CHARTA LEARNING · LESSON NOTES

Multiple regression: building it, testing it and using it to forecast

Quantitative Methods

1 A story: what moves a freight company's shares?

- Marlowe Freight: 60 months of returns
- Guesses: market, diesel price, interest rates
- One input cannot hold all three
- Multiple regression weighs them together

IN THE LESSON

- Meet Nadia Okafor, an analyst at a fictional firm called Fenwick Capital.
- Her committee holds shares in Marlowe Freight, a trucking company, and asks her a simple question.
- What makes this share go up and down?
- She has sixty months of the share's return above the risk free rate.
- A good guess is that the whole stock market matters.
- But Marlowe burns a lot of diesel, so fuel prices probably matter too, and perhaps interest rates.
- One input cannot hold all that.
- She needs a method that weighs several inputs at once, and separates one from another.
- That method is multiple linear regression.

2 Three jobs a regression can do

- Job 1: find which variables relate to Y
- Job 2: forecast Y
- Job 3: test a theory such as a factor model
- Same output, read differently for each job

IN THE LESSON

- Before she touches any data, Nadia must decide what the model is for, because the same output is read differently for each job.
- The first job is to find out which variables are related to the outcome.
- Does diesel matter, and in which direction?
- The second job is to forecast the outcome, to predict next month's return.
- The third job is to test a theory, such as whether a factor pricing model fits the data.
- These sound similar, but they are not.
- A model that tests a theory tells you what the past was made of.
- It does not by itself forecast, because to forecast you would need the future values of every input.

3 The steps, in order

Before fitting

Frame the question → Choose the variables → Is Y continuous?

After fitting

Estimate the model → Check the residuals → Use the model

IN THE LESSON

- Now the order of work, and the order matters.
- Along the top are three decisions made before any fitting.
- First, frame the question: is this explaining, forecasting or testing?
- Second, choose the dependent variable and the independent variables.
- Third, ask whether the dependent variable is continuous.
- A return can take any value, so ordinary least squares fits.
- But if the outcome is a yes or no, such as whether a bond defaults, a straight line is the wrong tool and logistic regression fits better.
- Along the bottom are three actions.
- Estimate the model.
- Then examine the residuals to check the assumptions.
- Only then use the model.
- The step people skip is the middle one.

4 Example 1: which job, which tool?

Three requests, three decisions

Does diesel matter?	identify, Y continuous	**least squares**
Next month's return?	forecast, Y continuous	**least squares**
Dividend cut: yes or no?	Y is yes or no	**logistic**

IN THE LESSON

- Let's practise on three requests that land on Nadia's desk.
- The first: do shares of Marlowe respond to diesel prices?
- That is the identify job, the return is continuous, so least squares.
- The second: what will Marlowe return next month?
- That is the forecast job, also least squares, but notice that she must supply next month's values for every input, which are unknown.
- The third: will Marlowe cut its dividend this year, yes or no?
- The outcome is a yes or no, so ordinary least squares is the wrong choice and logistic regression fits better.
- The trap on exam day is to jump from estimation straight to conclusions, or to read a tested theory as a forecast.

5 Many inputs, one outcome

- Y is explained by k inputs together
- Each input gets its own coefficient
- Least squares: smallest total squared miss
- Fitting together stops one input stealing credit

IN THE LESSON

- Think about how you would estimate the price of a flat.
- Size matters, so does the floor, and so does the distance from the station.
- If you only looked at size, you might wrongly give size credit that really belongs to location, because big flats may sit in the suburbs.
- Multiple regression handles this by fitting all the inputs together.
- Y is the dependent variable, the thing we explain.
- The X variables are the independent variables, the inputs, and there are k of them.
- Least squares picks the set of coefficients that makes the total of the squared misses as small as possible, exactly as in the simple case, only now the line becomes a flat surface.

