Model misspecification

Quantitative Methods

1 A confident instrument that is wrong

- A regression prints crisp numbers either way
- The printout cannot warn you; you must check
- Damage one: the coefficients are wrong
- Damage two: the precision is wrong

IN THE LESSON

- Imagine a kitchen scale that always shows a crisp, sharp number.
- Now suppose the scale sits on a tilted shelf.
- The display looks just as confident, but every reading is off.
- A regression is like that.
- It prints coefficients, standard errors and an R squared with the same crisp confidence whether the model is sound or not.
- The printout cannot warn you.
- You have to check.
- And there are two kinds of damage to keep apart.
- Sometimes the coefficients themselves are wrong.
- Other times the coefficients are fine, but the measure of how precise they are cannot be trusted.
- We start with the more serious kind.

2 What misspecification means

- Misspecified: wrong variables or wrong form
- Omitted variable: a relevant one left out
- Wrong form: a straight line where a curve belongs
- Wrong scaling, or pooling of different regimes

IN THE LESSON

- A model is misspecified when its form or its variables are wrong.
- Priya's boss asks her to estimate how sensitive a small company fund is to the market.
- She is about to meet four classic ways to go wrong.
- First, an omitted variable: a relevant factor is left out.
- Second, an inappropriate functional form: she fits a straight line where the true relationship curves.
- Third, inappropriate scaling: variables measured in units that make some huge and others tiny, which can create unequal spread or tangled variables.
- Fourth, inappropriate pooling: she mixes data from two different regimes, such as calm years and crisis years, as if they were one.
- We take them in turn.

3 Example 1: the variable left out

Fund return on the market alone (illustrative)

True market beta	the real exposure	**1.00**
True size exposure	left out of the model	**0.50**
Size slope on market	in this sample	**0.40**
Bias	0.50 × 0.40	**0.20**
Reported market beta	1.00 + 0.20	**1.20**

IN THE LESSON

- Here is the case.
- Priya regresses a small company fund's monthly return on the market alone.
- The fund really has a market sensitivity, or beta, of one point zero zero.
- It also has a size exposure of zero point five zero, meaning it gains when small companies beat large ones.
- In her sample, the size factor tends to rise when the market rises.
- If you regress size on the market, the slope is zero point four zero.
- Because she left size out, the market slope has to carry some of its effect.
- The extra amount is the size exposure times the size link to the market: zero point five zero times zero point four zero, which is zero point two zero.
- Her reported beta is one point two zero.

4 Where the extra beta comes from

Market beta

True market beta	**1.00**
Borrowed from size	**0.20**
Reported beta	**1.20**

Biased, and a bigger sample does not cure it

IN THE LESSON

- Here it is as a picture.
- The vertical axis is the market beta.
- The first bar is the truth, one point zero zero.
- The second bar is the extra that the regression borrowed from the size factor, zero point two zero.
- The third bar is what Priya actually sees, one point two zero.
- Now the key point.
- This extra amount is a bias, and it is also inconsistent.
- A bigger sample would not shrink it, because every sample from the same process shows the same borrowed amount.
- More data just makes her more sure of the wrong number.
- Notice too that the leftover errors may look perfectly ordinary, so this mistake hides well.

5 Example 1: what the wrong beta costs

Hedging a 10 million fund with the reported beta

Hedge needed	1.00 × 10 million	**10 million**
Hedge placed	1.20 × 10 million	**12 million**
Over-hedged by	12 − 10	**2 million**
Share of the right size	2 ÷ 10	**20%**

The size exposure stays open

IN THE LESSON

- Why does this matter to an investor?
- Suppose the fund is worth ten million and Priya hedges it by selling market futures in proportion to the beta.
- With the true beta of one point zero zero, she needs to sell ten million of the market.
- Using her reported beta of one point two zero, she sells twelve million.
- That is two million too much, which is twenty per cent more than needed.
- Meanwhile the size exposure she never modelled stays wide open.
- So the hedge is both too big in one place and missing in another.
- A neat result on the screen turned into a real loss of protection.
- The cure here is to add the size factor, and then the market slope comes back to about one.

