Chapter 20

Fitting and Testing

Sixteen chapters have built models and taken their parameters as given. This one asks where the numbers come from and how anyone would know a model is wrong. The two questions turn out to be different problems: fitting to today's prices is an inverse problem with no statistics in it, while fitting to a history is a statistical problem with a bias large enough to reverse a trading decision. Both have the same failure, which is a fit that looks excellent while determining almost nothing, and both have the same remedy, which is to measure how far the answer can move before anyone would notice. We finish with the one general test of whether a model fits at all; chapter 21 continues into the harder case where the parameter itself is never observed.

20.1 Choosing the Model in the First Place

Before a parameter can be fitted, something has to be chosen to fit, and these notes have now built enough models that the choice is real. This section is the one a reader is entitled to ask for and most texts omit: given a trade, which model, and why.

There is a single question underneath it, and it is the one chapter 15 ended on.

Remark (The question that decides it)**.** *What is this payoff a bet on, and does the candidate model have a view on that quantity or has it assumed one?*

Every entry in the table below is an application of that question. A model is adequate when the things the payoff depends on are things the model was calibrated to; it is dangerous exactly when the payoff is sensitive to something the model fixed by assumption — which is chapter 18's Gaussian copula, chapter 9's local volatility, and chapter 10's sticky strike, all over again.

Does it need a model at all?

The first question is whether to reach for one, and surprisingly often the answer is no.

If the payoff is a function of one rate or price observed on one date, its distribution on that date is the whole of the relevant information, and the option market quotes that distribution. Static replication then prices it exactly, as chapter 15 did for constant maturity swaps and cash settled

swaptions, and chapter 9's Breeden-Litzenberger did for anything written on a single equity fixing. No model is better than a model here, because a model can only add an assumption to information that was already complete.

A model becomes necessary when the payoff depends on more than one date, on the path, on more than one underlying, or on when the holder chooses to act. Those four are the whole of it.

Which model, and what each one is really for

Model	**Reach for it when**	**It assumes away**
Black-Scholes / Black-76	Quoting and hedging a vanilla	Everything about the smile
Local volatility (ch. 9)	The payoff depends only on terminal distributions	Volatility dynamics: the smile is predicted to slide, wrongly
SABR (ch. 10)	Interpolating one expiry's smile	Any consistency between expiries
SVI	Interpolating one expiry's smile so that it can be differentiated	Dynamics entirely — it is a curve, not a process
Heston (ch. 10)	A consistent surface across expiries; forward volatility	Fit to any single expiry's wings
LSV (ch. 11)	Barriers and cliquets: exact vanilla fit *and* live dynamics	Nothing much, at the cost of being slow and fiddly to build
Hull-White (ch. 8)	Bermudans and callables where the smile is second order	The smile entirely, and any decorrelation along the curve
Quasi-Gaussian / Cheyette (ch. 12)	Callables that are also smile sensitive	Realism in exchange for a finite Markov state
Market models (ch. 13)	Products written on the actual quoted rates; flexible correlation	Tractability: it is a simulation, so callables need regression
Markov functional	Callables needing an exact vanilla fit and speed	Any control over the dynamics beyond the one factor

Remark (The row that is not a model)**.** One entry in that table is a different kind of thing and the difference is the point of including it. Every other row is a process: it says what the underlying does, and a price follows. SVI — the stochastic volatility inspired parameterisation, which contains no stochastic volatility — says only what one expiry's smile *looks like*. It writes the total implied variance $w(k) = \sigma^2_{\text{imp}}(k)\,T$ at log-moneyness $k = \ln(K/F)$ as

$$w(k) = a + b\Big(\rho\,(k-m) + \sqrt{(k-m)^2 + s^2}\Big), \tag{20.1}$$

and stops. There is no underlying and no dynamics, so it cannot price a barrier and has nothing to say about tomorrow's smile.

The five parameters are readable. Vertical shift a is the level, horizontal shift m is where the vertex sits, s is how rounded the vertex is, b sets the size of the wings, and $\rho \in (-1, 1)$ tilts them: the wing slopes are $b(1-\rho)$ to the left and $b(1+\rho)$ to the right,[1] so a negative ρ is the equity skew.

[1] `svi::Svi::wing_slopes`, checked against the curve far out in each wing.

Structure (Why a hyperbola)**.** Expression (20.1) looks like an arbitrary functional form and is less arbitrary than it looks. It is a hyperbola, written in coordinates that name its degrees of freedom, and the constraints on a smile point at that shape without quite forcing it.