6 The multiple regression model

Y = b0 + b1X1 + b2X2 + … + bkXk + ε

Y: dependent variable	The thing you explain
b zero: intercept	Y when every X is zero
Slopes b one to b k	One slope for each input
X one to X k: the inputs	The k explanatory variables
Error term	Everything the inputs miss

IN THE LESSON

- Here is the model written down.
- Y equals b zero, plus b one times X one, plus b two times X two, and so on up to b k times X k, plus an error.
- Take it piece by piece.
- Y is what we explain.
- The coefficient b zero is the intercept, the expected value of Y when every input is zero.
- The coefficients b one to b k are the slopes, one per input.
- The inputs X one to X k are the explanatory variables.
- And the error, written with the Greek letter epsilon, is everything the inputs cannot explain.
- For Nadia, Y is Marlowe's excess return, and there are three inputs, so k is three.

7 Nadia's fitted model for Marlowe Freight

Marlowe excess return (%) on three inputs, n = 60

Variable (standard error)	Coefficient
Intercept (0.15)	0.20
Market excess return (0.12)	1.10
Diesel price change (0.08)	−0.30
Yield change, 10-year (0.50)	−0.60

IN THE LESSON

- Here is the output Nadia gets back, from sixty months of data.
- All returns and price changes are in per cent, and the interest rate change is in percentage points.
- The three inputs are the market's excess return, the monthly change in the diesel price, and the monthly change in the ten year yield.
- The fitted equation reads: Marlowe's excess return equals zero point two zero, plus one point one zero times the market, minus zero point three zero times the diesel change, minus zero point six zero times the yield change.
- The market coefficient is positive and large, as we would expect.
- Diesel is negative: dearer fuel hurts a trucker.
- The yield coefficient is also negative.

8 Example 2: what each coefficient means

Each slope is a partial effect: others held constant

Market up 1 point	1.10 × 1	**+1.10**
Diesel up 1 percent	−0.30 × 1	**−0.30**
Market +2 and diesel +5	2.20 − 1.50	**+0.70**
Intercept 0.20	return when all X = 0	**0.20**

Add slopes times changes; never add the intercept to a slope

IN THE LESSON

- Now read the numbers properly.
- The market slope is one point one zero.
- It says that if the market's excess return is one percentage point higher, with diesel and yields held constant, Marlowe's return is expected to be one point one percentage points higher.
- The diesel slope is minus zero point three zero: one percent dearer diesel, with the market and yields held constant, takes about zero point three off the return.
- That phrase, holding the others constant, is the key to every slope.
- Now suppose the market rises two points and diesel five percent, with yields unchanged.
- The expected change is two point two minus one point five, which is plus zero point seven.

9 A simple slope and a partial slope can disagree

Chart: Marlowe excess return (%) vs Change in diesel price (%)

- Diesel alone: +0.14
- Market held fixed: −0.30

IN THE LESSON

- Here is a surprise that the exam loves.
- Suppose Nadia had regressed Marlowe on diesel alone.
- She would have found a slope of plus zero point one four, a positive number, which seems to say that dearer diesel helps the share.
- Look at the picture.
- Along the bottom is the change in the diesel price, up the side is the share's excess return.
- The green line is that simple regression, rising gently.
- The red line is the diesel slope from the full model, falling.
- Why the difference?
- Diesel tends to rise in months when the economy is strong, and the market is strong too.
- Alone, diesel gets credit for the market's lift.
- Holding the market fixed, the true effect shows as negative.

10 What the regression takes for granted

- 1. Linear relationship
- 2. Constant error variance: homoskedasticity
- 3. Errors independent of each other
- 4. Errors normally distributed
- 5. Exogenous inputs: E(ε | X) = 0

IN THE LESSON

- Least squares will hand you coefficients whatever the data look like.
- It does not tell you whether you should believe the tests that come with them.
- Those tests rest on five assumptions.
- One, linearity: the relationship between Y and the inputs is a straight line in the coefficients.
- Two, homoskedasticity: the errors have the same variance everywhere.
- Three, independence: the errors are not linked to each other, so one month's miss tells you nothing about the next.
- Four, normality: the errors follow a bell shaped distribution.
- Five, exogeneity: the inputs are not correlated with the errors.
- Formally, the errors have a zero average whatever the inputs are, so the inputs may be random, but they must not be tied to the errors.

