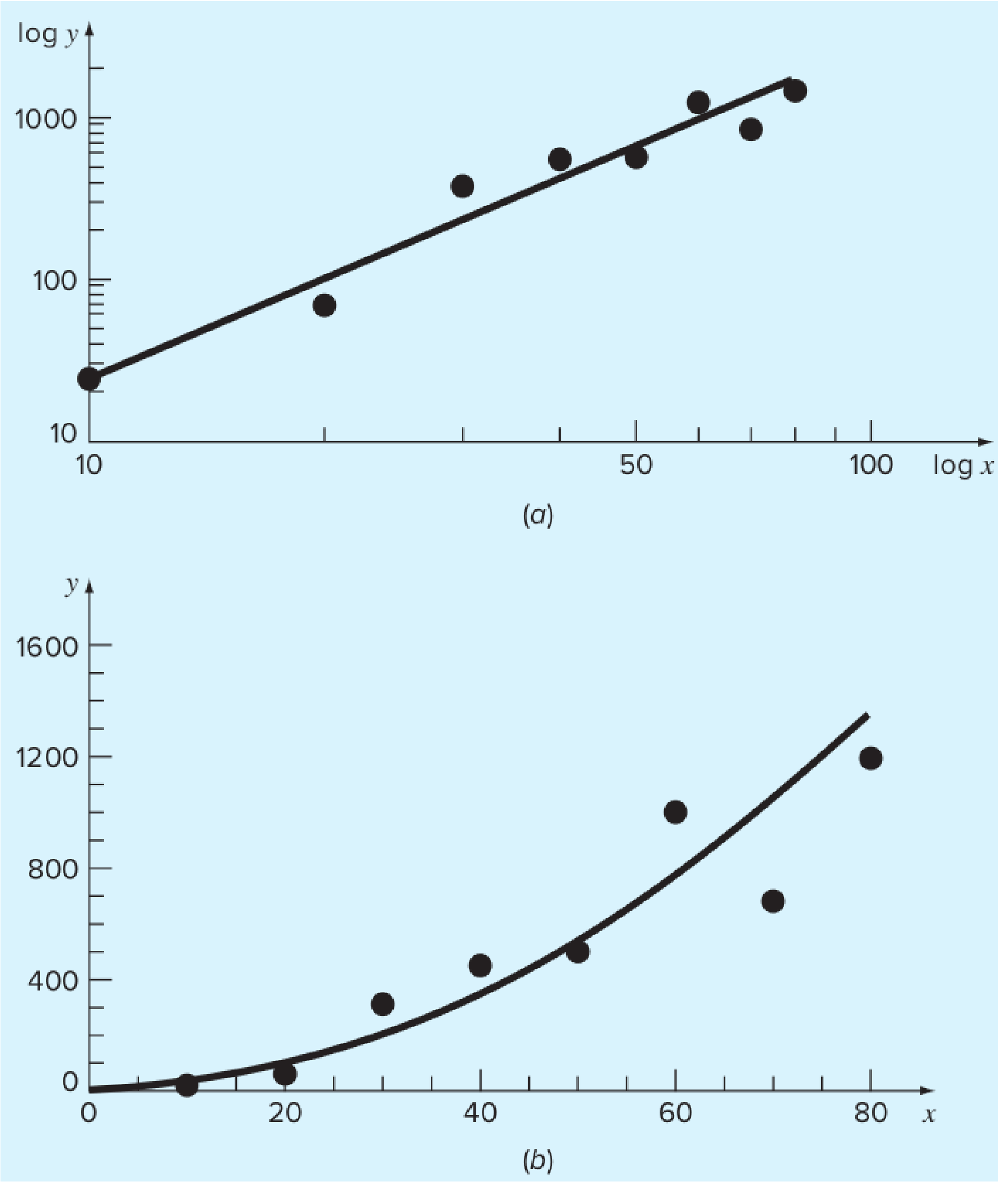
Log-means:

$$\bar{x} = \frac{12.606}{8} = 1.5757 \qquad \bar{y} = \frac{20.515}{8} = 2.5644$$