A smile must do three things. It must be smooth and convex in the middle, since chapter 9 makes the second derivative in strike a probability density. It must grow at most linearly in each wing, which is Lee's moment formula: total variance rising faster than $2|k|$ would price deep options as though the underlying had moments it cannot have. And the two wings are free to grow at different rates, because skew exists.

Those constraints narrow the field sharply. They rule out anything with quadratic wings, anything obliged to be symmetric, anything not convex. What they do not do is determine a single curve: adding to any admissible smile a smooth convex bump that dies away at both ends leaves the asymptotes and the convexity exactly as they were, so there are infinitely many curves meeting all three conditions, and (20.1) is not the general solution of them.

What can be said is smaller and still stands. Among *conics*, the condition is decisive: an ellipse has no real asymptote and a parabola has one, so a conic with two distinct real asymptotes is a hyperbola and can be nothing else. So (20.1) is the lowest-degree algebraic curve that does what a smile has to do, and its five parameters are exactly the count those conditions suggest — two asymptotic slopes, two shifts, one radius of curvature at the vertex. That the count matches is an observation rather than a derivation, and the form remains a choice: it is the minimal one, made once and kept because it fits, not a consequence with a proof behind it.

Which also says exactly where it can go wrong. The whole of Lee's condition is that neither asymptote is steeper than two, or $b(1 + |\rho|) \leq 2$, and it is a constraint on the fit rather than a property of the form. Violating it produces a smile whose implied density goes negative, and the density is what Dupire divides by — a fit outside the bound does not degrade the local volatility surface, it changes the sign of its denominator.[2]

Remark (The three families, and what actually separates them)**.** The table is long and the underlying taxonomy is short. Interest rate models differ in what they take as the state variable, and that choice, not the label, determines what each is good for.

Short rate models — Hull-White, and the quasi-Gaussian models of chapter 12 — carry a small Markov state. That makes them solvable on a lattice or a partial differential equation grid, which makes early exercise cheap, which is why callables are priced with them. The price of the small state is that the whole curve moves as a function of one or two numbers, so decorrelation between distant parts of the curve is not available.

Market models take the traded forward rates themselves as state variables. That is the honest choice — they model what is quoted, with as much correlation structure as one cares to specify — and it costs tractability, since the state is high-dimensional and simulation is the only route. Early exercise then needs a regression method, and every regression method is an approximation nobody can bound.

Markov functional models take neither, and instead impose that the state is one dimensional and choose the functional form to reproduce vanilla prices exactly. They are fast and they fit, and what one gives up is any independent control of the dynamics: having spent the freedom on the fit, there is none left for how the curve moves.

[2] `measured`, with the density obtained by differencing prices rather than from the parameters.

Choosing between them is choosing which of exact fit, rich dynamics, and cheap early exercise to give up, because no model in use gives all three.

Example 20.1 (Four trades, four answers)**.** - *A CMS cap.* One rate, one date per caplet. No model — replicate over the swaption smile, chapter 15. Reaching for Hull-White here obtains a worse answer more slowly, because the model's own smile replaces the market's.

- *A Bermudan swaption.* Several exercise dates, so the joint distribution matters and no set of European quotes supplies it. A short rate model, because early exercise is the binding constraint and a lattice is what makes it cheap.

- *An equity barrier.* The payoff depends on the path, and on the smile at every level the barrier might be crossed at. Local-stochastic volatility, chapter 11, because a pure local volatility model prices the barrier off dynamics it got wrong and a pure stochastic volatility model does not fit the vanillas it will be hedged with.

- *A callable range accrual.* Path dependent, smile sensitive and callable at once, which is the intersection that has no comfortable answer. Quasi-Gaussian if the callability dominates, a market model with regression if the smile does, and in either case a reserve against the difference between the two prices — which is the honest measure of the model risk, and larger than most desks admit.

Remark (The rule that survives when the table does not)**.** Models come and go and the table above will date. What does not is the discipline: price the trade in two models that disagree about the thing the payoff is most sensitive to, and reserve the difference. If the two agree, the choice did not matter and the reserve is zero. If they disagree, the gap is the part of the price that was never determined by the market — and it is better held as a number on the balance sheet than discovered later.

The rest of this chapter is about what to do once a model is chosen. This section was about not choosing one for a trade that never needed it.

20.2 Two Jobs With One Name

A desk says "calibration" for both of the following, and they have almost nothing in common.

Definition 20.1 (Calibration)**.** Choosing parameters so that a model reproduces today's quoted prices. Cross-sectional, done under $\mathbb{Q}$, repeated every morning.

Definition 20.2 (Estimation)**.** Choosing parameters so that a model matches the behaviour of a historical time series. Longitudinal, done under $\mathbb{P}$, and a question of statistical inference.