11 Which plot catches which failure

Each picture checks one assumption

Plot of the residuals	Warning sign
Against X: a curve	Not linear
Against fitted: a fan	Unequal spread
In time order: long runs	Serial correlation
Normal Q-Q: bent ends	Fat tails

IN THE LESSON

- The first check is to plot the residuals, the leftover misses.
- Each plot is evidence about one assumption.
- Plot the residuals against each input, and against the predicted values.
- If they bend in a curve, the relationship is not linear.
- If they spread out like a fan that widens to the right, the error variance is not constant.
- That is heteroskedasticity.
- Next plot the residuals in time order.
- Long runs of the same sign, a stretch of positives then a stretch of negatives, suggest serial correlation, errors linked to each other.
- Last, a normal quantile plot, which compares the residuals to a perfect bell shape.
- If the ends bend away from the line, the tails are fatter than normal.

12 Residuals against fitted values: the fan

Chart: Residual vs Fitted return

- Small misses
- Misses widen: a fan

IN THE LESSON

- Here is the fan, drawn.
- Along the bottom are the model's fitted values for Marlowe's return, from low to high.
- Up the side are the residuals, the misses.
- The grey line in the middle is zero, a perfect fit.
- Now the two curves.
- On the left, where fitted returns are low, the residuals hug zero, so the misses are small.
- As we move right, the two curves spread apart like a trumpet, so the misses get bigger.
- That widening is the signature of heteroskedasticity.
- A healthy plot would be an even band of constant width around zero.
- In Nadia's data, the typical miss in the lowest third of fitted values is under one point, and in the highest third nearly three.

13 Example 3: reading Nadia's four plots

Four residual plots for Marlowe

Residuals vs yield change	even band	**clean**
Residuals vs fitted	fan: 0.9, 1.6, 2.7	**unequal spread**
Residuals in date order	runs of 7 same sign	**serial correlation**
Normal Q-Q plot	ends bend away	**fat tails**

Standard errors unreliable, so t-statistics overstated

IN THE LESSON

- Time to read Nadia's four plots together.
- Against the yield change, the residuals form an even band: nothing wrong there.
- Against the fitted values, the band fans out, with typical misses of zero point nine, one point six and two point seven across the three thirds.
- That is heteroskedasticity.
- In time order, the residuals come in runs, seven months of positive misses in a row, then more negative ones: serial correlation.
- The normal plot bends away at both ends: fat tails.
- What do the failures cost?
- The coefficients are usually still unbiased, but the standard errors are unreliable, probably too small, so the t statistics look better than they are.
- And the clean plot proves nothing about the others.

14 Mistakes to avoid so far

- Slope read without holding others constant
- Partial slope assumed equal to simple slope
- One clean plot taken to clear the model
- Estimation straight to conclusions; units ignored

IN THE LESSON

- Let's stop to collect the traps from the first half.
- One.
- Reading a slope without saying the others are held constant, or adding the intercept to a slope.
- The intercept is a separate starting level, not part of any slope.
- Two.
- Assuming a partial slope must match the simple slope on the same variable.
- It can differ, even in sign, when the inputs move together.
- Three.
- Declaring the model fine because one residual plot looks clean.
- Each plot checks one assumption only.
- Four.
- Jumping from estimation to conclusions without examining the residuals, or presenting a tested theory as though it were a forecast.
- And finally, forgetting the units: a slope is in units of Y per unit of X.

15 How much does the model explain?

- SST: total movement in Y
- RSS: the part the regression explains
- SSE: the part left over (the residuals)
- SST = RSS + SSE

IN THE LESSON

- Now suppose the model is built and the residuals look acceptable.
- The next question is how much of Marlowe's movement the three inputs capture.
- The analysis of variance table, or anova, answers this by splitting the total movement into pieces.
- Take every month's return, measure how far it sits from the average return, square it, and add up.
- That total is the total sum of squares, S S T.
- Part of it is explained by the regression: the regression sum of squares, R S S.
- In this notation R S S is the explained part, not the residual part.
- The rest is the sum of squared errors, S S E, the part left unexplained.
- Always, total equals explained plus unexplained.

