MODULE 2

Method of Least Squares & Curve Fitting

BSM301 · Mathematics III

...finding the curve that best fits scattered real-world data.

What We'll Cover

From scattered data points to a single trustworthy trend line

1. **Why Curve Fitting?**

   Real data is noisy — we need the "best" curve through it

2. **Principle of Least Squares**

   Minimizing the sum of squared residuals

3. **Fitting a Straight Line**

   Normal equations — with full worked arithmetic

4. **Fitting a Parabola**

   When a line isn't flexible enough

5. **Fitting an Exponential Curve**

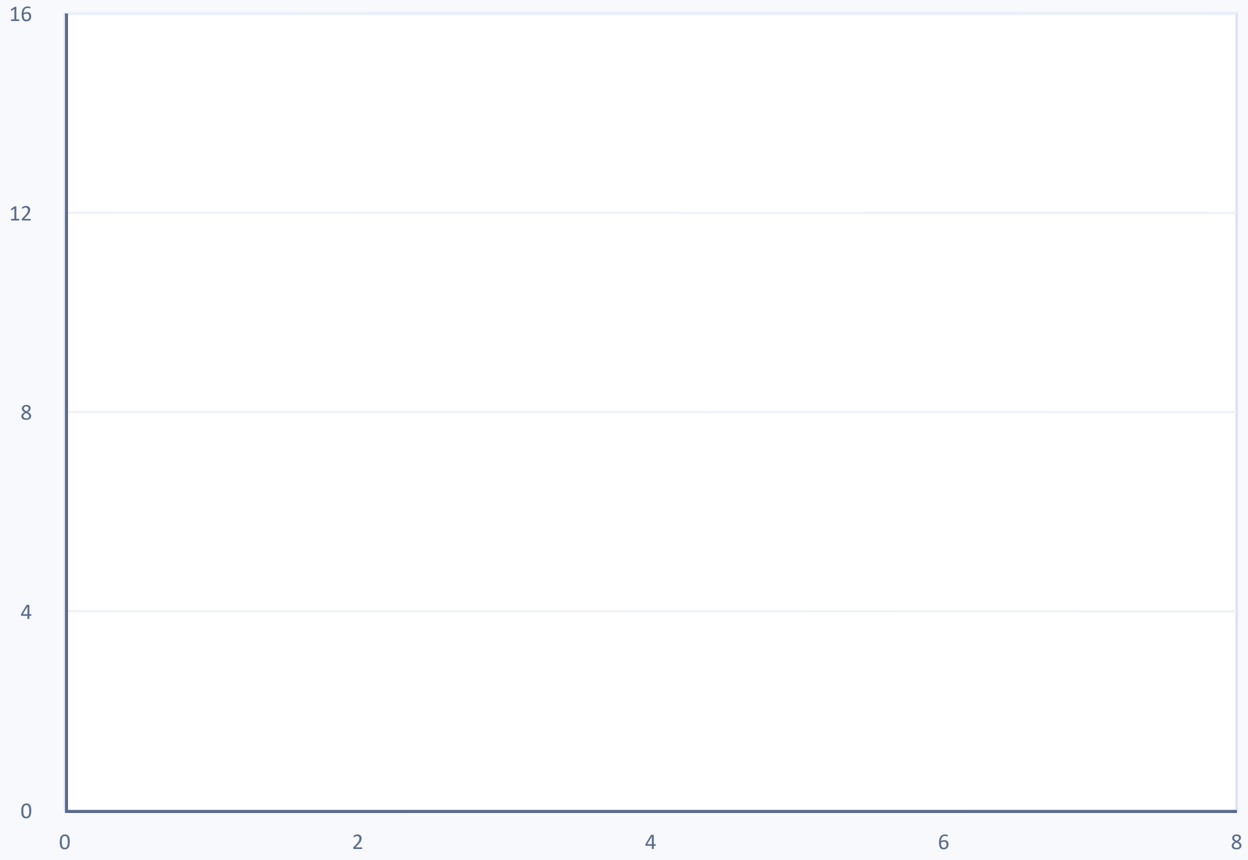
   Linearizing growth trends with logarithms

6. **Choosing the Right Curve**

   Comparing fits and reading residuals

Why Curve Fitting?