6 A line where a curve belongs

Chart: Dependent variable vs Independent variable

- True curve
- Fitted line

IN THE LESSON

- Now the second error, the wrong functional form.
- Suppose the true link between a stock's size and its trading cost is a curve that rises fast and then flattens out.
- Look at the picture.
- The horizontal axis is the independent variable, the vertical axis is the dependent variable.
- The green curve is the truth.
- The dashed gold line is the straight line Priya fitted.
- See how the line sits above the curve at both ends and below it in the middle.
- So her errors are not random.
- They are negative at the ends and positive in the middle, a pattern.
- A plot of the residuals against the variable would show a hill shape.
- That pattern is the warning, and the fix is to change the form, for example by taking a logarithm.

7 Four errors and their tell-tale signs

Error and what it leaves behind

Error	Tell-tale sign
Omitted variable	Biased slope
Wrong functional form	Residual pattern
Wrong scaling	Odd spread
Wrong pooling of regimes	Two clusters

IN THE LESSON

- Let us put the four errors side by side, since the exam loves this matching.
- An omitted variable that is correlated with an included one gives a biased and inconsistent slope.
- A wrong functional form leaves a pattern in the residuals, like the bowl we just saw.
- Wrong scaling can create unequal spread in the errors, which is heteroskedasticity, or tangled variables, which is multicollinearity.
- We will meet both shortly.
- Wrong pooling mixes regimes, and it often shows as two separate clusters of points in a residual plot, with a single line forced through both.
- Learn the sign for each, because the sign is often how the question is worded.

8 Prevention, and the R squared trap

- Reason from theory before you estimate
- Keep the model parsimonious
- Study the residuals for patterns and clusters
- Test out of sample; a high R^2 proves nothing

IN THE LESSON

- How do you prevent all this?
- Mostly before you press the button.
- Start with economic reasoning: ask what ought to drive the fund's return, and a size exposure for a small company fund is an obvious answer.
- Be parsimonious, which means use the simplest model that does the job, and do not stuff in variables just to push R squared up.
- After estimating, study the residuals for patterns and clusters.
- And test the model out of sample, which means on data it was not built from, to see if it still holds.
- Finally, the trap: a high R squared does not rule out misspecification.
- Priya's one variable model could have a splendid R squared and still give the wrong beta.

9 Errors with a changing spread

- Homoskedastic: error variance is constant
- Heteroskedastic: error variance changes
- Unconditional: unrelated to the Xs; no real harm
- Conditional: moves with the Xs; the real problem

IN THE LESSON

- Next problem.
- Think of forecasting how much people spend on holidays.
- For a student, you can predict the spend quite closely, within a few dollars.
- For a wealthy traveller, your guess could be off by thousands.
- The size of the miss grows with the size of the thing you are explaining.
- In regression we call a steady spread of errors homoskedasticity.
- A spread that changes is heteroskedasticity, and it comes in two kinds.
- Unconditional means the changing spread has nothing to do with the independent variables.
- Conditional means the spread moves with the independent variables, perhaps growing as they grow.
- Priya now has eighty quarters of data, and she suspects the second kind.

10 Residual spread that fans out

Chart: Residual vs Independent variable

- Upper edge
- Lower edge
- Average error: zero

IN THE LESSON

- Here is the conditional kind, drawn.
- The horizontal axis is an independent variable.
- The vertical axis is the residual, the leftover error.
- The dashed grey line at zero is where the residuals average out.
- Now the two solid lines, the green upper edge and the gold lower edge, mark how far the residuals typically land.
- On the left they hug zero.
- On the right they spread wide apart, like a fan.
- So the average error is zero everywhere, which is why the coefficients stay unbiased, but the typical size of the error grows with the variable.
- Least squares treats every observation as equally reliable, though the ones on the right are far noisier.
- The usual standard error formula assumes one single spread that fits none of them.

