# Influential observations, dummy variables, logistic regression and trend models

Quantitative Methods

## 1 A few points can steer the whole line

- Leverage: unusual X values, far from their means
- Outlier: unusual Y value, far from the fitted line
- Influential: removing it would move the estimates
- High leverage alone is not influence

IN THE LESSON

- Imagine thirty people voting on where to eat.
- Twenty-nine want pizza and one shouts for sushi.
- Now imagine the vote counts more for whoever stands furthest from the middle of the room.
- Suddenly the person at the far end of the hall can swing the result.
- A regression line works a little like that second vote.
- Tomas fits the price to book ratio of thirty regional banks on two variables, so k is two and n is thirty.
- Three banks look odd.
- Leverage here means unusual independent variables, and has nothing to do with debt.
- An outlier is unusual in the dependent variable.
- And an influential point is one that would move the estimates if you removed it.

## 2 Three odd banks, drawn

*Chart: Dependent variable vs Independent variable*

- P: far out, on the line
- Q: outlier, mid-range
- R: both, drags the fit

IN THE LESSON

- Here is the picture.
- Along the bottom is an independent variable, and up the side is the dependent variable.
- The grey dots are the ordinary banks, and the green line is the fit made without the three odd ones.
- Start with bank P, far out on the right but sitting right on the line.
- It has high leverage, but it agrees with everyone else, so removing it changes almost nothing.
- Next, bank Q, in the middle of the data but high above the line.
- It is an outlier, yet being in the middle it has little pull on the slope.
- Finally bank R, far out on the right and far below the line.
- It has both.
- Put it back and the red line is dragged towards it.
- That is influence.

## 3 Turning the picture into three rules

**Leverage h > 3 × (k + 1) ÷ n**
**Outlier: studentized residual > critical t (n − k − 2 df)**
**Cook's D > 0.5 look, 1 likely, or 2 × √(k ÷ n)**

| | |
|---|---|
| **h: leverage** | Distance in X, from 0 to 1 |
| **Studentized residual** | Miss, with the point left out |
| **Cook's distance** | Do the estimates move if cut? |
| **k and n** | k variables, n observations |

IN THE LESSON

- Now the rules of thumb that turn the picture into numbers.
- First, leverage, written h.
- It runs from zero to one, and an observation is flagged when it is above three times k plus one, all divided by n.
- Here k is the number of independent variables and n the number of observations.
- Second, the outlier test.
- Re-estimate the model without the point, see how far the point sits from that new line, and divide by its standard error.
- That gives the studentized residual.
- Compare it with the critical t value, with n minus k minus two degrees of freedom.
- Third, Cook's distance, which blends the first two.
- Values above zero point five deserve a look, values above one are likely influential.

## 4 Example 1: set the cut-offs first

*30 banks, k = 2 independent variables*

| | | |
|---|---|---|
| **Leverage cut-off** | 3 × (2 + 1) ÷ 30 | **0.30** |
| **Critical t** | 26 df, 5% two-tailed | **2.06** |
| **Cook's D: look** | above | **0.5** |
| **Cook's D: influential** | above | **1.0** |
| **Cook's D: 2√(k ÷ n)** | 2 × √(2 ÷ 30) | **0.52** |

IN THE LESSON

- Example one.
- Tomas sets his cut-offs before he looks at any bank.
- With two variables and thirty observations, the leverage cut-off is three times two plus one, which is nine, divided by thirty, giving zero point three.
- For the outlier test he needs the critical t value with thirty minus two minus two, so twenty-six, degrees of freedom.
- At the five per cent level, the table value he uses is two point oh six.
- For Cook's distance, anything above zero point five is worth a look, and anything above one is likely influential.
- The other rule of thumb, two times the square root of k over n, gives two times the square root of two over thirty, about zero point five two.

