

Regression and Survey Scales

Holding things fixed, yes or no outcomes, and whether a scale hangs together

Statistics & Data Analysis Fundamentals
Stat.BD

Mentor: Muhtasim Munif Fahim, MSc

CHAPTER 6

Regression and Survey Scales

A regression says how much an outcome changes when one thing changes and the others stay put. This chapter fits one for a number and one for a yes or no outcome. It also builds the stress score the course has used since Chapter 1, and reads a published results table row by row.

In the simulated class, each extra hour of daily screen time goes with about 0.88 more points on the stress score. Among students who sleep the same number of hours, it is about 0.64. Chapter 5 ended on the question this raises: where did the rest go? That chapter could hold residence fixed by splitting the class into three panels. Sleep cannot be held fixed that way, because it is a number with dozens of values, and splitting by it would leave a handful of students in each slice. Regression holds a number fixed without splitting anyone off.

The chapter has five parts. Section 1 draws a straight line through a scatterplot and reads its two numbers. Section 2 adds more predictors, which is what “controlling for” means, and checks what the model assumes. Section 3 turns to an outcome that is yes or no, a CGPA of 3.00 or above, and makes odds ratios concrete. Section 4 builds a score from five questionnaire items and asks whether they hang together, with Cronbach’s alpha. Section 5 reads the logistic regression table of a published Bangladeshi survey and finds a row that disagrees with the paper’s own counts.

The data are the class survey of Chapter 1. Until the real answers are coded, every figure drawn from it comes from a **simulated** stand-in of 120 students, generated by a program from a fixed seed. They are not real students. A few examples are constructed on purpose, with made-up numbers, and are labelled as constructed where they appear.

What this chapter establishes

1. What a regression line is, and how to read its intercept and slope in words, with units.
2. How residuals measure what the line gets wrong, and why the line is the one that makes their squares smallest.
3. What R^2 measures, and how a slope is tested and given an interval.
4. What a coefficient means in a model with several predictors, and why the screen-time slope shrank once sleep was added.
5. How to enter a categorical predictor with dummy variables.
6. What a linear regression assumes, how to check it with a residual plot, and what the variance inflation factor warns about.
7. Why “controlling for” a variable is not a cure for confounding.
8. Why a yes or no outcome needs logistic regression, and how odds, log odds and odds ratios fit together.

9. Why an odds ratio multiplies the odds, not the probability, and why it is not “times as likely”.
10. How many predictors a logistic model can carry.
11. How to score a scale with a reverse-worded item, and what Cronbach’s alpha does and does not show.
12. How to read a published regression table, and how to check a row against the paper’s own counts.

Notation

y, x	the outcome and a predictor
$\hat{y}$	the outcome predicted by the model (“ y -hat”)
a, b	the intercept and the slope of a line
$b_1, b_2, \dots$	the coefficients of several predictors
k	the number of predictors; in Section 4, the number of items
e	a residual, observed minus predicted
R^2	the share of the variation in y that the model accounts for
SE, t , p	as in Chapter 4
VIF	the variance inflation factor
p	in Section 3, also a probability; the meaning is clear from context
odds, OR, RR	odds, an odds ratio and a risk ratio
$\ln, e^b$	the natural logarithm, and its inverse
EPV	events per variable
α	Cronbach’s alpha (not the significance level of Chapter 4)
s^2	a variance, as in Chapter 2

Contents

What this chapter establishes	1
Notation	2
1 Simple linear regression	6
1.1 From a correlation to a line	6
1.2 Intercept and slope, in words	6
1.3 Using the line to predict	7
1.4 Residuals	7
1.5 Why this line? Least squares	8
1.6 R^2 : how much of the spread the line accounts for	9
1.7 Is the slope more than luck?	10
1.8 Reading the coefficient table	10
1.9 Three ways a regression line misleads	11
2 Multiple regression	12
2.1 Adding a second predictor	12
2.2 What “holding sleep fixed” looks like	12
2.3 Why the slope shrank	13
2.4 Four predictors at once	14
2.5 Each coefficient in one sentence	14
2.6 Predicting for one student	15
2.7 R^2 , adjusted R^2 , and the F test	16
2.8 A categorical predictor: dummy variables	17
2.9 What a linear regression assumes	17
2.10 Checking the assumptions: the residual plot	18
2.11 When predictors overlap: the variance inflation factor	19
2.12 Controlling for is not magic	19
2.13 Reporting a multiple regression	20
3 Logistic regression	20
3.1 When the outcome is yes or no	20
3.2 Why a straight line fails	20