Linear regression coefficients for log variables

$$a_1 = \frac{8(33.622) - 12.606(20.515)}{8(20.516) - (12.606)^2} = 1.9842$$

$$a_0 = 2.5644 - 1.9842(1.5757) = -0.5620$$



$$\log y = -0.5620 + 1.9842 \log x$$

$$\alpha_2 = 10^{-0.5620} = 0.2741$$

$$\beta_2 = 1.9842$$

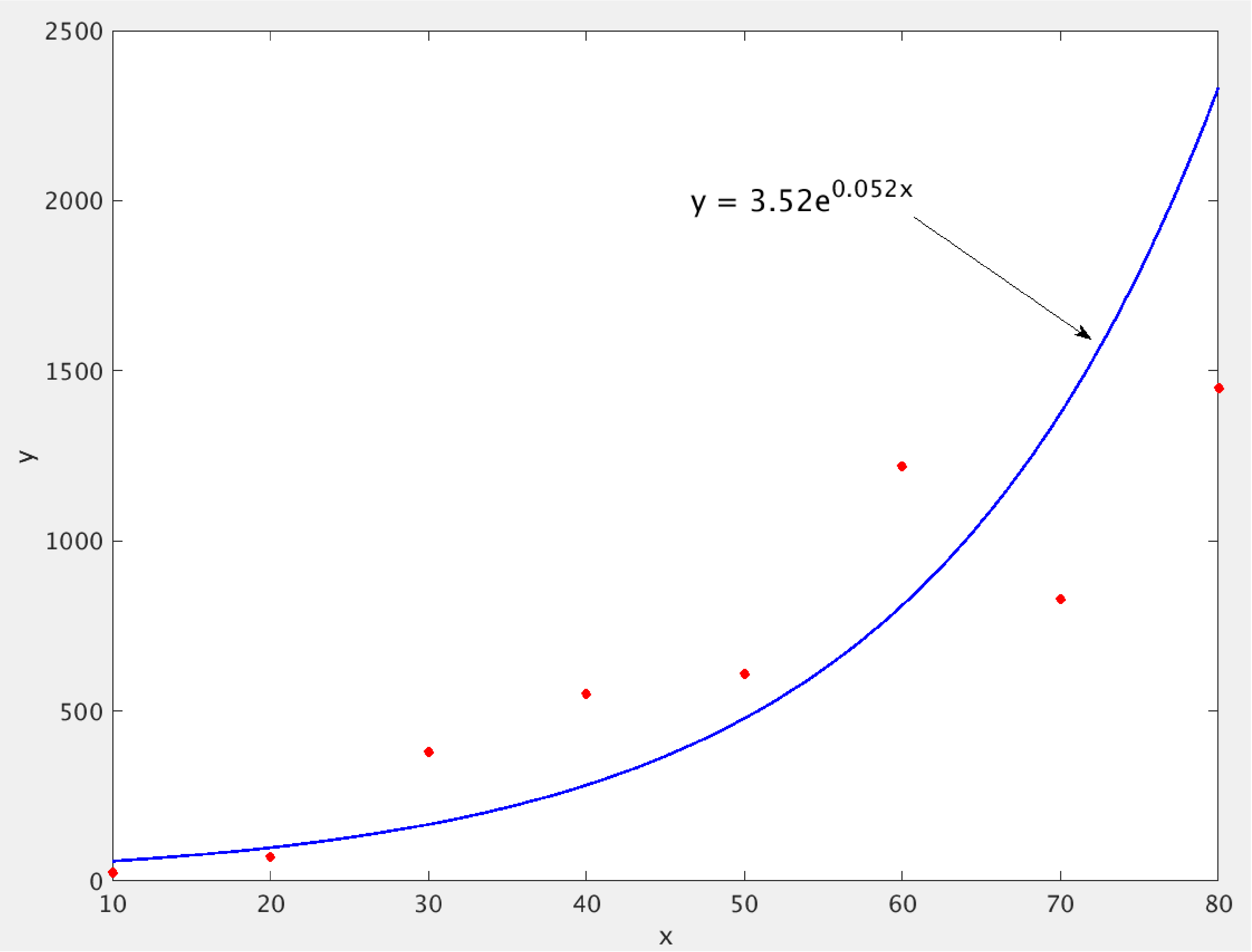
$$F = 0.2741 v^{1.9842}$$

- So which model is better? Linear fit or power equation fit?
- The answer must come from an engineering knowledge of the system you are studying.
- From first principle fluid mechanics calculations we can show under certain idealised conditions that the drag force on an object is proportional to the velocity squared.
- This indicates that our power model equation may be a more realistic mathematical model for our system.