## 5 Example 1: judge banks P, Q and R

*Leverage h, studentized residual t, Cook's D*

| | | |
|---|---|---|
| **P: leverage only** | h 0.42, t 0.5 | **D 0.06** |
| **Q: outlier only** | h 0.06, t 3.2 | **D 0.16** |
| **R: both flags** | h 0.38, t 2.8 | **D 1.28** |
| **Influential** | Cook's D above 1 | **R only** |

IN THE LESSON

- Now judge each bank.
- Bank P has leverage of zero point four two, above the cut-off of zero point three, but a studentized residual of only zero point five, and a Cook's distance of just zero point oh six.
- High leverage, but on the line, so not influential.
- Bank Q has leverage of only zero point oh six but a studentized residual of three point two, above two point oh six, so it is an outlier.
- Its Cook's distance is only zero point one six.
- Bank R has leverage of zero point three eight and a residual of two point eight, so both flags are up, and its Cook's distance is one point two eight, above one.
- Only R is influential.
- Before deleting anything, check R for a data error.

## 6 Putting a category into a regression

- Dummy = a 0 or 1 switch for a category
- n categories need n − 1 dummies
- Intercept dummy: shifts the line up or down
- Slope dummy: dummy × variable, tilts the line

**IN THE LESSON**

- Regression likes numbers, but a lot of finance is made of categories.
- A bond is an industrial, a utility or a financial company.
- A dummy variable is a light switch, either off, zero, or on, one.
- Tomas now models the spread, in basis points, that sixty bonds pay over government bonds.
- The main independent variable is leverage, meaning debt over earnings.
- But he suspects sector matters, so he adds switches.
- There are three sectors, and the rule is one fewer dummy than categories, so two dummies.
- The sector left out is the base group, and every dummy coefficient is a difference from that base group.
- A dummy that shifts the line is an intercept dummy.
- A dummy multiplied by leverage, which lets the slope differ, is a slope dummy.

## 7 Coding three sectors with two dummies

*The omitted sector is the base group*

| Sector | UTL, FIN |
|---|---|
| Industrial (base) | 0, 0 |
| Utility | 1, 0 |
| Financial | 0, 1 |
| Add a third dummy? | Trap |

**IN THE LESSON**

- Here is the coding.
- An industrial bond has zero for both switches.
- A utility bond has one for the utility dummy and zero for the financial one.
- A financial bond has zero for the utility dummy and one for the financial one.
- The base group needs no switch: it is what you get when everything is off.
- Now the trap.
- Suppose Tomas also added an industrial dummy.
- For every single bond, the three dummies would add up to exactly one, which is the same as the column of ones that the intercept uses.
- The computer cannot tell them apart, so the regression breaks.
- That is perfect collinearity, known as the dummy variable trap.
- With n categories, use n minus one dummies.

## 8 The model with both kinds of dummy

**Spread = b0 + b1 × LEV + b2 × UTL + b3 × FIN + b4 × (FIN × LEV)**

| | |
|---|---|
| **b0: intercept** | Industrial, zero leverage |
| **b1: slope** | Industrial slope |
| **b2: UTL dummy** | Utility shift in level |
| **b3: FIN dummy** | Financial shift in level |
| **b4: FIN × LEV** | Financial shift in slope |

**IN THE LESSON**

- Here is Tomas's model written down.
- Spread equals b zero, plus b one times leverage, plus b two times the utility dummy, plus b three times the financial dummy, plus b four times the financial dummy multiplied by leverage.
- Read it group by group.
- For an industrial bond both dummies are zero.
- For a utility bond the intercept becomes b zero plus b two.
- For a financial bond the intercept becomes b zero plus b three, and the slope becomes b one plus b four.
- So b two and b three are intercept shifts, and b four is a slope shift.
- Each is a difference from the industrial base group, and a t test on each tells you whether the difference is real.