Real measurements never lie perfectly on a curve — we need the one that fits best



e.g. sensor test time (x) vs. raw reading (y)

The Problem

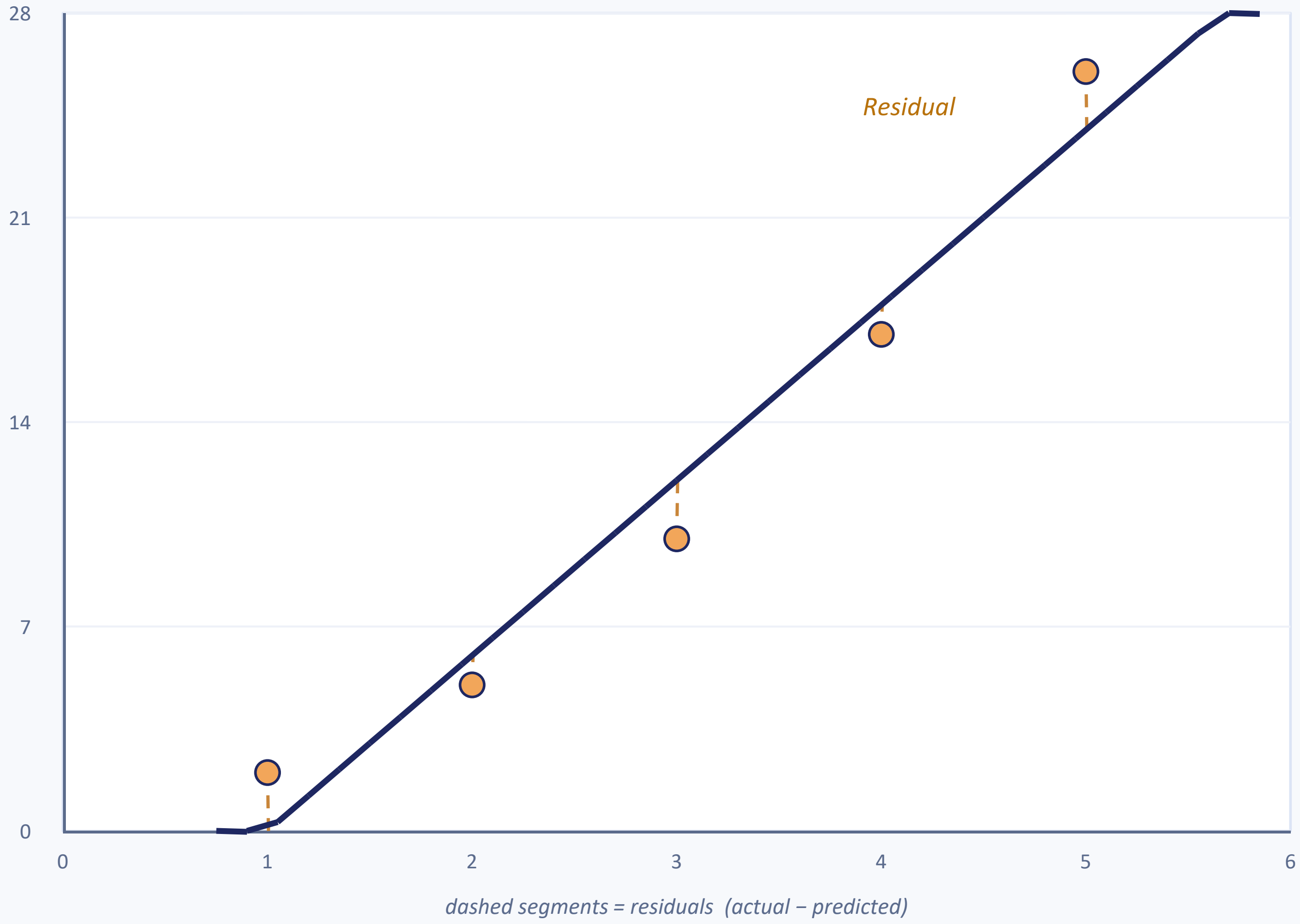
No single straight line, parabola, or exponential curve can pass through every noisy point exactly.