16 R squared and adjusted R squared

$R^2 = RSS \div SST$
Adj. $R^2 = 1 - [(n - 1) \div (n - k - 1)] \times (1 - R^2)$

RSS: explained	Explained part of the movement
SST: total	Total movement in Y
n: observations	Number of observations
k: number of inputs	Number of inputs
Adjusted R squared	Can fall when a variable is added

IN THE LESSON

- The first score is R squared.
- It is the explained share: R S S divided by S S T.
- Here is its flaw.
- R squared never falls when you add another variable, even a useless one, because least squares could always give the new variable a coefficient of zero.
- So a bigger model always looks at least as good.
- Adjusted R squared fixes this.
- It equals one minus the fraction n minus one over n minus k minus one, times one minus R squared.
- Here n is the number of observations and k the number of inputs.
- Adding a variable costs a degree of freedom, and adjusted R squared falls if the new variable does not earn its place.
- Watch the count: it is n minus k minus one, not n minus k.

17 Example 4: Marlowe's fit, then a fourth input

n = 60, k = 3: SST 400, RSS 260, SSE 140

R squared	260 ÷ 400	**0.650**
Adjusted R squared	1 − (59 ÷ 56) × 0.35	**0.631**
Four inputs: R squared	261 ÷ 400	**0.653**
Four inputs: adjusted	1 − (59 ÷ 55) × 0.3475	**0.627**

R squared rose, adjusted R squared fell

IN THE LESSON

- Here is Nadia's analysis of variance table.
- The total sum of squares is four hundred, the regression sum of squares is two hundred and sixty, and the sum of squared errors is one hundred and forty.
- There are sixty months and three inputs, so the errors have fifty six degrees of freedom.
- R squared is two hundred sixty over four hundred, which is zero point six five zero.
- Adjusted R squared is one minus fifty nine over fifty six, times zero point three five, which is zero point six three one.
- Now she adds a fourth input, and S S E only falls to one hundred and thirty nine.
- R squared rises to zero point six five three, but adjusted R squared falls to zero point six two seven.
- The new variable was not worth its cost.

18 The overall F-statistic

$F = (RSS \div k) \div [SSE \div (n - k - 1)]$

RSS: explained	Explained part of the movement
SSE: unexplained	Unexplained part
k: number of inputs	Number of inputs
n: observations	Number of observations
F-statistic	Compare with the F table

IN THE LESSON

- R squared tells you how much is explained.
- It does not tell you whether that is more than luck would give.
- For that, the overall F statistic tests the null hypothesis that every slope is zero at the same time.
- It is the regression sum of squares divided by k, all over the sum of squared errors divided by n minus k minus one.
- The top is the explained variation per input.
- The bottom is the unexplained variation per remaining degree of freedom.
- A big ratio means the inputs together explain far more than noise would.
- You compare it with a critical value from the F table, with k and n minus k minus one degrees of freedom.
- If F is above the critical value, reject: at least one slope is not zero.

19 Example 5: are the three slopes all zero?

F test with 3 and 56 degrees of freedom

Explained per input	260 ÷ 3	**86.67**
Unexplained per df	140 ÷ 56	**2.50**
F-statistic	86.67 ÷ 2.50	**34.7**
Critical F (3, 56)	table, 5% level	**2.77**

34.7 > 2.77: reject, slopes not all zero

IN THE LESSON

- Let's run the test for Nadia's three input model.
- The numerator is the regression sum of squares, two hundred sixty, divided by three, which is eighty six point seven.
- The denominator is the sum of squared errors, one hundred forty, divided by fifty six, which is two point five.
- Divide: eighty six point seven over two point five is thirty four point seven.
- The critical value from the table, with three and fifty six degrees of freedom, is about two point seven seven.
- Thirty four point seven is far above it, so we reject.
- The slopes are not all zero together.
- Notice what this does not say.
- It does not say each input belongs, or which one matters.

20 Comparing models: AIC and BIC

AIC = n ln(SSE ÷ n) + 2(k + 1)
BIC = n ln(SSE ÷ n) + ln(n)(k + 1)

SSE: unexplained	Unexplained: smaller is better
n: observations	Number of observations
k: number of inputs	Number of inputs
AIC	Forecasting: lower is better
BIC	Fit: lower is better

IN THE LESSON

- Adjusted R squared is one way to compare models.
- Two more are common.
- The Akaike information criterion, A I C, equals n times the natural log of S S E over n, plus two times k plus one.
- The Schwarz Bayesian criterion, B I C, has the same first term, but its penalty is the natural log of n, times k plus one.
- For both, lower is better.
- Each rewards a smaller sum of squared errors and penalises extra variables.
- The B I C penalty is heavier once n is eight or more, so it favours smaller models.
- As a rule, use the A I C when the model is for forecasting, and the B I C when you care about best fit.
- Neither one is a test with a critical value.