Chapter 6 drew the measure distinction and this is where it bites hardest. Calibration has *no sampling error* — there is no sample. The quotes are what they are, and if the model reproduces them there is nothing left to be uncertain about in the statistical sense. Its difficulty is entirely that the map from parameters to prices may not be invertible.

Estimation has the opposite profile. The model may be inverted perfectly well; the trouble is that a finite history is a finite sample, and the estimator built on it can be badly biased in a direction nobody checks.

Remark (Why the confusion is expensive)**.** A parameter fitted one way is routinely used as though it had been fitted the other. Correlation is the standard offender: a correlation estimated from history is a $\mathbb{P}$ quantity, a correlation calibrated to quanto or spread option prices is a $\mathbb{Q}$ quantity, and they differ by a risk premium exactly as chapter 17's forward differed from the expected spot. Substituting one for the other is not an approximation. It is a category error that happens to produce a number.

20.3 A Perfect Fit That Determines Nothing

Take the SABR model of chapter 10 and the cleanest possible test of whether calibration recovers parameters: generate the quotes *from a known SABR model*, so the model is exactly right and there is no noise, no bid-offer and no misspecification. If the parameters cannot be recovered here, they cannot be recovered anywhere.

A five year expiry, a forward at 3%, seven strikes spanning two percent either side, and $(\alpha, \rho, \nu) = (0.30, -0.30, 0.40)$ with β fixed at 0.5. Starting the optimiser deliberately far away, at $(0.15, +0.40, 1.20)$, it converges and reproduces every quote to within 10^{-5} volatility points. A perfect fit.

Now ask the only question that matters: how far can each parameter be moved, re-optimising the others, before the fit degrades by a tenth of a volatility point — an amount several times smaller than any real bid-offer?[3]

Parameter	Fitted	Range that still fits	Width
α	0.300	[0.295, 0.304]	0.008
ρ	−0.300	[−0.999, −0.276]	0.72
ν	0.400	[0.070, 0.647]	0.58

The level parameter α is pinned to about one percent of itself. The correlation is not pinned at all: it can be anything from -0.28 to -1.0, and every one of those values reproduces the whole smile to inside a tenth of a point. The volatility of volatility can be anywhere in a nine-fold range.

Remark (What is and is not being claimed)**.** Not that the optimiser failed — it returned the generating parameters exactly. And not that the minimum is non-unique: with noiseless quotes it is unique, sitting at the truth.

The claim is that the minimum is not identified *to the precision of the data.* Every quote in a real market arrives with a bid-offer around it, and once the objective is flat within that width, the parameters inside the flat region are indistinguishable on the evidence. Reporting one of them to two decimal places is a statement about the optimiser's stopping rule. With a realistic half-point bid-offer the regions above get wider still.

Remark (The identification problem, for the fourth time)**.** This is the shape these notes keep meeting. Chapter 16: a spread pins down $\lambda(1 - R)$ and neither factor, so recovery is fixed by convention. Chapter 18: the marginals are pinned and the copula is free. Chapters 10 to 12: the smile pins the marginals and leaves the dynamics open. Here: the quotes pin the level and leave the shape parameters loose.

[3] α against ρ, and the width of ρ.