We need a rule that picks the **“best”** curve — one that stays as close as possible to every point, on average.

That rule is the **Method of Least Squares**.

The Principle of Least Squares

Minimize the sum of the squared vertical distances from every point to the curve



Residual $e_i = \mathbf{y_i - \hat{y}_i}$

Minimize: $\mathbf{S = \Sigma e_i^2}$

Why square the residuals?

Squaring makes every residual positive (so points above and below the curve don't cancel out) and penalizes large errors far more than small ones — which is exactly what a "best fit" should do.

Fitting a Straight Line

$y = a + bx$ — the two normal equations that pin down a and b

$$\Sigma y = na + b\Sigma x \qquad \Sigma xy = a\Sigma x + b\Sigma x^2$$

Solving these two equations simultaneously gives:

Slope

$$b = (n\Sigma xy - \Sigma x\Sigma y) / (n\Sigma x^2 - (\Sigma x)^2)$$

Intercept

$$a = (\Sigma y - b\Sigma x) / n$$

How to use them: build a small table of x, y, xy, and x^2 for every data point, sum each column, then plug the totals (n, Σx, Σy, Σxy, Σx^2) straight into the two formulas above — no calculus needed once the sums are ready.

Worked Example — Straight-Line Fit

EXAMPLE

Fitting $y = a + bx$ to five calibration readings

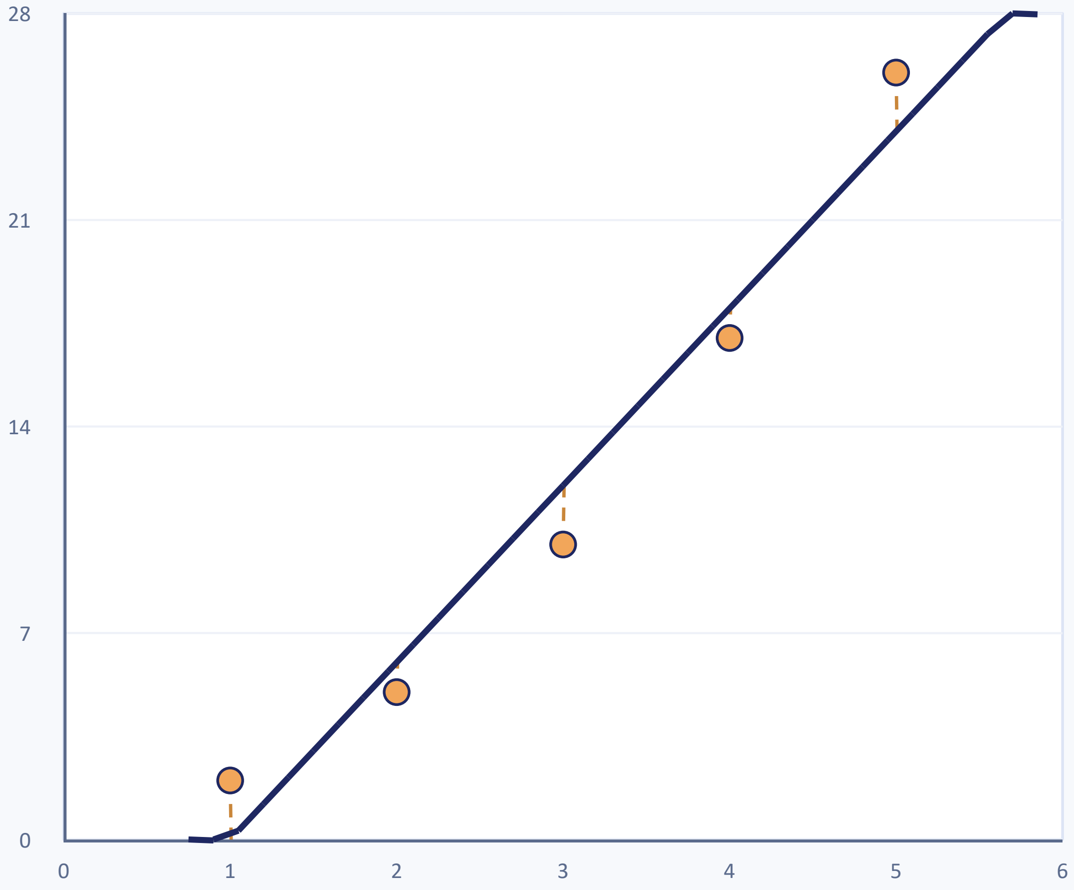
x	1	2	3	4	5
y	2	5	10	17	26
xy	2	10	30	68	130
x^2	1	4	9	16	25