21 Example 6: three inputs or four?

n = 60: n ln(SSE ÷ n) is 50.84 for k = 3

AIC, three inputs	50.84 + 2 × 4	**58.8**
BIC, three inputs	50.84 + 4.09 × 4	**67.2**
AIC, four inputs	50.41 + 2 × 5	**60.4**
BIC, four inputs	50.41 + 4.09 × 5	**70.9**

Lower is better: keep the three-input model

IN THE LESSON

- Compare Nadia's two models with these criteria.
- For three inputs, S S E over n is one hundred forty over sixty, which is two point three three.
- Its natural log is zero point eight four seven, times sixty is fifty point eight four.
- Add the A I C penalty, two times four, which is eight: A I C is fifty eight point eight.
- For the B I C, the penalty is the natural log of sixty, four point zero nine, times four, which gives B I C of sixty seven point two.
- With four inputs, S S E is one thirty nine, so the first term is fifty point four one.
- Then A I C is sixty point four and B I C is seventy point nine.
- Both are higher, so both prefer the three input model.
- All three criteria agree here.

22 Testing one coefficient

t = (estimate − hypothesised value) ÷ standard error

Estimated coefficient	From the regression output
Value you test against	Often zero, but not always
Standard error	Reported beside the estimate
Degrees of freedom	n − k − 1
t-statistic	Beyond critical value: reject

IN THE LESSON

- So the model as a whole is worth keeping.
- Now Nadia asks about a single coefficient.
- Does diesel really matter?
- The t statistic is the estimate, minus the value you are testing against, divided by the standard error of the estimate.
- It simply counts how many standard errors the estimate sits away from that value.
- The degrees of freedom are n minus k minus one.
- Two things to remember.
- The value you test against need not be zero.
- If the question is whether a stock's market beta equals one, you test against one.
- And the standard error reported in the output is for that single coefficient only.
- If the t statistic is beyond the critical value, in either direction, reject.

23 Example 7: four single coefficient tests

Critical t, 56 df, 5% two-tailed: about 2.00

Diesel = 0?	−0.30 ÷ 0.08	**−3.75**
Market beta = 1?	(1.10 − 1) ÷ 0.12	**0.83**
Yield = 0?	−0.60 ÷ 0.50	**−1.20**
Intercept = 0?	0.20 ÷ 0.15	**1.33**

Only diesel is significant on its own

IN THE LESSON

- Here are four tests on Nadia's output, using a critical t value of about two point zero zero for fifty six degrees of freedom.
- First, does diesel matter?
- Minus zero point three zero, divided by zero point zero eight, is minus three point seven five.
- That is beyond minus two, so reject: diesel is significant.
- Second, is Marlowe's market beta equal to one?
- Here we test against one, not zero.
- One point one zero minus one, over zero point one two, is zero point eight three.
- That is well inside the range, so we cannot reject a beta of one.
- Third, the yield.
- Minus zero point six zero over zero point five zero is minus one point two.
- Also inside, so it is not significant.
- Fourth, the intercept.
- Zero point two zero over zero point one five is one point three three, also inside.

24 Testing several coefficients together

- Unrestricted model: keeps all k inputs
- Restricted model: drops q of them
- Does the error rise by more than chance?
- Separate t-tests are not the same question

IN THE LESSON

- Now a harder question.
- Nadia's boss suggests that diesel and yields add little, and that a simple market only model would do.
- That is a claim about two coefficients at once.
- Why not just look at the two t statistics?
- Because separate tests are not the same as a joint test.
- The inputs can be linked to each other, and two tests together do not have the confidence level of one.
- The joint test compares two models.
- The unrestricted model keeps every input.
- The restricted model drops the q inputs in question, which forces their coefficients to zero.
- If dropping them makes the fit much worse, the inputs were doing real work.