EXAMPLE 7 Fit an exponential model to the data using linearisation.

$$y = a_0 + a_1 x: \quad a_0 = \bar{y} - a_1 \bar{x} \qquad a_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

i	x_i	y_i
1	10	25
2	20	70
3	30	380
4	40	550
5	50	610
6	60	1,220
7	70	830
8	80	1,450


2500
2000
1500
1000
500
0
10
20
30
40
50
60
70
80
y
x
y = 3.52e^0.052x

7.3 Built-In Matlab Functions

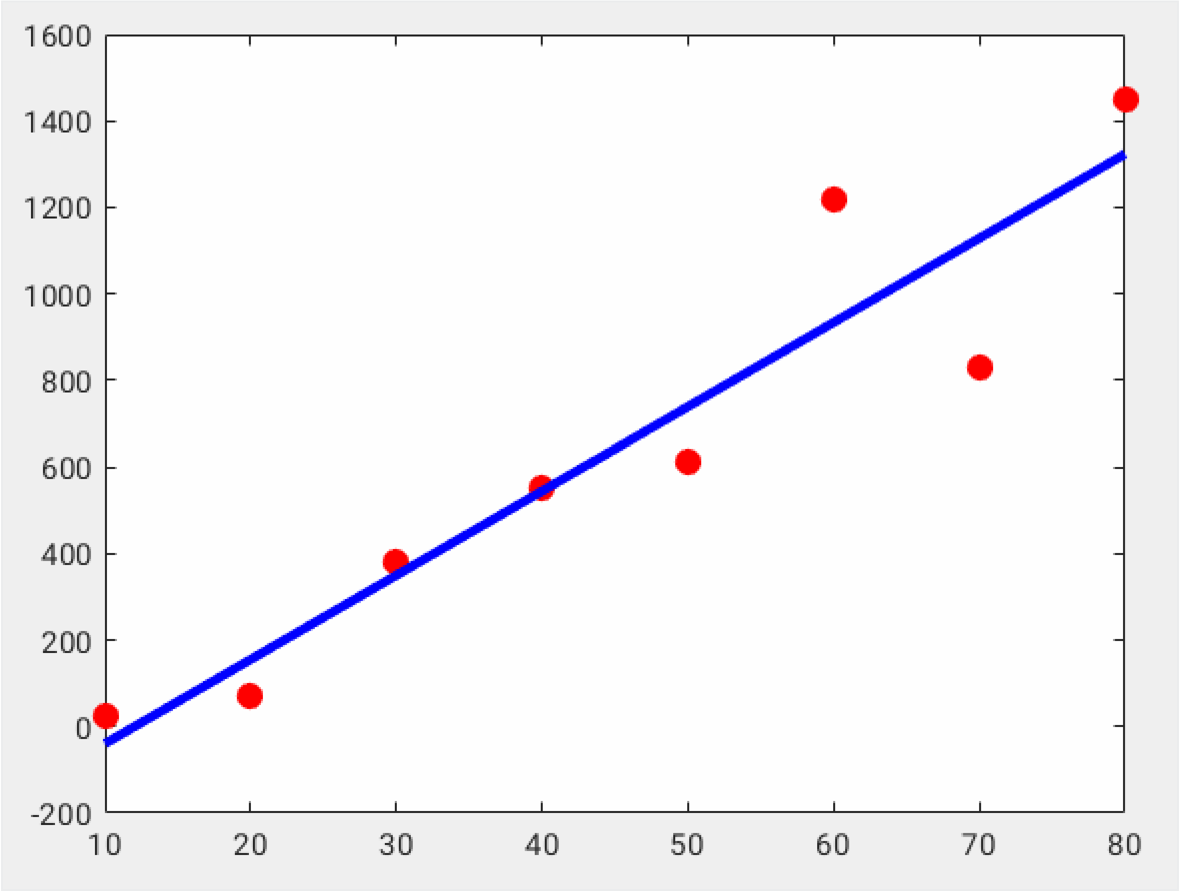
- The **polyfit** and **polyval** functions in Matlab can be used to do linear least-squares regression as follows:

```
>> x = [10 20 30 40 50 60 70 80];
>> y = [25 70 380 550 610 1220 830 1450];
>> a = polyfit(x,y,1)

a =

    19.4702 -234.2857
```