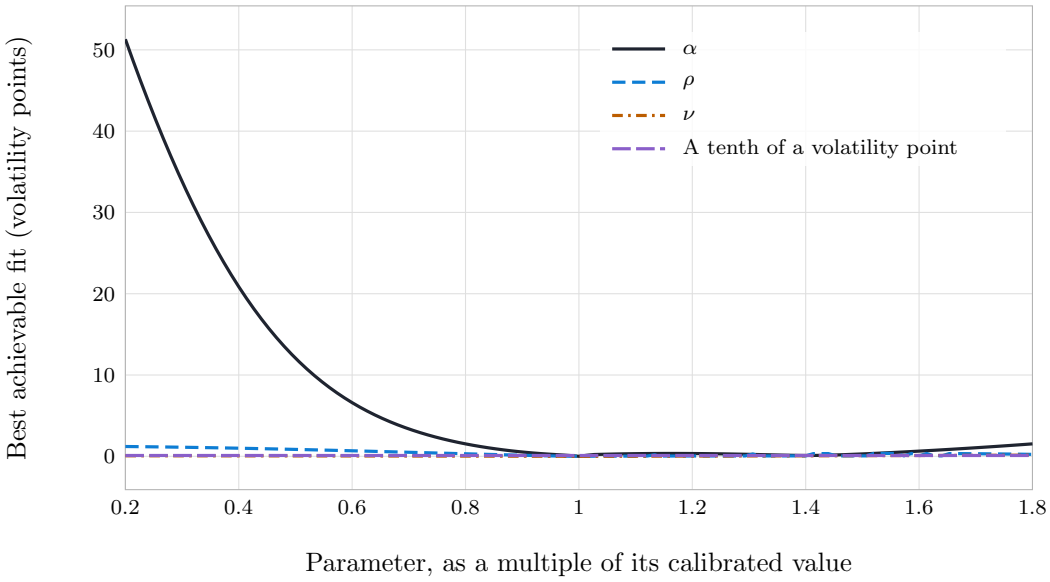

Figure 20.1: Each curve holds one parameter away from its calibrated value and re-optimises the other two, so it shows the best fit still achievable. The shape is the answer. α has a sharp minimum — move it and nothing recovers, because every quote constrains the level. ρ and ν have long flat floors running well below the tenth-of-a-point line, so a wide range of them is indistinguishable from the truth. A calibration report quoting $\rho = -0.30$ is reporting where the optimiser stopped, not something that was measured.

Drawn by `profile` and `identifiable_range`.

Each time the resolution is the same and it is not statistical. Either bring more information — more instruments, of a kind that depends on the unidentified quantity — or fix the parameter by convention and be honest that it was fixed. The one thing that never works is a better optimiser, because there is nothing wrong with the optimiser.

Example 20.2 (Bringing more information)**.** The constructive half, and it is measurable. Widening the quoted strikes, holding everything else fixed, changes the range of ρ that survives:[4]

Strikes quoted	ρ still fitting	Width
$\pm 0.5\%$	$[-0.999,\ -0.209]$	0.79
$\pm 1\%$	$[-0.999,\ -0.252]$	0.75
$\pm 2.5\%$	$[-0.314,\ -0.287]$	0.03

A thirtyfold improvement, from nothing but asking for quotes further out of the money. Correlation is a statement about the wings, so the wings are where the information about it lives, and a set of near-the-money quotes simply does not contain it. This is what "bring more information" means concretely, and it is a conversation with the broker rather than with the model.

Remark (Why β is fixed by convention)**.** Chapter 10 noted that β is conventionally set rather than fitted and left it there. The table above is the justification. A single expiry's smile already fails to pin ρ and ν; adding a fourth parameter that also controls shape makes the flat region larger still. β is chosen from a view about how the smile should move when the forward moves — which is dynamic information, and therefore precisely what a single day's quotes cannot contain.

20.4 The Bias Nobody Checks

Now the other job. The canonical estimation problem in these notes is the one chapter 22 will build its trades on: a spread is believed to mean revert, and the trade is sized by how fast.

Model it as an Ornstein-Uhlenbeck process,

$$dX_t = \kappa(\theta - X_t)\,dt + \sigma\,dW_t, \tag{20.2}$$

where κ is the reversion rate and $\ln 2/\kappa$ the half-life. Sampled at spacing Δ, (20.2) is an autoregression,

$$X_{i+1} = c + \phi X_i + \varepsilon_i, \qquad \phi = e^{-\kappa\Delta}, \tag{20.3}$$

so regressing each observation on the previous one recovers ϕ and hence κ. This is also the maximum likelihood estimator, since the transition is Gaussian. It is what every desk does.

It is biased, and severely.

Calculation 20.3 (Where the bias comes from and what it scales with)**.** The least squares estimate of an autoregressive coefficient is biased downwards — the classical result is

$$\mathbb{E}[\hat{\phi}] - \phi \approx -\frac{1 + 3\phi}{n}$$

for n observations.

[4] the narrowing, checked.

Least squares is unbiased under *strict exogeneity*: the errors must be uncorrelated with the regressor at every date, past and future, not merely at the same one. Regressing x_t on x_{t-1} satisfies the weaker condition and fails the stronger. It satisfies $\mathbb{E}[x_{t-1}\varepsilon_t] = 0$, since x_{t-1} is fixed before ε_t arrives — which is why the estimator is consistent, and why the problem disappears as the sample grows. But $x_{t-1} = \phi x_{t-2} + \varepsilon_{t-1}$ is built out of the earlier shocks, so it is correlated with ε_{t-1}, ε_{t-2} and all the rest. Consistency is what predeterminedness buys; unbiasedness is what strict exogeneity would have bought, and it is not available.

That says the estimator is biased without saying which way, and the sign has two sources. The estimate is a ratio,

$$\hat{\phi} = \frac{\sum (x_t - \bar{x})(x_{t-1} - \bar{x})}{\sum (x_{t-1} - \bar{x})^2},$$

and its denominator is not independent of its numerator: a shock raises x at its own date and, through the recursion, at every date after, so it inflates the sum of squares in a way that is tied to the errors above. The ratio is pulled down. That much happens even when the mean is known, and it gives a bias of about $-2\phi/n$.