25 The joint F-test

F = [(SSE restricted − SSE unrestricted) ÷ q] ÷ [SSE unrestricted ÷ (n − k − 1)]

SSE restricted	Error when q inputs are dropped
SSE unrestricted	Error with every input kept
q: coefficients dropped	How many coefficients tested
k: number of inputs	Inputs in the unrestricted model
F-statistic	Compare with F table: q, n − k − 1

IN THE LESSON

- Here is the statistic.
- The numerator is the restricted model's S S E minus the unrestricted model's S S E, divided by q, the number of coefficients dropped.
- That is the extra error per restriction.
- The denominator is the unrestricted S S E divided by n minus k minus one, where k counts the inputs in the unrestricted model.
- Compare the result with the critical F value having q and n minus k minus one degrees of freedom.
- If it is bigger, reject the claim that all the dropped coefficients are zero.
- The most common slip is to put the restricted model's S S E in the denominator.
- The denominator must come from the model that is allowed to be right if the claim is false.

26 Example 8: drop diesel and yield

Restricted SSE 157, unrestricted SSE 140, q = 2

Extra error per restriction	(157 − 140) ÷ 2	**8.50**
Unrestricted error per df	140 ÷ 56	**2.50**
F-statistic	8.50 ÷ 2.50	**3.40**
Critical F (2, 56)	table, 5% level	**3.16**

3.40 > 3.16: reject, at least one coefficient is not zero

IN THE LESSON

- Nadia fits the market only model on the same sixty months.
- Its sum of squared errors is one hundred fifty seven, compared with one hundred forty for the full model.
- She has dropped q equals two coefficients, and the full model has k equals three inputs, so n minus k minus one is fifty six.
- The numerator: one fifty seven minus one forty is seventeen, divided by two, which is eight point five.
- The denominator: one forty over fifty six, which is two point five.
- F is eight point five over two point five, which is three point four.
- The critical value for two and fifty six degrees of freedom is about three point one six.
- Three point four beats it, so we reject: dropping the two inputs hurts.

27 One sample, two verdicts

F-statistic

Correct F	**3.40**
Wrong denominator	**3.03**
Critical value	**3.16**

Rejection: at least one coefficient is not zero

IN THE LESSON

- Look at the bars.
- The first is the correct F statistic, three point four.
- The second is what you get if you slip and use the restricted sum of squared errors in the denominator.
- That is one fifty seven over fifty six, which is two point eight, so the statistic becomes eight point five over two point eight, or three point zero three.
- The third bar is the critical value, three point one six.
- The correct statistic is above the line, so we reject.
- The slipped statistic falls below it, so we would wrongly keep the claim.
- Now what does rejection tell us?
- Only that at least one of the two coefficients is not zero.
- It does not say both are non zero, and it does not say which one.

28 Making a forecast from the fitted model

$\hat{Y} = b0 + b1X1 + b2X2 + \ldots + bkXk$

Forecast of Y	The point forecast
b zero: intercept	Always include the intercept
Every coefficient	Even if not significant
Forecast inputs	Same units as the fitting data

IN THE LESSON

- Now the payoff.
- To forecast, Nadia plugs forecast values of every input into the estimated equation.
- The predicted Y equals b zero hat, plus b one hat times X one, and so on, with the hats showing these are estimates.
- The rule that catches people: use every estimated coefficient, including the intercept, and including any slope that was not significant.
- The equation was estimated with all of them present, so the other coefficients are right only alongside them.
- If you want a model without a variable, you must re-estimate it without that variable.
- Also watch units.
- If the model was fitted with the yield change in percentage points, enter it in percentage points, not as a decimal.

29 Example 9: forecast Marlowe's return

Market −3.0, diesel −5.0, yield +0.4

Intercept		**0.20**
Market	1.10 × (−3.0)	**−3.30**
Diesel	−0.30 × (−5.0)	**+1.50**
Yield	−0.60 × 0.4	**−0.24**
Forecast	0.20 − 3.30 + 1.50 − 0.24	**−1.84**

IN THE LESSON

- Here is a scenario for the committee.
- The market falls three percent, the diesel price falls five percent, and the ten year yield rises zero point four percentage points.
- Build it term by term.
- Start with the intercept, zero point two zero.
- The market term is one point one zero times minus three, which is minus three point three zero.
- The diesel term is minus zero point three zero times minus five, which is plus one point five zero: a fall in a cost is good news.
- The yield term is minus zero point six zero times zero point four, which is minus zero point two four.
- Add them: zero point two minus three point three plus one point five minus zero point two four is minus one point eight four.
- The expected excess return is minus one point eight four percent.