$n = 5$ $\Sigma x = 15$ $\Sigma y = 60$ $\Sigma xy = 240$ $\Sigma x^2 = 55$

Step-by-Step

1. Slope: $b = (5\times240 - 15\times60) / (5\times55 - 15^2)$
2. $b = 300 / 50 = 6.00$
3. Intercept: $a = (60 - 6\times15) / 5 = -6.00$

Fitted line: $y = -6 + 6x$



Sum of squared residuals (SSE) = 14.0 — notice the line still misses every point.

Fitting a Second-Degree Parabola

$y = a + bx + cx^2$ — when a straight line just isn't flexible enough

Three Normal Equations

$$\Sigma y = na + b\Sigma x + c\Sigma x^2$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$$

$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$$

How to solve:

extend the summary table with columns for x^3, x^4, and x^2y, sum each column, then solve the three simultaneous equations for a, b, and c — by substitution, elimination, or matrix methods.

When to reach for a parabola: when a scatter plot curves — rising then flattening, or accelerating — rather than tracking a straight trend.

Worked Example — Parabola Fit

EXAMPLE

Same five readings — does a parabola fit better than the line?

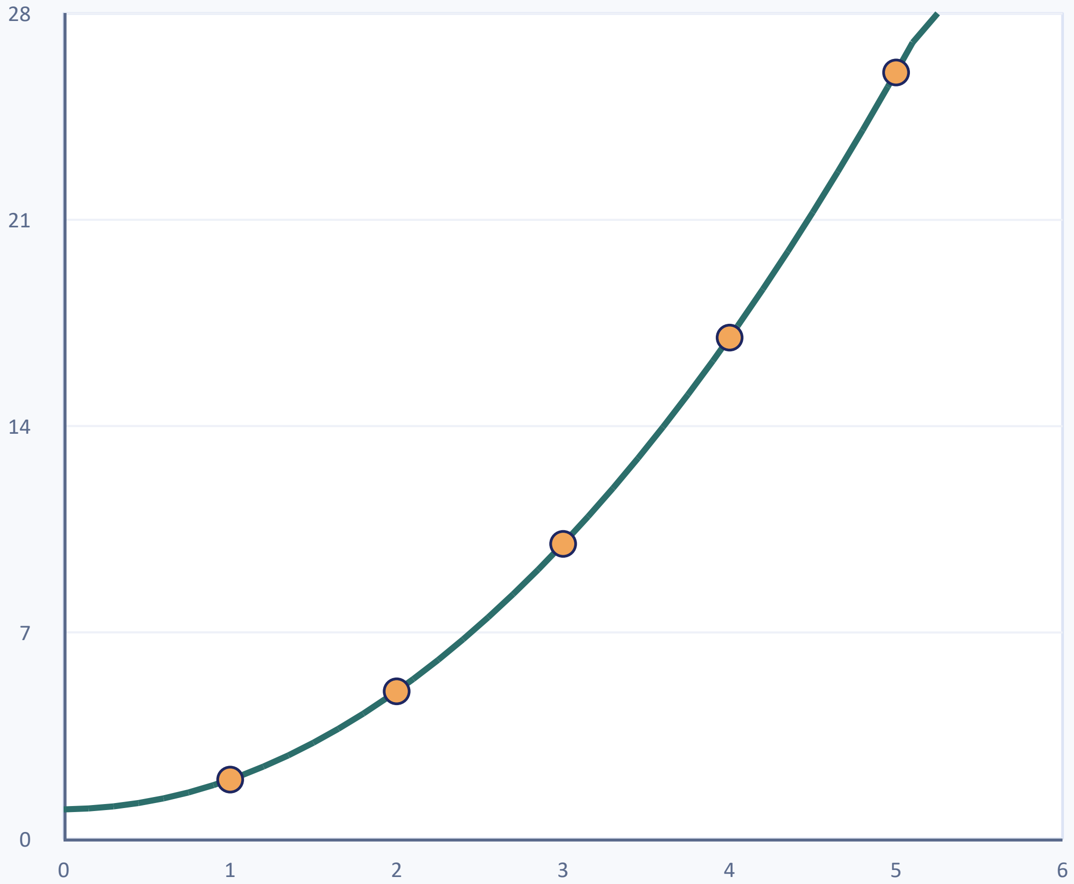
$n=5$ $\Sigma x=15$ $\Sigma x^2=55$ $\Sigma x^3=225$ $\Sigma x^4=979$