Estimating the mean adds the rest, and is the easier half to picture. $\bar{x}$ is computed from the same path, and a sample mean drawn from one realisation of a persistent series sits closer to that realisation than the true mean does — the path drags its own average along with it. Measured against a centre that has followed it, the series appears to return sooner than it truly does. That is what takes the bias from $-2\phi/n$ to the $-(1+3\phi)/n$ above, and at ϕ near one it roughly doubles it. Both halves are measurable, and both are there: simulating an autoregression at $\phi = 0.95$ over two hundred steps, the two fits differ by a factor of about 2.4.[5]

A downward bias in ϕ is an *upward* bias in $\kappa = -\ln \hat{\phi}/\Delta$: the series looks less persistent than it is, so it looks to revert faster than it does.

Now propagate it. With Δ small, $\phi \approx 1$, and

$$\delta\kappa \approx -\frac{\delta\phi}{\phi\Delta} \approx \frac{4}{n\Delta} = \frac{4}{T},$$

where $T = n\Delta$ is the length of the history in years. The Δ has cancelled. The bias depends on the *span* of the data and not on how finely it was sampled, and it does not depend on κ either.

Remark (Read that again, because it is the practical point)**.** The instinct on discovering an estimate is noisy is to get more data, and in practice that means higher frequency data, because history has a fixed length and the tick archive does not. Equation above says that does nothing. Going from daily to hourly multiplies the number of observations by eight and leaves the bias exactly where it was.

The reason is that mean reversion is a statement about a timescale. A history informs you about it only by containing several of them, and sampling the same three years more finely does not contain more three-year periods. Only a longer history does.

Example 20.3 (The size of it)**.** Simulating (20.2) with $\kappa = 1$ — a half-life of eight months — and estimating it the standard way, in `quant/src/estimation.rs`:

[5] `measured`, against Kendall's two constants.

History	Daily	Hourly	Bias $\times T$
1y	6.10	6.04	5.1
2y	3.42	3.42	4.8
5y	1.91	1.93	4.6
10y	1.43	1.44	4.3
20y	1.21	1.21	4.2
40y	1.10	1.11	4.0

A single year of daily data reports a mean reversion of 6.1 against a truth of 1. The last column is the prediction $\delta\kappa \cdot T \approx 4$, and it holds across the range.

In half-lives, which is how a trader thinks: the truth is 8.3 months. Two years of data says 2.4 months. Ten years says 5.8. The estimate does not merely have error bars around the right answer — it is systematically, and largely, on the optimistic side.

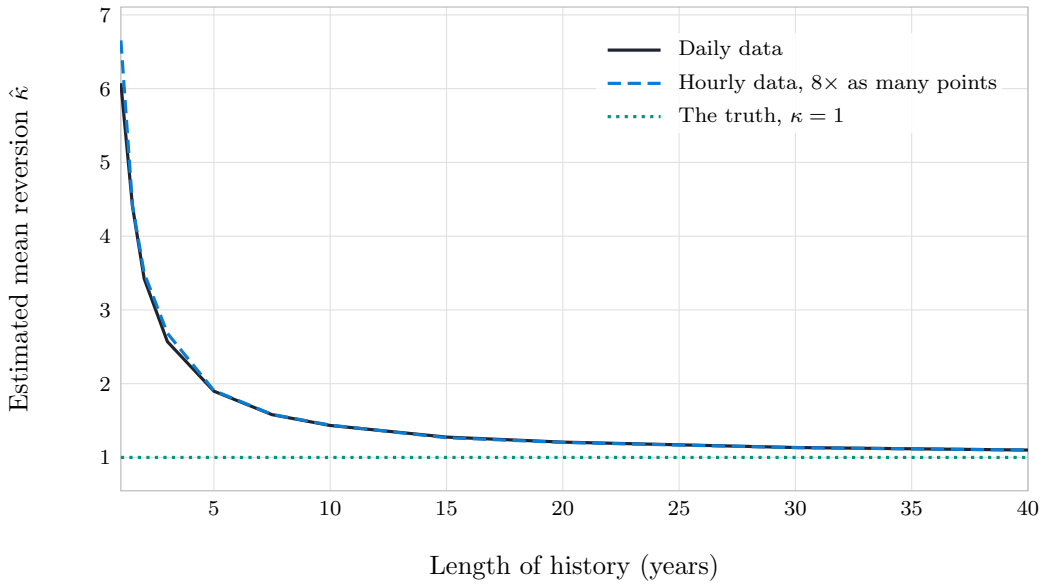

Figure 20.2: The estimated mean reversion against the length of history used, for a process whose true value is the flat line. Two things to read. The daily and hourly curves lie on top of each other, though one uses eight times as many observations — more data of the wrong kind is not more information. And the convergence is slow: it takes decades of history to get within ten percent of a reversion rate whose own half-life is under a year. Move the slider to a slower reverting process and the situation gets worse, because what matters is how many half-lives the window contains.
Drawn by `fit_ou` and `mean_reversion_estimate`.