slope **_y_-intercept**

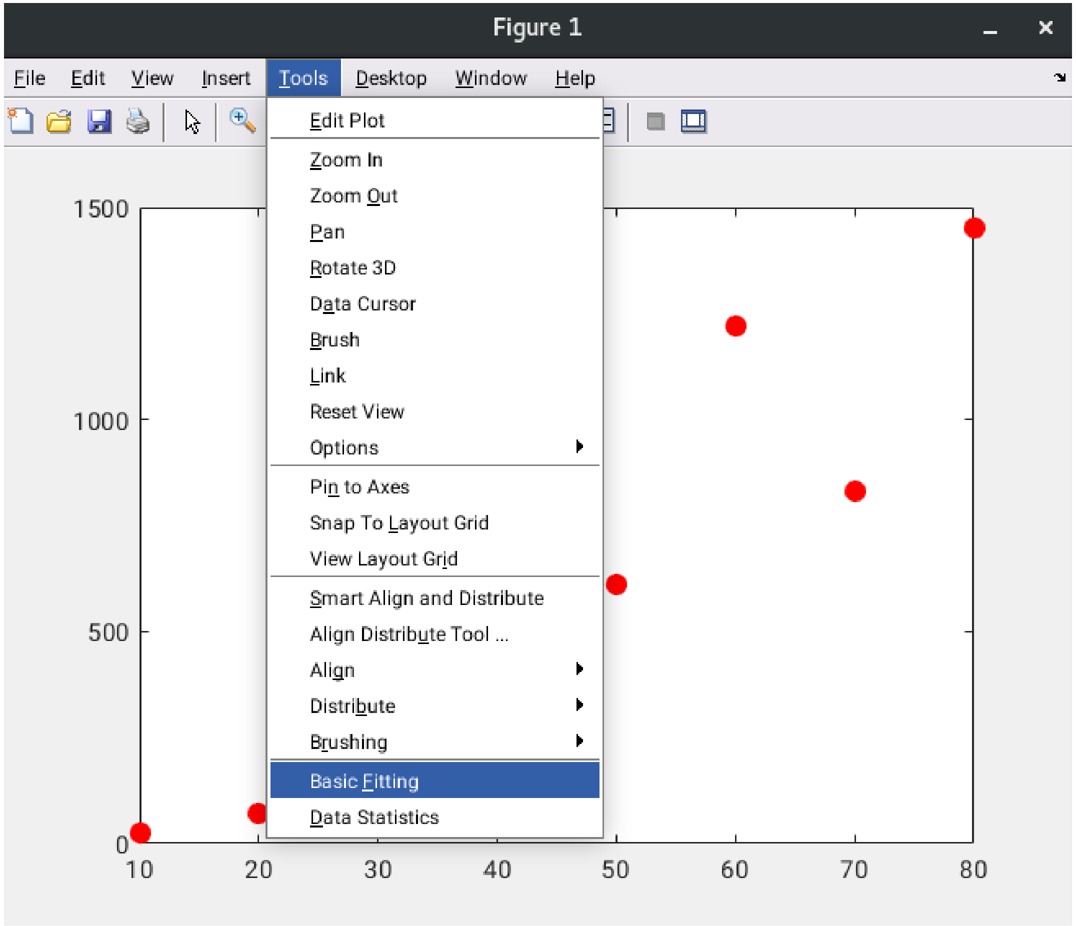
$$y = -234.2857 + 19.4702x$$

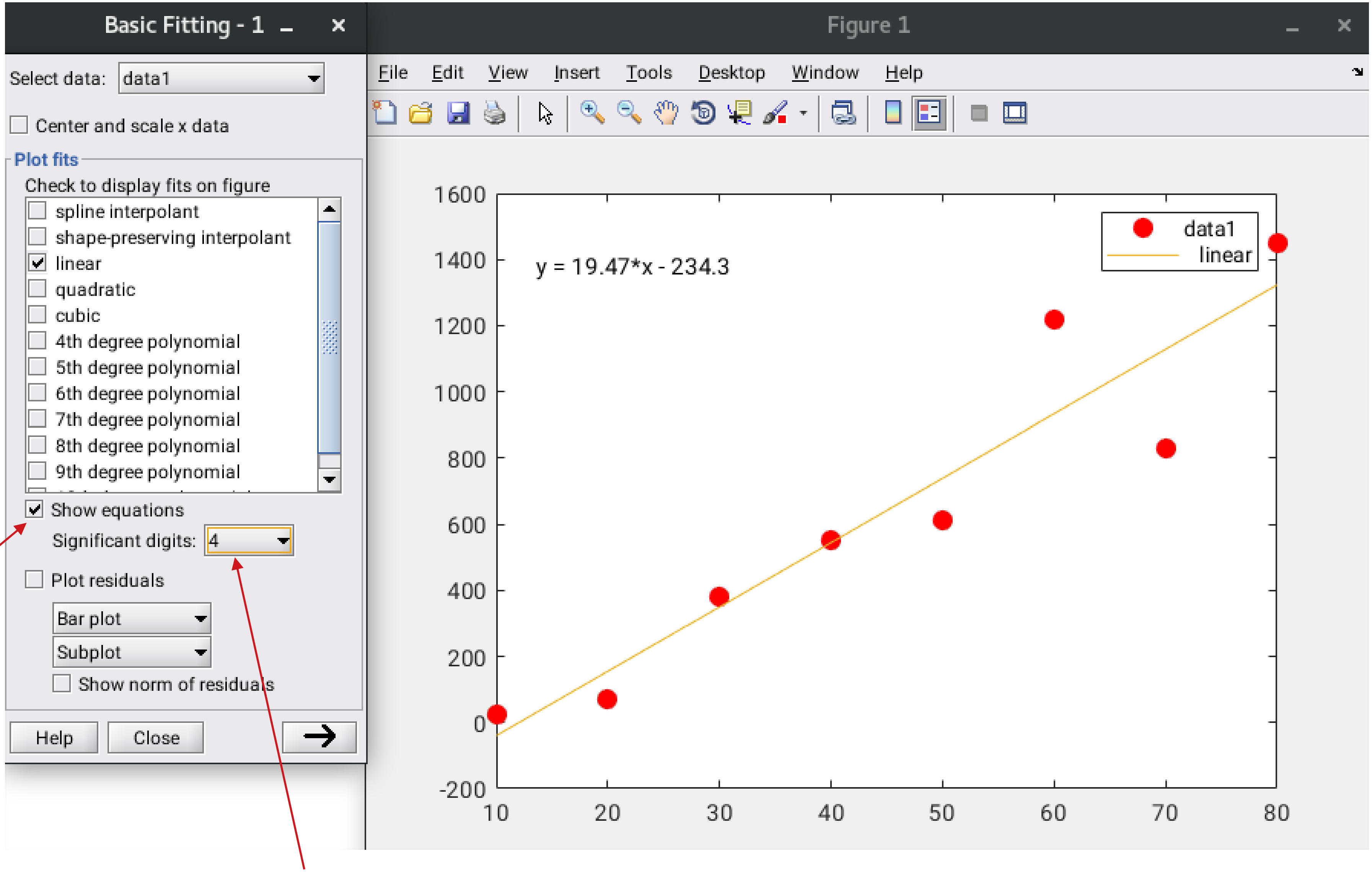
```
>> plot(x,y,'or','Markersize',8,'Markerfacecolor','r')
>> hold on
>> X = linspace(10,80);
>> Y = polyval(a,X);
>> plot(X,Y,'b','Linewidth',3)
```



7.4 Matlab Graphical Method

- You can also use the Matlab graphical fitting tool.





Basic Fitting - 1
Select data: data1
Center and scale x data
Plot fits
Check to display fits on figure
spline interpolant
shape-preserving interpolant
linear
quadratic
cubic
4th degree polynomial
5th degree polynomial
6th degree polynomial
7th degree polynomial
8th degree polynomial
9th degree polynomial
Show equations
Significant digits: 4
Plot residuals
Bar plot
Subplot
Show norm of residuals
Help
Close
Figure 1
File
Edit
View
Insert
Tools
Desktop
Window
Help
y = 19.47*x - 234.3
data1
linear
1600
1400
1200
1000
800
600
400
200
0
-200
10
20
30
40
50
60
70
80

If >> a = polyfit(x,y,2) in Matlab gives a = [-3 0 2], what is the fitting equation?

$$y = -3 + 2x$$

$$y = -3x + 2$$

$$y = -3x^2 + 2$$

$$y = -3 + 2x^2$$

$$y = -3x^2 + 2x$$

$$y = -3x + 2x^2$$

Tc 0

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

If t is the independent variable, r is the dependent variable, then which command fits the data to a power equation?

- q = polyfit(log10(r) , t , 1)
- q = polyfit(r , log10(t) , 1)
- q = polyfit(log10(r), log10(t) , 1)
- q = polyfit(t , log10(r) , 1)
- q = polyfit(log10(t) , r , 1)
- q = polyfit(log10(t), log10(r) , 1)

Tc 0

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**