$\Sigma y=60$ $\Sigma xy=240$ $\Sigma x^2y=1034$

Solving the 3 Normal Equations

1. 1) $60 = 5a + 15b + 55c$
2. 2) $240 = 15a + 55b + 225c$
3. 3) $1034 = 55a + 225b + 979c$
4. Eliminate a, b → $c = 1$, $b = 0$, $a = 1$

Fitted parabola: $y = 1 + x^2$



SSE = 0.0 — the parabola passes through every point exactly!

Fitting an Exponential Curve

$y = a\cdot e^{bx}$ — turn a curved trend into a straight one first

The Trick: Take Logarithms

1. Start: $y = a\cdot e^{bx}$
2. Take ln of both sides: $\ln(y) = \ln(a) + bx$
3. Relabel: $Y = A + bx$ where $Y = \ln(y)$, $A = \ln(a)$

This is just a straight line in disguise! Fit $Y = A + bx$ to $(x, \ln y)$ using the exact same normal equations from the straight-line case — then recover the original curve with $a = e^A$.

Where this shows up: malware-variant growth, blockchain network hash-rate trends, viral-post share counts — anything that grows (or decays) by a constant percentage each step.

EXAMPLE

Worked Example — Exponential Fit

Malware variant count observed monthly — does it double each month?

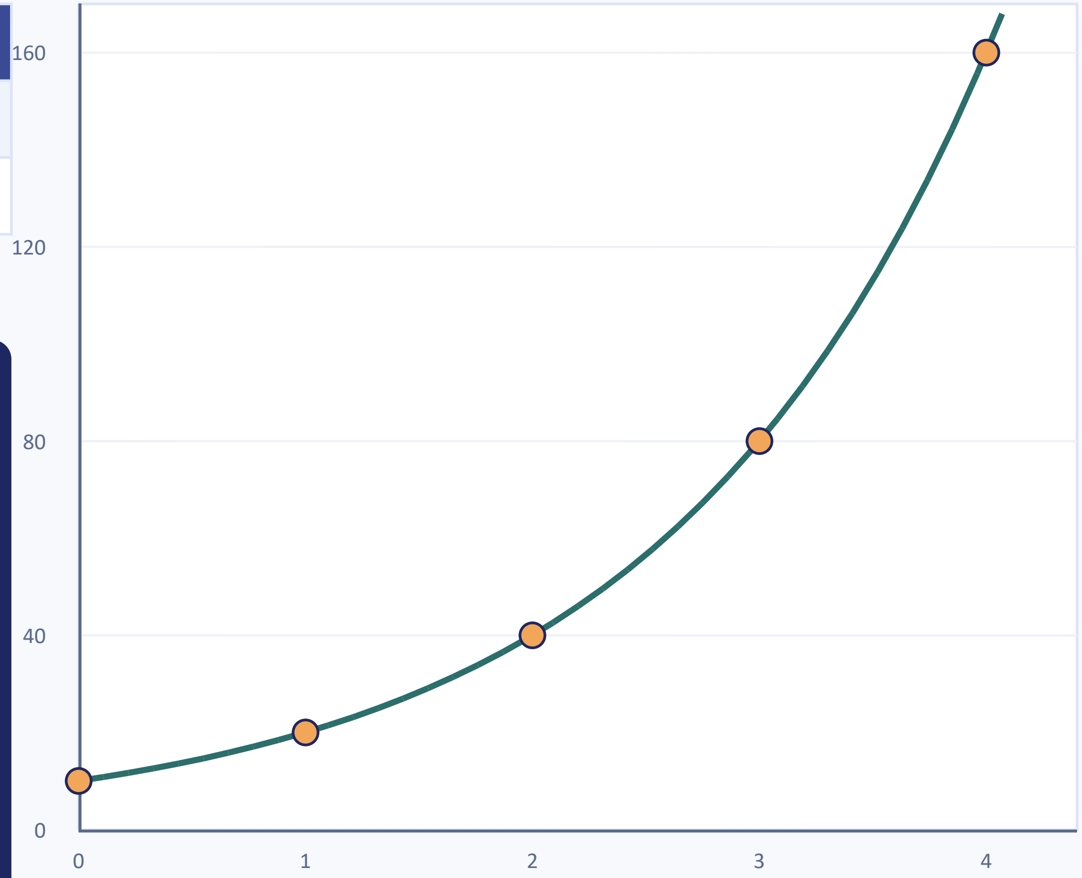
t	0	1	2	3	4
N	10	20	40	80	160
Y=ln N	2.303	2.996	3.689	4.382	5.075