Remark (What this does to a trade)**.** Chapter 22 will size relative value trades on the assumption that a spread returns to fair value on a known timescale, and this is the estimator that finds it. Suppose the true half-life is eight months and two years of data suggests two and a half.

Everything downstream inherits the error, and in the same direction. The expected holding period is understated by a factor of three, so the expected annualised return is overstated by the same factor. The probability of the spread moving further against the position before it converges

is understated, so the stop is set too tight and the position too large. And the capital charge, which depends on how long the exposure persists, is understated as well.

None of that is a failure of the strategy. It is one biased estimator propagating through every number that depends on it.

Exercise. Show from (20.3) that a series with no mean reversion at all, $\kappa = 0$, still produces a strictly positive $\hat{\kappa}$ with probability approaching one on a finite sample. Deduce that "the regression found mean reversion" is not by itself evidence of any, and propose what would be. (Hint: the null hypothesis $\phi = 1$ is a unit root, and the distribution of $\hat{\phi}$ under it is not the usual one — this is why unit root tests have their own critical values.)

20.5 Testing Rather Than Estimating

Everything above estimates a parameter. This section asks the prior question, which a chapter called fitting and testing owes the reader and which the estimation above quietly assumes: does the data support mean reversion at all?

It matters because the answer is frequently no, and because chapter 22's trades rest on it. A convergence trade is a bet that a spread returns to a level. If the spread is a random walk the bet has no edge whatever, and telling the two apart is a hypothesis test rather than an estimation.

Definition 20.4 (The Dickey-Fuller test)**.** Regress the increment on the level,

$$x_t - x_{t-1} = c + b\,x_{t-1} + \varepsilon_t, \tag{20.4}$$

and test the null $b = 0$, a random walk, against $b < 0$, mean reversion. The statistic is the ordinary t ratio on b.

Remark (Why the critical value is not from a t table)**.** Equation (20.4) looks like a regression and its statistic does not have a t distribution.

Standard regression theory needs the regressor to be well behaved as the sample grows. Here the regressor is the series' own lagged level, and *under the null* that series is a random walk, whose variance grows without bound. The usual limit theory does not apply, the statistic's distribution is skewed to the left, and using -1.65 from a normal table would reject far too often. Dickey and Fuller tabulated the correct distribution by simulation; with an intercept and no trend the five per cent point is about -2.86.

This is the standard trap in testing for mean reversion, and it arises because the null hypothesis is non-stationary. Testing a hypothesis under which nothing converges requires distribution theory in which nothing converges either.

Calculation 20.5 (How much data it takes)**.** Take a spread that genuinely reverts with a one-year half-life, observe it daily, and count how often the test finds the reversion that is really there.[6]

[6] `estimation::inference_quality`, with the test's size checked against a near random walk first, since a power figure means nothing without it.

Sample	Rejects random walk	Median half-life	90% interval
2 years	6%	0.26	0.09 – 1.61
5 years	9%	0.49	0.19 – 1.64
10 years	19%	0.67	0.31 – 1.61
20 years	54%	0.81	0.45 – 1.52
40 years	98%	0.90	0.59 – 1.43

The test's size is five per cent, so the first row is barely distinguishable from rejecting at random. Two years of daily data on a spread that certainly reverts is powerless to say so.

Remark (Sampling faster does not help)**.** The natural response to the first row is to collect more observations, and it does not work. Observing the same five years ten times as often — hourly rather than daily, fifty times the data — leaves the power essentially unchanged.[7]

The reason is the same one chapter 14 gives for the estimation bias. What the test needs is elapsed *half-lives*, because mean reversion is a statement about how the series behaves over intervals comparable to $1/\kappa$, and sampling within an interval says nothing about what happens across it. Fifty times the observations over the same calendar span produces no additional half-lives, and the information about κ does not increase.

Separate this from the usual observation that more data is better. More data along the time axis is better. More data within a fixed window is, for this purpose, almost worthless — and the distinction is invisible if one thinks of a sample only as a number of rows.

Structure (What this does to a horizon)**.** Read the last two columns of Calculation 20.5 against chapter 22's use of them, because the consequence is stronger than that chapter claims.