## 9 Example 2: read the output

*60 bonds; spread in basis points; critical t about 2.00*

| | | |
|---|---|---|
| **Intercept b0** | industrial, zero leverage | **90 bp** |
| **Leverage b1** | industrial slope | **20 bp** |
| **UTL b2** | t = −2.4, significant | **−25 bp** |
| **FIN b3** | t = 2.9, significant | **+35 bp** |
| **FIN × LEV b4** | t = 2.5, significant | **+12 bp** |

**IN THE LESSON**

- Example two.
- Here is the output.
- The intercept is ninety basis points: that is an industrial bond with zero leverage.
- Each extra turn of leverage adds twenty basis points for industrials.
- The utility dummy is minus twenty-five, with a t statistic of minus two point four, beyond the critical value of about two.
- So utilities pay twenty-five less than comparable industrials.
- The financial dummy is plus thirty-five, with a t statistic of two point nine, so financials start thirty-five higher.
- And the interaction, financial times leverage, is plus twelve, with a t statistic of two point five, so each turn of leverage costs a financial bond twelve more than an industrial one.
- All three differences are significant.

## 10 Three lines, one picture

*Chart: Spread (bp) vs Leverage (debt ÷ EBITDA)*

- Financial
- Industrial
- Utility

**IN THE LESSON**

- Here are the three lines.
- Leverage runs along the bottom and spread goes up the side.
- The middle line is the industrial bonds, starting at ninety and rising twenty for every turn of leverage.
- The utility line sits exactly twenty-five below it all the way along, because the utility dummy changes only the intercept.
- The financial line starts higher, at one hundred and twenty-five, and climbs faster, thirty-two for every turn, so it fans away from the others as leverage rises.
- That fan is what the slope dummy does.
- Without it, the financial line would be parallel to the others.

## 11 Example 2: spreads at leverage of three

*Plug leverage = 3 into each group's own line*

| | | |
|---|---|---|
| **Industrial** | 90 + 20 × 3 | **150 bp** |
| **Utility** | 65 + 20 × 3 | **125 bp** |
| **Financial** | 125 + 32 × 3 | **221 bp** |
| **Financial minus industrial** | 35 + 12 × 3 | **71 bp** |

**IN THE LESSON**

- Put leverage equal to three into each group's equation.
- Industrial: ninety plus twenty times three, which is one hundred and fifty basis points.
- Utility: the intercept is ninety minus twenty-five, sixty-five, so sixty-five plus sixty is one hundred and twenty-five.
- Financial: the intercept is ninety plus thirty-five, one hundred and twenty-five, and the slope is twenty plus twelve, thirty-two.
- Thirty-two times three is ninety-six, so the spread is two hundred and twenty-one.
- The gap between financial and industrial is seventy-one.
- That is the thirty-five from the intercept dummy, plus twelve times three from the slope dummy.
- The exam trap: with a slope dummy in the model, the financial dummy alone is the gap only at zero leverage, not at every leverage.

## 12 When the answer is yes or no

- Outcome is yes or no: coded 1 or 0
- A straight line can predict below 0 or above 1
- Logistic models the log odds instead
- Estimated by maximum likelihood, not least squares

**IN THE LESSON**

- Now a different kind of question.
- Will this borrower default?
- Will this company receive a takeover bid?
- The answer is yes or no, so the dependent variable is itself a dummy, one for the event and zero for no event.
- Tomas's credit team wants the chance that a borrower defaults.
- If he runs an ordinary regression on zeros and ones, the straight line will happily predict a probability of minus ten per cent for a safe borrower and one hundred and twenty per cent for a risky one.
- That is nonsense.
- The fix is logistic regression.
- Instead of modelling the probability directly, it models the log odds, which can be any number at all, and then converts that back to a probability that always stays between zero and one.

## 13 A line escapes; the S-curve cannot

*Chart: Probability of default vs Leverage (debt ÷ EBITDA)*

- p = 1
- p = 0
- Straight line escapes
- Logistic S-curve

**IN THE LESSON**

- Here is the problem drawn.
- Leverage runs along the bottom and the probability of default goes up the side, with the two dashed lines marking zero and one.
- The straight line, the kind least squares would fit, starts below the zero line and finishes above the one line, which is impossible.
- Now the S shaped curve, the logistic curve.
- It rises slowly at low leverage, steeply in the middle, and flattens as it nears one.
- It never crosses either dashed line.
- Notice the trade-off.
- A one-unit rise in leverage always adds the same amount to the log odds, but it moves the probability most in the middle of the S, and hardly at all at the two ends.