11 What it does to inference

- Coefficients stay unbiased
- Standard errors unreliable, usually too small
- t tests and the F test reject too often
- Cure: robust (White-corrected) standard errors

IN THE LESSON

- So what are the consequences?
- With conditional heteroskedasticity the coefficient estimates are still unbiased.
- The line is not tilted.
- What breaks is the standard error.
- It becomes unreliable, and usually too small.
- A standard error that is too small makes a t statistic too big, so the t tests, and also the F test, reject a true null hypothesis too often.
- You see significance that is not really there.
- With the unconditional kind, none of this happens.
- The usual standard errors stay valid, so it is not treated as a serious problem, although robust errors are a safe choice anyway.
- The correction for the conditional kind is to use robust standard errors, also called White corrected standard errors.

12 Detecting it: the Breusch-Pagan test

Breusch-Pagan statistic = n × R^2 vs chi-square with k df

n: observations	Quarters, months or years of data
R^2 of the auxiliary fit	From squared residuals on the Xs
Chi-square critical value	Table value at your level
k: degrees of freedom	Count of independent variables

IN THE LESSON

- To detect conditional heteroskedasticity we use the Breusch-Pagan test.
- The idea is simple.
- If the spread of the errors depends on the independent variables, then the squared residuals should be explained by them.
- So, step one, run your original regression and square the residuals.
- Step two, regress those squared residuals on the same independent variables.
- Step three, take the R squared from that second regression, called the auxiliary regression, and multiply it by n, the number of observations.
- Step four, compare the result with a chi-square critical value, with k degrees of freedom, where k is the number of independent variables.
- If the statistic is bigger, reject the null of no conditional heteroskedasticity.

13 Example 2: running the test

80 quarters, three Xs; slope 0.84, standard error 0.30

Squared residuals on the Xs	auxiliary R^2	**0.12**
Test statistic	80 × 0.12	**9.6**
Critical chi-square	3 df, 5%	**7.81**
Decision	9.6 > 7.81	**Reject**

IN THE LESSON

- Priya has eighty quarterly observations and three independent variables.
- Her key slope is zero point eight four, with an ordinary standard error of zero point three zero.
- She squares the residuals and regresses them on the three variables.
- The R squared of that auxiliary regression is zero point one two.
- The test statistic is eighty times zero point one two, which is nine point six.
- The chi-square critical value with three degrees of freedom at five per cent is seven point eight one, a table value given in this example.
- Nine point six is larger, so she rejects the null.
- She has conditional heteroskedasticity.
- The slope itself is still unbiased, but its standard error cannot be trusted.

14 Example 2: the same slope, a new verdict

Slope 0.84 before and after the correction

OLS t-statistic	0.84 ÷ 0.30	**2.80**
Robust standard error	White-corrected	**0.45**
Robust t-statistic	0.84 ÷ 0.45	**1.87**
Critical t, about	76 df, 5%	**1.99**
Verdict	1.87 < 1.99	**Not significant**

IN THE LESSON

- Now the cure.
- With the ordinary standard error, the t statistic is zero point eight four divided by zero point three zero, which is two point eight zero.
- That looks significant, since it is above the critical value of about one point nine nine.
- Priya now recomputes with White corrected standard errors, and the corrected standard error of the slope is zero point four five.
- The robust t statistic is zero point eight four divided by zero point four five, which is one point eight seven.
- That is below one point nine nine, so the slope is not significant.
- Look at what moved and what did not.
- The slope stayed at zero point eight four.
- Only the verdict on it changed.

15 Three t numbers side by side

t-statistic

OLS t	**2.80**
Robust t	**1.87**
Critical	**1.99**

Same slope, 0.84; only the standard error changed

IN THE LESSON

- Here are those three numbers as bars.
- The vertical axis is the t statistic.
- The first bar is the ordinary least squares t statistic, two point eight zero.
- It stands well above the critical value, so the naive reading is that the slope is significant.
- The second bar is the robust t statistic, one point eight seven.
- It falls short.
- The third bar is the critical value itself, about one point nine nine.
- The bar to beat.
- Notice that the whole change came from the bottom of the fraction.
- The top, the slope, never moved.
- A too small standard error flatters the result, and the robust version removes the flattery.
- This is the Type one error that heteroskedasticity quietly invites.