There the argument was that mean reversion is estimated too fast, so a horizon is planned too short, with the illustration taken at thirty per cent fast. The table says thirty per cent is conservative: at five years of daily data the median estimated half-life is about half the truth, and at ten years about two thirds of it. A desk with a decade of history on a spread will typically believe it reverts half again as fast as it does.

Worse is the interval. At five years the ninety per cent range for the half-life runs from about a fifth of a year to about a year and two thirds — a factor of eight. A horizon is not a quantity that data of that length determines, and a position sized on the point estimate is sized on a number the sample barely constrains.

Which leads to an uncomfortable but honest place. The structural reason chapter 22 requires for a trade is not merely a nice supplement to the statistics; it is doing most of the work, because the statistics on a realistic sample cannot establish either that the spread reverts or how quickly. A trader who says "this spread has a six-month half-life" is reporting a point estimate from a sample that could not distinguish six months from two years, and a trader who says "this spread is wide because a regulation forces somebody to hold the other side" is making a claim the data can actually support.

20.6 Does the Model Fit At All?

Both sections so far assume the model is right and ask what its parameters are. The prior question is whether to use it, and there is a general answer.

[7] `measured, within six percentage points`.

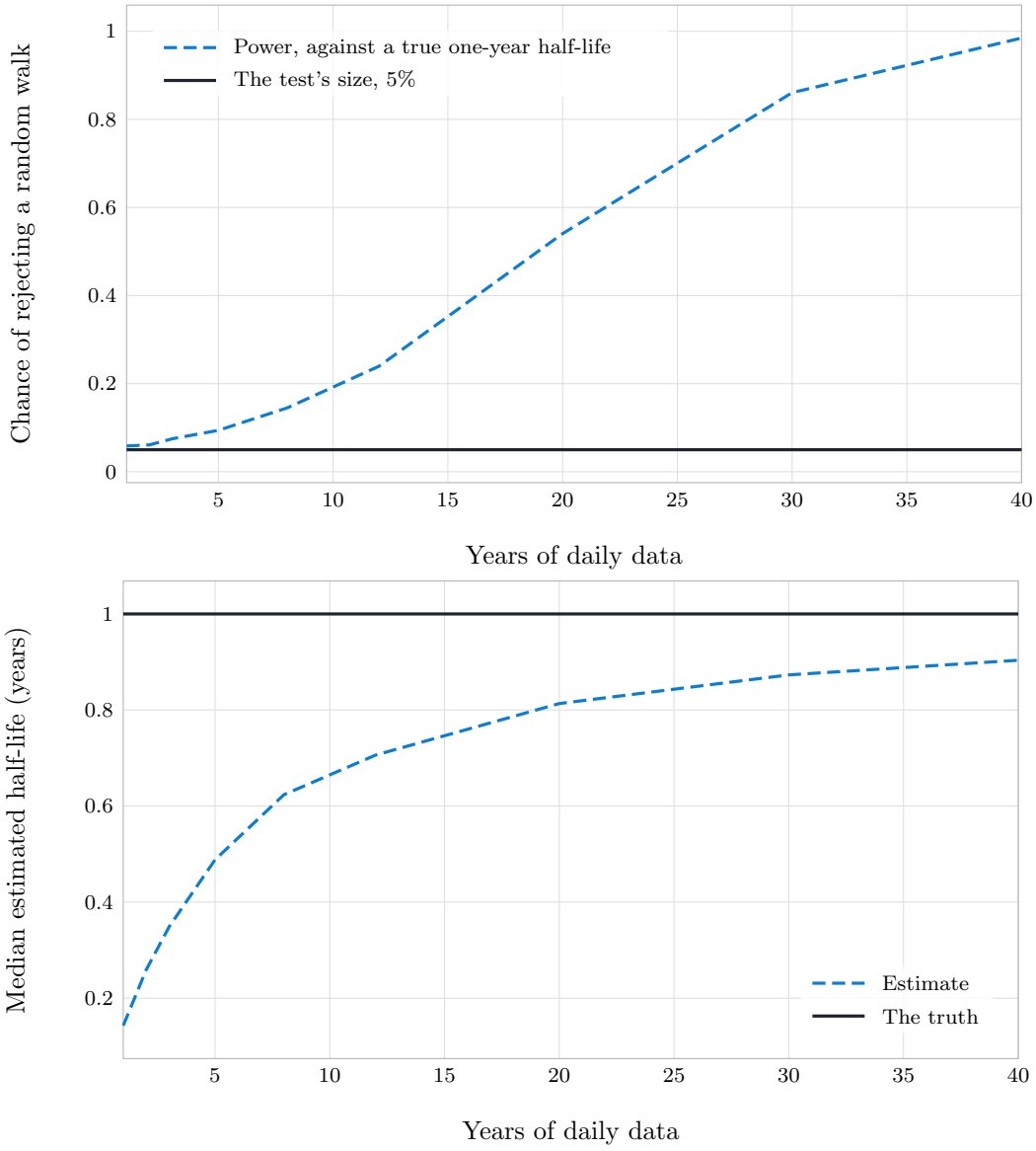

Figure 20.3: Above: the chance of rejecting a random walk, against the length of the record, for a spread that genuinely reverts with a one-year half-life. The flat line is the test's own size — the rate at which it rejects when there is nothing to find — and the two are hard to tell apart until the record is several years long. Below: the median estimated half-life against the truth, approaching it from beneath and still short of it after forty years. Both curves are functions of the calendar span and not of the number of observations, which is the subject of the next remark.

Drawn by `inference_quality`.

Theorem 20.6 (Probability integral transform)**.** *Let X have continuous distribution function F. Then $U = F(X)$ is uniform on $[0, 1]$.*

Proof. $\mathbb{P}(U \le u) = \mathbb{P}(F(X) \le u) = \mathbb{P}(X \le F^{-1}(u)) = F(F^{-1}(u)) = u.$ □

Remark (Why this is the whole test)**.** Trivial to prove and unreasonably useful. If a model claims a conditional distribution for tomorrow's observation given everything known today, then feeding each realised observation through that day's claimed distribution function must produce numbers that are uniform on $[0, 1]$ and independent of each other. Uniform because of Theorem 20.6; independent because each used all the information available at the time, so nothing predictable can be left.

Two properties, both testable, and *no reference to what the model is.* It applies to a diffusion, a jump process, a copula, a neural network, or a trader's intuition written down as a distribution. Anything that outputs a conditional distribution can be tested this way, and anything that does not output one cannot be tested at all — which is itself a useful criterion when someone offers you a model.