n=5 Σt=10 ΣY=18.444 ΣtY=43.820 Σt²=30

Step-by-Step

1. Slope b: $(5\times43.82 - 10\times18.44) / (5\times30 - 10^2) = 0.6931$
2. Intercept A: $(18.44 - 0.69\times10) / 5 = 2.3026$
3. Recover a: $a = e^{2.3026} = 10.00$

Fitted curve: $N = 10\cdot e^{0.693t} = 10\cdot2^t$



b = ln 2, so the variant count exactly doubles every month.

Choosing the Right Curve

The same 5 points, three different models — and very different fits

Model	Equation	Fitted to Our Data	SSE	Verdict
Straight Line	$y = a + bx$	$y = -6 + 6x$	14.0	Systematic curvature left over
Parabola	$y = a + bx + cx^2$	$y = 1 + x^2$	0.0	Exact fit — matches the true pattern

The lesson: a lower SSE means a better fit — but always look at the residual pattern too. If residuals still show a clear curve or trend, a straight line is the wrong model, no matter how small its SSE looks in isolation.

Use a straight line when…

the scatter plot looks like a constant, linear trend

Use a parabola when…

the trend curves — accelerating, decelerating, or U-shaped

Use an exponential when…

the quantity grows or decays by a constant percentage each step

Curve Fitting in Practice

The same three models, applied across your specialization tracks

Internet of Things

Straight Line

Calibrating a raw sensor reading against a reference thermometer

Parabola

Modeling sensor drift that accelerates as a battery ages

Exponential

Capacity fade of a battery over repeated charge cycles

Cyber Security

Straight Line

Tracking a steady week-over-week rise in phishing attempts

Parabola

Modeling an attack's ramp-up and eventual plateau

Exponential

Forecasting near-term growth of a new malware family

Blockchain

Straight Line

Tracking steady weekly growth in network hash rate

Parabola

Modeling gas-fee spikes that build up and taper off

Exponential

Projecting active-wallet growth during early adoption

Key Takeaways

1. The Method of Least Squares picks the curve that minimizes the sum of squared residuals — no other curve does better on average.

2. For a straight line, two normal equations (from Σy and Σxy) solve directly for the slope and intercept.

3. A parabola adds one more normal equation (from $\Sigma x^2 y$) and can fit curvature a straight line simply cannot.

4. An exponential curve becomes a straight line after taking logarithms — the same normal equations then apply.

5. Next: Module 3 — sampling and sampling distributions, using these same summarizing ideas on random samples instead of fixed data.