16 Errors that remember

- Serial correlation: errors linked across time
- Positive: a high error tends to follow a high one
- Negative: errors tend to alternate in sign
- Common in time series such as monthly returns

IN THE LESSON

- Third problem.
- Think of a hot streak in sport.
- A team that surprised its fans last week tends to surprise them again this week.
- Surprises come in runs.
- Errors in a regression can do the same.
- Serial correlation means the errors are correlated across observations.
- It is common in time series, such as monthly returns, because shocks carry forward.
- If a high error tends to follow a high error, that is positive serial correlation.
- If errors tend to flip in sign from one period to the next, that is negative.
- Least squares assumes the errors are unrelated to one another.
- Priya's next regression uses sixty monthly returns on a bond fund, so the assumption is under strain.

17 Runs of the same sign

Chart: Residual vs Month

- Long positive run
- Long negative run

IN THE LESSON

- Here are sixty monthly residuals in date order, smoothed.
- The horizontal axis is the month.
- The vertical axis is the residual.
- The dashed grey line is zero.
- Look at the green line.
- It crosses zero only a few times.
- It stays above for months, then below for months.
- If the errors were unrelated, you would expect a change of sign about every other month, some thirty crossings in all.
- Here there are only a handful.
- So each error carries information about the next one.
- That means sixty of them hold less independent evidence than sixty unrelated errors would.
- The standard error formula counts every one as fresh news, so it overstates how precise the estimates are.

18 What it does to inference

- Standard errors usually too small if positive
- t-statistics inflated; Type I errors rise
- Coefficients stay consistent...
- ...unless a lagged dependent variable is a regressor

IN THE LESSON

- Now the consequences.
- With positive serial correlation, the conventional standard errors are typically too small in the usual time series settings.
- In fact they are unreliable whenever errors are serially correlated.
- Too small a standard error means an inflated t statistic, so you reject the null too often.
- That is a higher chance of a Type one error.
- What about the coefficients?
- Usually they stay consistent.
- There is one important exception.
- If a lagged value of the dependent variable, say last month's return, is one of the regressors, then the errors are tied up with that regressor and the coefficients become inconsistent too.
- Watch for that set-up, because it changes which tests are valid.

19 Detecting it: the Durbin-Watson test

Durbin-Watson ≈ 2 × (1 − r)
below dL: reject; above dU: do not reject

Durbin-Watson statistic	Near 2 means no correlation
r: first-order correlation	Each residual with the one before
Lower bound dL	Below it: reject no correlation
Upper bound dU	Above it: do not reject

IN THE LESSON

- The first test is the Durbin-Watson statistic.
- It tests for first-order serial correlation, meaning each error linked to the one just before.
- A handy approximation is two times one minus r, where r is the correlation between each residual and the previous one.
- With no correlation, r is zero and the statistic is about two.
- With strong positive correlation, r heads toward one and the statistic falls toward zero.
- Tables give two bounds.
- If the statistic is below the lower bound, reject the null of no positive serial correlation.
- If it is above the upper bound, do not reject.
- Between the two bounds, the test is inconclusive.

20 Example 3: running Durbin-Watson

60 months, two Xs; slope 0.62, standard error 0.20

Residual correlation r	first order	**0.40**
Durbin-Watson	2 × (1 − 0.40)	**1.20**
Lower bound dL	60 obs, 2 Xs, 5%	**1.51**
Decision	1.20 < 1.51	**Reject**

IN THE LESSON

- Priya's bond fund regression has sixty monthly observations and two independent variables.
- The slope is zero point six two, with an ordinary standard error of zero point two zero.
- The first-order correlation of the residuals is zero point four zero.
- So the Durbin-Watson statistic is about two times one minus zero point four zero, which is one point two zero.
- That is well below two.
- The table gives a lower bound of one point five one for this sample size and number of variables at five per cent, a value given in this example.
- One point two zero is below one point five one, so she rejects the null of no positive serial correlation.
- Her errors do echo.