## 14 The logistic model and its two conversions

**ln( p ÷ (1 − p) ) = b0 + b1X1 + … + bkXk**
**p = 1 ÷ (1 + e^−(log odds))**

| | |
|---|---|
| **p: probability** | Chance of the event, 0 to 1 |
| **Odds** | p ÷ (1 − p) |
| **Log odds** | Natural log of the odds |
| **e to the b** | Factor the odds are times |

**IN THE LESSON**

- Here is the logistic model.
- Odds are the probability of the event divided by the probability of no event.
- If the chance of default is sixty per cent, the odds are sixty over forty, one and a half to one.
- The log odds are the natural log of that, and the model says the log odds equal b zero plus b one X one and so on: an ordinary line.
- To get back to a probability, work out the log odds, then p equals one over one plus e raised to minus the log odds.
- A slope is the change in the log odds for a one-unit rise in that variable.
- Raise e to the slope, and you have the factor by which the odds are multiplied.

## 15 Example 3: from output to probability

*Log odds = −1.5 + 0.5 LEV − 0.6 COV + 0.8 CYC*

| | | |
|---|---|---|
| **Borrower** | LEV 4, COV 1.5, CYC 1 | |
| **Log odds** | −1.5 + 2.0 − 0.9 + 0.8 | **0.40** |
| **Odds** | e^0.40 | **1.49** |
| **Probability** | 1 ÷ (1 + e^−0.40) | **0.60** |

**IN THE LESSON**

- Example three.
- Tomas's default model uses leverage, interest coverage, and a dummy for cyclical sectors.
- The estimated log odds are minus one point five, plus zero point five times leverage, minus zero point six times coverage, plus zero point eight times the cyclical dummy.
- One borrower has leverage of four, coverage of one and a half, and is in a cyclical sector.
- The log odds are minus one point five, plus two, minus zero point nine, plus zero point eight, which comes to zero point four.
- The odds are e to the zero point four, about one point four nine.
- And the probability is one over one plus e to the minus zero point four, about zero point six oh, or sixty per cent.
- Odds and probability are different numbers.

## 16 Example 4: what a coefficient does

*Start: log odds 0.40, odds 1.49, probability 0.60*

| | | |
|---|---|---|
| **Leverage up by 1** | odds × e^0.5 | **2.46** |
| **New probability** | 2.46 ÷ (1 + 2.46) | **0.71** |
| **Same, not cyclical** | log odds 0.40 − 0.80 | **p = 0.40** |
| **Odds ratio for CYC** | e^0.8 | **2.23** |

**IN THE LESSON**

- Example four, what the coefficients do.
- Start with the borrower at log odds of zero point four and odds of one point four nine.
- If leverage rises by one, the log odds rise by zero point five, so the odds are multiplied by e to the zero point five, which is about one point six five.
- The odds go from one point four nine to two point four six.
- The probability is two point four six over three point four six, which is zero point seven one.
- That is up about eleven points, not zero point five and not sixty-five per cent.
- Now take the same borrower but outside a cyclical sector.
- The log odds fall by zero point eight, to minus zero point four, the odds are zero point six seven, and the probability is forty per cent.
- The cyclical odds ratio is e to the zero point eight, about two point two three.

## 17 Testing the model: the likelihood ratio

**LR = −2 × (LL restricted − LL unrestricted)**
**Compare with chi-square, df = number of restrictions**

| | |
|---|---|
| **LR statistic** | Big value: variables matter |
| **Restricted LL** | Fit with the variables dropped |
| **Unrestricted LL** | Fit with all the variables |
| **Chi-square** | Table value, 5% level |

**IN THE LESSON**

- How do we test whether the model as a whole is any good?
- In least squares we used an F test.
- For logistic regression we use the likelihood ratio test.
- Maximum likelihood gives every model a log-likelihood, a score that is negative, where closer to zero is better.
- The unrestricted model has all the variables.
- The restricted model drops the ones being tested, for instance all three slopes, leaving only the intercept.
- The statistic is minus two times the restricted log-likelihood minus the unrestricted log-likelihood.
- Compare it with a chi-square value whose degrees of freedom equal the number of restrictions.
- A large statistic rejects the restricted model.