30 Four ways to get the forecast wrong

Forecast for the same scenario, done four ways

Method	Forecast (%)
Correct	−1.84
Diesel sign flipped	−4.84
Intercept forgotten	−2.04
Yield term dropped	−1.60

IN THE LESSON

- Each slip gives a number that looks just as plausible, which is why the arithmetic should be written out.
- The correct forecast is minus one point eight four.
- If Nadia flips the sign of the diesel input, treating the fall as a rise, she gets minus four point eight four, which is three too low.
- If she forgets the intercept, she gets minus two point zero four.
- If she drops the yield term because its t statistic was only minus one point two, she gets minus one point six zero.
- That is wrong too, because the equation without yield was never estimated.
- The fourth slip is units, such as entering a yield change as a decimal when the model uses percentage points.
- Check each step.

31 How sure is the forecast?

Chart: Forecast return (%) vs Market excess return (%)

- Forecast
- Interval flares out
- Input at its average

IN THE LESSON

- A forecast is a point, and the real outcome will land somewhere around it.
- Two sources of uncertainty surround it.
- The first is the error term itself, the scatter that the model never explained.
- The second is sampling error in the coefficients, because they were estimated from only sixty months.
- Look at the picture.
- Along the bottom is the input, here the market return.
- The solid line is the forecast.
- The dashed lines are the interval, and see how they flare out away from the middle, where the input sits at its sample average.
- For Nadia the software reports a forecast standard error of one point seven zero.
- Minus one point eight four, plus or minus two times one point seven, runs from minus five point two four to plus one point five six.

32 Recap: the formulas to remember

$Y = b0 + b1X1 + \ldots + bkXk + \varepsilon$ Adj. $R^2 = 1 - [(n-1) \div (n-k-1)](1 - R^2)$
$F = (RSS \div k) \div [SSE \div (n-k-1)]$ $t = (b - b\ hyp) \div SE$ $\hat{Y} = b0 + \Sigma\ bjXj$

Slopes b one to b k	Partial effect of each input
Adjusted R squared	Penalises extra variables
F-statistic	Overall or joint test
t-statistic	One coefficient, any value
Forecast of Y	Use every coefficient

IN THE LESSON

- Let's pull it together.
- A multiple regression explains Y with k inputs, fitted by least squares, and each slope is a partial effect, holding the others constant.
- Check the assumptions with residual plots before you trust any test.
- R squared is explained over total, and adjusted R squared uses n minus k minus one.
- The overall F compares explained per input with unexplained per degree of freedom.
- A single t tests one coefficient against any value.
- A joint F compares the restricted and unrestricted models, with the unrestricted S S E in the denominator.
- And a forecast uses every coefficient, intercept included.
- Lower A I C and B I C are better.

33 Try three quick questions

1. n = 50, k = 4, R^2 = 0.60. What is adjusted R^2?
   Answer: $1 - (49 \div 45) \times 0.40 = 0.564$
2. b = 1.5, SE = 0.2. Test b = 1 with critical t = 2.
   Answer: t = 0.5 ÷ 0.2 = 2.5: beyond 2, reject
3. $\hat{Y} = 2 + 3X1 - 4X2$. Forecast with X1 = 1, X2 = −1.
   Answer: 2 + 3 + 4 = 9

IN THE LESSON

- Now try three questions yourself, pausing after each.
- Question one.
- A regression has fifty observations and four inputs.
- R squared is zero point sixty.
- What is adjusted R squared?
- One minus forty nine over forty five, times zero point four, is zero point five six four, about zero point five six.
- Question two.
- A coefficient is one point five with a standard error of zero point two.
- Test whether it equals one, with a critical value of two.
- The t statistic is zero point five over zero point two, which is two point five.
- That is beyond two, so we reject.
- Question three.
- A model is Y equals two, plus three times X one, minus four times X two.
- Forecast Y when X one is one and X two is minus one.
- Two plus three plus four is nine.