21 Example 3: the correction

Slope 0.62 before and after the correction

OLS t-statistic	0.62 ÷ 0.20	**3.10**
Newey-West error	allows for serial correlation	**0.34**
Newey-West t-statistic	0.62 ÷ 0.34	**1.82**
Critical t, about	57 df, 5%	**2.00**
Verdict	1.82 < 2.00	**Not significant**

IN THE LESSON

- Now the cure for serial correlation: Newey-West standard errors, which are consistent when errors are serially correlated and also allow for heteroskedasticity.
- With the ordinary standard error, the t statistic is zero point six two divided by zero point two zero, which is three point one zero.
- Highly significant, apparently.
- With the Newey-West standard error of zero point three four, the t statistic is zero point six two divided by zero point three four, which is one point eight two.
- The critical value is about two point zero zero, so the slope is no longer significant.
- The corrected standard error is one point seven times bigger.
- The estimate never changed.
- The evidence behind it had been counted twice.

22 Two tests, one cure, one trap

What each tool covers

Tool	Reach
Durbin-Watson	One lag only
Breusch-Godfrey	Several lags
With a lagged dependent variable	Durbin-Watson invalid
Newey-West errors	The cure

IN THE LESSON

- The Durbin-Watson test has limits.
- It covers first-order correlation only, one lag.
- The Breusch-Godfrey test is wider.
- You regress the residuals on the original regressors and on p lagged residuals, and test whether the lagged residuals matter.
- That lets you look at several lags at once, which helps with monthly or quarterly data where the link may sit at a seasonal lag.
- Now the trap.
- If a lagged dependent variable is among the regressors, the Durbin-Watson test is invalid.
- Breusch-Godfrey is still fine.
- And the cure, in either case, is Newey-West standard errors.
- Keep that little grid in your head.

23 Four mistakes to avoid

- R^2 for Breusch-Pagan is from the auxiliary fit
- Robust errors never change the coefficients
- Durbin-Watson is invalid with a lagged dependent variable
- A high R^2 does not rule out misspecification

IN THE LESSON

- Before the last problem, four slips that cost marks.
- One.
- In the Breusch-Pagan test, using the R squared of the original regression.
- The right one comes from the auxiliary regression of squared residuals on the independent variables.
- Using the wrong one gives a large meaningless number.
- Two.
- Thinking robust standard errors change the coefficients.
- They never do.
- They correct only the standard errors.
- Three.
- Using the Durbin-Watson test when a lagged dependent variable is in the model.
- It is invalid there.
- Four.
- Believing that a high R squared protects you from misspecification.
- A model can fit nicely and still have the wrong variables, the wrong shape, or standard errors you cannot trust.

24 Variables that move together

- Multicollinearity: Xs highly correlated with each other
- Coefficients stay unbiased
- Standard errors swell; t-statistics shrink
- Model fits well, yet no single X looks useful

IN THE LESSON

- The last problem.
- Imagine two rowers pulling the same boat forward in perfect rhythm.
- The boat speeds up.
- But if you ask how much each rower contributed, you cannot say, because they never pull differently.
- Multicollinearity is this situation in a regression.
- Two or more independent variables are highly correlated with each other.
- The coefficients stay unbiased.
- The trouble is precision.
- The data cannot separate variables that always move together, so the standard errors swell, and the individual t statistics shrink.
- The strange result is a model that fits well as a whole, even though no single variable looks useful.
- Priya's final regression shows it.