## 18 Example 5: is the model jointly significant?

*Intercept only versus all three variables: 3 restrictions*

| | | |
|---|---|---|
| **Restricted log-likelihood** | intercept only | **−250.0** |
| **Unrestricted log-likelihood** | all three variables | **−238.2** |
| **LR statistic** | −2 × (−250.0 + 238.2) | **23.6** |
| **Chi-square, 3 df, 5%** | table value | **7.81** |
| **Decision** | 23.6 above 7.81 | **Reject** |

IN THE LESSON

- Example five.
- Tomas tests all three slopes together.
- The restricted model, with only an intercept, has a log-likelihood of minus two hundred and fifty.
- The full model has a log-likelihood of minus two hundred and thirty-eight point two.
- The statistic is minus two times minus two hundred and fifty minus minus two hundred and thirty-eight point two.
- The bracket is minus eleven point eight, and minus two times that is twenty-three point six.
- There are three restrictions, so we look at the chi-square table with three degrees of freedom.
- The five per cent value, as he reads it, is seven point eight one.
- Twenty-three point six is far above seven point eight one, so he rejects the restricted model.
- The variables are jointly significant.

## 19 Mistakes to avoid so far

- High leverage is not influence: check Cook's D too
- Dummies: n − 1 of them, each a gap from the base group
- Odds are not probability; a slope is in log odds
- Logistic: maximum likelihood and the LR test

IN THE LESSON

- One: treating high leverage as influence.
- A point far out in X that sits on the line changes little, so look at Cook's distance too.
- Two: in a dummy regression, using as many dummies as categories, or reading a dummy coefficient as a level.
- It is always a gap against the base group, and with a slope dummy in the model the intercept dummy is the gap only at zero.
- Three: reading odds as probability.
- Odds of one point four nine are not a probability of one point four nine, and a slope is a change in log odds, not in probability.
- Four: using least squares thinking for logistic regression.
- The model is estimated by maximum likelihood and tested with a likelihood ratio.

## 20 A trend: the series as a function of time

- Linear trend: adds the same amount each period
- Log-linear trend: grows by the same rate each period
- The only independent variable is t, the period number
- Forecast t = last sample period + periods ahead

IN THE LESSON

- Now time.
- Tomas has forty quarters of revenue, in millions of dollars, for a homeware retailer called Fernlea Stores, numbered t equals one for the first quarter up to t equals forty.
- He wants a forecast using nothing but the passage of time.
- Picture two staircases.
- On the first, every step adds the same amount.
- On the second, every step multiplies by the same amount, say three per cent.
- At the start they look alike.
- Later the second one climbs much faster.
- The first is a linear trend and the second is a log-linear trend.
- The only independent variable is t, a simple counter.
- Count carefully: two quarters after quarter forty is quarter forty-two.

## 21 The two trend models

**Linear: y = b0 + b1 × t**
**Log-linear: ln y = b0 + b1 × t, so y = e^(b0 + b1 × t)**
**Growth per period = e^b1 − 1**

| | |
|---|---|
| **t: period number** | Count it carefully |
| **Slope b1** | Amount, or log growth |
| **e to the power** | Turns a log back to units |
| **Growth per period** | Not quite equal to b1 |

IN THE LESSON

- Here are the two models.
- The linear trend says revenue equals b zero plus b one times t.
- The slope is the amount added each period.
- The log-linear trend says the natural log of revenue equals b zero plus b one times t.
- Because the left side is a log, the forecast comes out as a log, and to get back to dollars you raise e to that number.
- So revenue equals e to the power of b zero plus b one t.
- In this model the slope is a growth rate in logs, and the actual growth per period is e to the b one, minus one.
- It is close to b one only for small slopes.