Example 20.4 (Catching a volatility that moves)**.** Take four thousand daily returns generated with genuinely stochastic volatility, and test them against a constant-volatility lognormal model — fitted, generously, to the realised variance, so the overall level is right and only the shape is wrong.

The transformed values come out far from uniform, with too many near 0 and 1 and too few in the middle: the fat tails and quiet middle that a single volatility cannot produce. The Kolmogorov-Smirnov statistic rejects at any conventional level. Run the same test on data that really was generated with constant volatility and it passes, so the rejection means something.

Note what the *shape* of the deviation tells you. A U-shaped histogram says the model's tails are too thin. A hump says they are too fat. A tilt says the drift is wrong. The test does not merely reject; it says in which direction to look next.

Remark (Chapter 24's backtest will be a special case)**.** Value at risk backtesting, which chapter 24 comes to, counts the days on which the loss exceeded the reported quantile and asks whether there were too many. In this language: the model claims a conditional distribution, the transform of each day's loss should be uniform, and the exception indicator is just checking whether the transform exceeds 0.99 as often as it should.

Which is why that test is not allowed to stop at the count: the exceptions must also be independent. That is the second half of the transform's conclusion, and it is the half that catches the models that fail worst, since a model that is wrong about volatility clustering will produce the right number of exceptions arriving all in the same week.

Remark (In-sample fit is not evidence)**.** The deepest point in this chapter, and chapter 9 has already made it once. A local volatility model reproduces every European option price exactly. It is nonetheless wrong, in the specific sense that its predictions about how the smile moves are contradicted the next day.

A model with enough parameters fits anything. What distinguishes a model that works is out-of-sample and *dynamic* performance: not whether it reproduces today's surface but whether the hedge it prescribes actually removes the risk, and whether the surface it predicts for tomorrow is the surface that arrives. Chapter 10's decomposition of delta is the concrete form of that test, and it is the one an in-sample fit can never pass or fail.

20.7 What To Take Away

Three things, and they are the same thing seen from different angles.

First, a good fit is not a measurement. Ask of every fitted parameter how far it could move before anyone would notice, and report that width alongside the number. Most of the time nobody has computed it, and often it is embarrassing.

Second, when a parameter turns out to be unidentified, the fix is more information or an honest convention. It is never a better optimiser.

Third, the test of a model is out-of-sample and dynamic. The probability integral transform gives a general way to run that test for anything that emits a conditional distribution, and refusing to emit one should be treated as a property of the model rather than a detail of its implementation.

None of this is specific to derivatives, and none of it has changed in fifty years.

References

- Kendall, M. G. (1954). Note on bias in the estimation of autocorrelation. *Biometrika*, 41(3-4), 403–404.

- Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39(4), 863–883.

- Cont, R. (2006). Model uncertainty and its impact on the pricing of derivative instruments. *Mathematical Finance*, 16(3), 519–547.