25 The telltale output

Homebuilder index, 60 months, three Xs (illustrative)

Output	Value
R^2 and F-statistic	0.62; 30.5
t, mortgage rate change	1.1
t, Treasury yield change	0.9
t, market return	4.2

IN THE LESSON

- Here is Priya's output.
- She regresses sixty monthly returns of a homebuilder index on the change in mortgage rates, the change in a ten year Treasury yield, and the market return.
- Look at the whole model first.
- R squared is zero point six two, and the F statistic is thirty point five, significant at any usual level.
- The model clearly explains a lot.
- Now the individual t statistics.
- The mortgage rate has one point one, the Treasury yield has zero point nine, and only the market return, at four point two, clears a critical value of about two.
- Two rate variables that both matter to homebuilders, and neither is significant.
- A high R squared and a significant F test, with few or no significant t statistics, is the classic sign.

26 Measuring it: the VIF

VIF = 1 ÷ (1 − R^2)
R^2 from regressing that X on the other Xs

VIF: variance inflation factor	Times the variance is inflated
R^2 of one X on the other Xs	How well the others predict it
Above 5	Needs investigation
Above 10	Serious

IN THE LESSON

- To measure the problem we use the variance inflation factor, written V I F, one for each independent variable.
- Here is how it is built.
- Take one of the independent variables and regress it on all the other independent variables.
- Record the R squared of that regression.
- If the others predict it well, that R squared is high.
- Then the V I F is one divided by one minus that R squared.
- It tells you how many times the variance of that coefficient is inflated by the overlap.
- Two rules of thumb apply.
- Above five, the variable needs investigation.
- Above ten, the problem is serious.
- These are rules of thumb, not formal tests.
- Pairwise correlations can miss the problem when three or more variables combine, but the V I F catches it.

27 Example 4: three VIFs

Each X regressed on the other two

Mortgage rate	R^2 0.92: 1 ÷ 0.08	**12.5**
Treasury yield	R^2 0.91: 1 ÷ 0.09	**11.1**
Market return	R^2 0.10: 1 ÷ 0.90	**1.1**
Verdict	two rates above 10	**Serious**

IN THE LESSON

- Priya runs the three helper regressions.
- First, regress the mortgage rate change on the Treasury yield change and the market return.
- The R squared is zero point nine two, which is not surprising, since mortgage rates largely follow Treasury yields.
- The V I F is one divided by zero point zero eight, which is twelve point five.
- Second, regress the Treasury yield on the other two.
- R squared is zero point nine one, so the V I F is one divided by zero point zero nine, which is eleven point one.
- Third, regress the market return on the other two.
- R squared is only zero point one zero, so its V I F is one divided by zero point nine zero, about one point one.
- Both rate variables are above ten, so the verdict is serious.

28 How far each variance is inflated

VIF

Mortgage rate	**12.5**
Treasury yield	**11.1**
Market return	**1.1**

Above 10 is serious; above 5 deserves a look

IN THE LESSON

- Here are the three variance inflation factors as bars.
- The vertical axis is the V I F.
- The tallest bar is the mortgage rate, at twelve point five.
- The Treasury yield is close behind, at eleven point one.
- Both tower over the usual thresholds of five and ten.
- The short bar is the market return, at about one point one.
- A value near one means almost nothing else in the model predicts that variable, so its coefficient is not inflated at all.
- This also explains why the market return was the only significant variable.
- It is the only one standing on its own.
- The two rates are, in effect, one piece of information entered twice.

29 How the VIF climbs with overlap

Chart: VIF vs R^2 of one X on the others

- VIF 5
- VIF 10

IN THE LESSON

- This curve shows why the V I F is nasty at the high end.
- The horizontal axis is the R squared of one variable on all the others, from zero to one.
- The vertical axis is the V I F.
- At low overlap the curve is nearly flat.
- An R squared of zero point six gives a V I F of only two point five.
- Then it climbs.
- At zero point eight the V I F is five, the first rule of thumb, marked on the curve.
- At zero point nine it is ten, the second.
- And as R squared heads to one, the curve shoots upward, because that variable is almost an exact copy of the others.
- Priya's mortgage rate, at zero point nine two, is on the steep part.