## 22 Example 6: the linear forecast

*Fernlea revenue, $ million, t = 1 to 40*

| | | |
|---|---|---|
| **Fitted linear trend** | revenue = 13.6 + 1.51 × t | |
| **Sample ends** | t = 40 | |
| **Two quarters ahead** | t = 40 + 2 | **t = 42** |
| **Linear forecast** | 13.6 + 1.51 × 42 | **$77.0m** |
| **Wrong t of 41 gives** | 13.6 + 1.51 × 41 | **$75.5m** |

IN THE LESSON

- Example six.
- Tomas fits a linear trend to the forty quarters and gets revenue equals thirteen point six, plus one point five one times t.
- He wants the forecast two quarters after the sample ends.
- The last quarter is forty, so two quarters later is quarter forty-two.
- Put t equals forty-two into the line.
- One point five one times forty-two is sixty-three point four two, and adding thirteen point six gives seventy-seven point oh two.
- The linear forecast is about seventy-seven million dollars.
- If he had used t equals forty-one by mistake, he would have got seventy-five point five one, which is the forecast for the wrong quarter.

## 23 Example 6: the log-linear forecast

*ln(revenue) = 3.00 + 0.035 × t*

| | | |
|---|---|---|
| **Forecast period** | t = 42 | |
| **In logs** | 3.00 + 0.035 × 42 | **4.47** |
| **Back to $ million** | e^4.47 | **$87.4m** |
| **Growth per quarter** | e^0.035 − 1 | **3.56%** |
| **Gap to linear forecast** | 87.36 − 77.02 | **$10.34m** |

IN THE LESSON

- Now the log-linear fit to the same quarters.
- The natural log of revenue equals three point oh oh, plus zero point oh three five times t.
- For t equals forty-two, zero point oh three five times forty-two is one point four seven, and adding three gives four point four seven.
- That is not revenue, it is the log of revenue.
- To get back to dollars, raise e to four point four seven, which is about eighty-seven point four.
- The growth per quarter is e to the zero point oh three five, minus one, which is three point five six per cent.
- So the two models forecast seventy-seven and eighty-seven for the same quarter.
- The log-linear forecast is ten million higher.

## 24 Same start, different futures

*Chart: Revenue ($m) vs Quarter, t*

- Log-linear
- Linear
- End of data

**IN THE LESSON**

- Here is why the choice matters.
- Quarters run along the bottom and revenue goes up the side.
- The straight line is the linear trend and the curve is the log-linear trend.
- Inside the forty quarters of data, shown to the left of the red line, the two stay fairly close, which is why both can look like a good fit.
- Beyond the end of the sample they separate.
- At quarter forty-two the gap is about ten million dollars.
- By quarter sixty the linear trend says about one hundred and four million, while the log-linear trend says about one hundred and sixty-four.
- The further out you forecast, the more the shape decides.

## 25 The U in the linear residuals

*Chart: Residual ($m) vs Quarter, t*

- positive
- negative
- positive
- zero

**IN THE LESSON**

- Back to the choice.
- Which trend should Tomas trust?
- Revenue curves upward, so he plots the residuals of the linear trend, quarter by quarter.
- Quarters run along the bottom and the residual, in millions of dollars, goes up the side.
- The grey line is zero.
- Follow the curve from left to right.
- It starts above zero, dips below it for the long middle stretch, and climbs above zero again at the end.
- So the residuals are positive, then negative, then positive.
- That U-shape tells you the straight line is the wrong shape.
- If revenue compounds at a steady rate, taking logs turns the curve into a straight line against t, and a log-linear trend fits it without that pattern.

## 26 How to choose between the two

*Let the pattern of growth decide*

| What you see | Use |
|---|---|
| Same amount added each period | Linear |
| Same % growth, curves upward | Log-linear |
| Linear residuals shaped like a U | Log-linear |
| A high $R^2$ by itself | No help |

**IN THE LESSON**

- So here is the rule.
- If the series climbs by about the same amount each period, the plot is roughly a straight line and a linear trend fits.
- If it grows by about the same percentage, the plot curves upward, and a log-linear trend is better.
- Your best clue is the residual plot.
- Residuals that are positive at both ends and negative in the middle mean the linear model has the wrong shape.
- And R squared does not settle it.
- Both models can show a high R squared on the same data, so a high R squared can never be the reason you choose one.
- Choose by the pattern of growth, and by what the residuals look like.

## 27 Example 7: what the residuals still say

*40 quarters; Durbin-Watson lower bound 1.44 at 5%*

| | | |
|---|---|---|
| **Linear: R², DW** | R² 0.97 | **DW 0.38** |
| **Log-linear: R², DW** | R² 0.99 | **DW 0.82** |
| **Linear: correlation r** | 1 − 0.38 ÷ 2 | **0.81** |
| **Log-linear: correlation r** | 1 − 0.82 ÷ 2 | **0.59** |
| **Verdict** | both below 1.44 | **Serial corr.** |

IN THE LESSON

- Example seven.
- Tomas checks both fits with the Durbin-Watson statistic, which asks whether each residual is correlated with the one before it.
- A value near two means no serial correlation.
- A value well below two means positive serial correlation.
- Roughly, the statistic is two times one minus r, where r is the correlation between neighbouring residuals.
- The linear trend has an R squared of zero point nine seven and a Durbin-Watson of zero point three eight, so r is about zero point eight one.
- The log-linear trend does better, with an R squared of zero point nine nine and a Durbin-Watson of zero point eight two, so r is about zero point five nine.
- But both are below the lower critical bound of one point four four.
- Even the better model has serially correlated residuals.

## 28 Where trend models fall short

- Assume the past pattern carries on for ever
- Residuals often serially correlated: check Durbin-Watson
- Then t-tests and forecast intervals are unreliable
- Next step: an autoregressive model

IN THE LESSON

- Trend models are simple, which is both their charm and their weakness.
- First, they assume the past pattern goes on.
- Three and a half per cent a quarter says Fernlea can compound for ever, which is rarely believable.
- Second, the residuals of a trend model tend to be correlated with each other.
- When that happens the standard errors are wrong, so the t statistics and the confidence intervals cannot be trusted, even if the coefficients look fine.
- The Durbin-Watson test spots it.
- The usual next step is an autoregressive model, in which the series is explained by its own past values.
- But we cannot do that until we know the series is stable.
- That is the last idea.

## 29 Why a series needs a leash

- Covariance stationary: stable mean, variance, covariances
- A stationary series is pulled back to a long-run level
- Trend, changing variance or unit root: not stationary
- Prices and index levels are rarely stationary

IN THE LESSON

- Picture two walkers.
- One is a dog on a long lead, held by an owner who stands still.
- The dog wanders left and right but always comes back to the same spot.
- The other is a wanderer in an open field, with no owner and nowhere to return to.
- Where they are tomorrow depends on where he is today, and the longer he walks the further he can drift.
- The dog is a stationary series.
- A series is covariance stationary when three things are constant and finite over time: its mean, its variance, and its covariance with its own past values.
- A trending series, a series whose variance keeps changing, and a series with a unit root, which is the wanderer's walk, all fail the test.

## 30 The three conditions

**$E(y_t) = \mu$, the same for every t**
**$Var(y_t) = \sigma^2$, constant and finite**
**$Cov(y_t, y_{t-s})$ depends only on the gap s**

| | |
|---|---|
| **Constant mean** | No trend in the mean |
| **Constant variance** | Same spread at every date |
| **Stable covariances** | Depends on gap, not on date |

IN THE LESSON

- Here are the three conditions written down.
- One: the expected value of the series is the same constant, mu, at every time.
- Two: the variance is a constant, finite number.
- Three: the covariance between the series today and its value s periods ago depends only on the gap s, and not on the calendar date.
- If any of these fails, the series is not covariance stationary.
- Why does it matter?
- Autoregressive models estimate one relationship between a value and its past, and that only means something if it is the same in every part of the sample.