30 Example 4: what the inflation costs

Standard error inflated by the square root of VIF

Mortgage rate	√12.5	**3.54 times**
Treasury yield	√11.1	**3.33 times**
Mortgage t if it stood alone	1.1 × 3.54	**about 3.9**

Roughly: the estimate itself is unchanged

IN THE LESSON

- What does a V I F do to a standard error?
- The variance is inflated by the V I F, so the standard error is inflated by its square root.
- For the mortgage rate, the square root of twelve point five is about three point five four.
- Its standard error is about three and a half times what it would be if the variable stood alone.
- For the Treasury yield, the square root of eleven point one is about three point three three.
- Now try this.
- The mortgage rate's t statistic was one point one.
- Shrink its standard error by that factor, holding everything else roughly equal, and the t statistic would be about three point nine.
- That is the significance the overlap was hiding.
- Treat it as a rough guide.

31 Remedies and the traps

- Drop or combine the linked variables
- Use a different proxy, or gather more data
- Pairwise correlations can miss three-way links
- A joint F test shows the group still matters

IN THE LESSON

- What can Priya do?
- The remedies all attack the overlap.
- She can drop one of the linked variables, keeping the mortgage rate or the Treasury yield but not both.
- She can combine them, for instance into one average rate change.
- She can try a different proxy that is less tangled with the others.
- Or she can collect more data, especially from periods when the two rates moved differently.
- Two traps.
- Dropping a variable does not prove it was irrelevant.
- A joint F test on the two rate coefficients would still show that rates as a group matter.
- And the coefficients were never biased, only imprecise.
- The standard errors were honestly saying the data could not separate the two.

32 Recap

- Omitted variable: biased, inconsistent slope
- Heteroskedasticity: Breusch-Pagan; robust errors
- Serial correlation: DW, Breusch-Godfrey; Newey-West
- Multicollinearity: high R^2, weak t; VIF above 5 or 10

IN THE LESSON

- Let us pull it together.
- Misspecification means wrong variables or wrong form.
- An omitted variable correlated with an included one biases the slope and makes it inconsistent, and the cure is theory, parsimony, residual checks and testing out of sample.
- Conditional heteroskedasticity leaves coefficients unbiased but makes standard errors too small.
- Detect it with Breusch-Pagan, n times the auxiliary R squared against chi-square, and correct it with robust errors.
- Serial correlation also shrinks standard errors.
- Use Durbin-Watson for one lag, Breusch-Godfrey for several, and Newey-West errors to correct.
- Multicollinearity inflates standard errors, shows as a high R squared with weak t statistics, is measured by the V I F, and is treated by dropping or combining variables.

33 Try three quick questions

1. 100 obs, 2 Xs, auxiliary R^2 0.09, chi-square 5.99. Conclusion?
   Answer: 100 × 0.09 = 9 > 5.99: reject; conditional heteroskedasticity.
2. Last month's return is a regressor. Which serial correlation test is valid?
   Answer: Breusch-Godfrey; Durbin-Watson is invalid here.
3. A variable regressed on the other Xs gives R^2 of 0.80. Its VIF?
   Answer: 1 ÷ 0.20 = 5: needs investigation.

IN THE LESSON

- Now try three quick questions yourself.
- Pause after each one.
- Question one.
- A regression with one hundred observations and two independent variables has an auxiliary R squared of zero point zero nine.
- The five per cent chi-square value with two degrees of freedom is five point nine nine.
- What do you conclude?
- The statistic is one hundred times zero point zero nine, which is nine.
- That is above five point nine nine, so reject: conditional heteroskedasticity.
- Question two.
- A model includes last month's return as a regressor.
- Which serial correlation test is valid?
- Breusch-Godfrey, not Durbin-Watson.
- Question three.
- A variable regressed on the others gives an R squared of zero point eight zero.
- What is its V I F?
- One divided by zero point two zero is five, so it needs investigation.