## 31 The level wanders, the change returns

*Chart: Index value vs Time*

- Index level: drifts
- Monthly change: returns to its mean

IN THE LESSON

- Here is a stylised picture.
- Time runs along the bottom.
- The upper line is the level of an equity index.
- It drifts upward, so any average depends on which stretch you pick, and it has no level to return to.
- A shock stays in the level for good.
- The lower line is the monthly change in the same index, measured as the difference in logs.
- Whatever happened last month, it keeps returning to the same average, shown by the grey line, with roughly the same spread.
- That is the mean reversion of a stationary series.
- The usual remedy for a non-stationary series is exactly this: take first differences, or log differences, and model the changes instead of the levels.

## 32 Example 8: a meaningless R squared

*240 months: index level on house price level*

| | | |
|---|---|---|
| **Levels regression** | $R^2$ 0.88, t 41.8 | **Looks great** |
| **Both series** | trend upward | **Not stationary** |
| **Latest index move** | ln(1,284 ÷ 1,262) | **1.73%** |
| **Changes on changes** | $R^2$ 0.005, t 1.1 | **Nothing left** |
| **Conclusion** | two trends, one direction | **Spurious** |

IN THE LESSON

- Example eight.
- Tomas regresses the level of an equity index on the level of a house price index, using two hundred and forty months.
- The R squared is zero point eight eight and the t statistic on the slope is forty-one point eight.
- But both series trend upward, so neither is covariance stationary, and the output may be meaningless.
- He transforms both into log differences.
- The latest index move is the natural log of twelve eighty-four over twelve sixty-two, which is one point seven three per cent.
- Regressing changes on changes gives an R squared of just zero point oh oh five, and a t statistic of one point one.
- The relationship has vanished.
- It was two trends heading in the same direction, a spurious regression.

## 33 Recap: the rules to remember

- Influence: leverage, studentized residual, Cook's D
- Dummies: n − 1, gaps from the base; slope dummy tilts
- Logistic: odds × e^b; p = 1 ÷ (1 + e^−log odds)
- Trend: linear or log-linear, convert logs back, count t
- Stationary or difference it; DW flags serial correlation

IN THE LESSON

- Let's pull it together.
- Influence: leverage above three times k plus one over n, a studentized residual above the critical t, and Cook's distance above zero point five or one.
- Only a point that moves the estimates is influential.
- Dummies: use n minus one, read each as a gap from the base group, and add a slope dummy when the slope can differ.
- Logistic: log odds are linear, the probability is one over one plus e to minus the log odds, and e to the slope multiplies the odds.
- Trends: linear adds an amount, log-linear compounds a rate, convert the log forecast back, and count t carefully.
- Residual patterns and Durbin-Watson judge the fit.
- And a series must be covariance stationary, so difference it if it is not.

## 34 Try three quick questions

1. Three variables, 50 observations. Leverage cut-off?
   **Answer:** 3 × (3 + 1) ÷ 50 = 0.24
2. Log odds are −1.0. What is the probability?
   **Answer:** 1 ÷ (1 + e^1) = 0.27 (the odds are 0.37)
3. ln(sales) = 2.00 + 0.05t. Forecast at t = 20?
   **Answer:** e^3.00 = 20.09, not 3

**IN THE LESSON**

- Now try three quick questions yourself.
- Pause after each one.
- Question one.
- A regression has three independent variables and fifty observations.
- What is the leverage cut-off?
- Three times three plus one, which is twelve, divided by fifty, gives zero point two four.
- Question two.
- The log odds of an event are minus one.
- What is the probability?
- One over one plus e to the one.
- Since e to the one is about two point seven two, the probability is one over three point seven two, about zero point two seven.
- Not the odds of zero point three seven.
- Question three.
- A log-linear trend gives the natural log of sales as two plus zero point oh five times t.
- What is the forecast at t equals twenty?
- Two plus one is three, and e to the three is about twenty point oh nine.
- Not three.