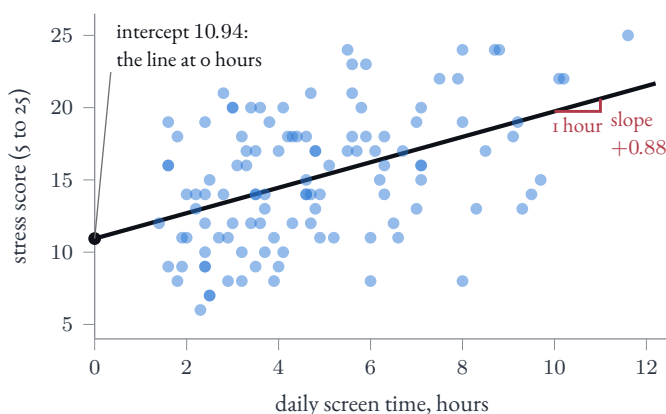
3.3	Odds and the odds ratio	21
3.4	Probability, odds and log odds	21
3.5	The logistic curve	22
3.6	Reading the logistic output	23
3.7	An odds ratio multiplies the odds, not the probability	24
3.8	An odds ratio is not “times as likely”	24
3.9	How many predictors can a logistic model carry?	25
3.10	Reporting a logistic regression	25
4	Survey scales and Cronbach’s alpha	26
4.1	Five statements, one score	26
4.2	How the class answered	26
4.3	Scoring one student	27
4.4	Sum or average, and what the score can be treated as	27
4.5	Do the items move together? Item–total correlations	28
4.6	Cronbach’s alpha	28
4.7	Forget to reverse s4	29
4.8	Alpha if an item is deleted, and alpha with fewer items	29
4.9	What alpha does not tell you	30
4.10	Reporting a scale	30
5	Reading a published regression table	31
5.1	Islam 2020, Table 3	31
5.2	“1.8 times more likely to be depressed”?	31
5.3	Eight intervals, and what they can and cannot say	32
5.4	Check a row against the counts: living with family	33
5.5	Writing a finding from a regression table	34
6	Fitting a regression: the procedure	34
7	Chapter summary	36
8	Key terms	37
9	Exercises	38
10	Solutions	42

11	Our arithmetic and constructed numbers	46
12	References and further reading	46
13	For students in the Stat.BD course	47

1 Simple linear regression

1.1 From a correlation to a line

Chapter 5 measured how closely stress and screen time follow a straight line: $r = 0.45$. Regression goes one step further and draws the line. Figure 1 shows it, through the same cloud of students.



Regression

a model that predicts an outcome from one or more predictors, and says how the prediction changes as each predictor changes

Figure 1: Stress score against daily screen time, with the regression line. The dark dot is the intercept; the red step is one hour's rise. **Simulated** class data, $n = 115$ students who gave both answers.

The line has an equation:

$$\hat{y} = a + bx, \quad \widehat{\text{stress}} = 10.94 + 0.88 \times \text{screen hours} \quad (1.1)$$

- $\hat{y}$ the predicted outcome; the hat means “predicted by the line”
- a the *intercept*: the predicted outcome when x is zero
- b the *slope*: the change in the predicted outcome for each one-unit increase in x
- x the predictor

The equation is a recipe. Start at 10.94, and add 0.88 for every hour of screen time.

Correlation had no direction: the r of stress with screen time is the r of screen time with stress. Regression has one. Stress is the *outcome*, on the vertical axis, and screen time is the *predictor*, across. Choosing which is which is the first decision in any regression. Section 1.9 shows that the other choice gives a different line.

Outcome, predictor

the variable a model predicts (also called the dependent or response variable), and a variable used to predict it (independent or explanatory)

1.2 Intercept and slope, in words

A slope is only meaningful with its units and a sentence around it:

The slope in words

For each extra hour of daily screen time, the predicted stress score is 0.88 points higher. The units are stress points per hour; measure screen time in minutes and the same line has a slope of $0.88 \div 60 = 0.015$ points per minute.

The intercept, 10.94, is the predicted stress score at zero hours of screen time. Nobody in the class reports zero; the smallest answer is 1.4 hours. The intercept anchors the line, but here it describes no real student. That is common, and it is not a fault: an intercept is often outside the data, and is then not worth interpreting.

“Screen time raises stress by 0.88”?

The slope compares *different* students who differ by an hour of screen time. It does not say what would happen if one student added an hour. That needs an experiment. Everything in Chapter 5 about association and causation applies to a slope from a survey, however precisely it is estimated.

1.3 Using the line to predict

To predict, put a value of x into equation (1.1).

Example 1.1 • Predicted stress at 4 and at 8 hours

Screen time	Predicted stress
4 hours	$10.94 + 0.88 \times 4 = 14.46$
8 hours	$10.94 + 0.88 \times 8 = 17.98$

Four more hours, $4 \times 0.88 = 3.52$ more points. The difference between two predictions is always the slope times the difference in x .

A prediction is the *average* stress of students with that screen time, not a forecast for one person. Two students with 4 hours each can sit far apart, as the scatter in Figure 1 shows.

Stay inside the data

The class runs from 1.4 to 11.6 hours of screen time. At 20 hours the line predicts $10.94 + 0.88 \times 20 = 28.5$, on a scale that stops at 25. A straight line goes on for ever; the data do not. Predicting outside the range of the data is called *extrapolation*, and a line has no way to warn against it.

1.4 Residuals

Every student has a residual: the vertical gap between their point and the line. It is what the line gets wrong about them.

Residual

observed outcome minus the outcome the model predicts

$$e = y - \hat{y} = \text{observed} - \text{predicted} \quad (1.2)$$

e the residual for one person

$y, \hat{y}$ their observed and predicted outcomes

Example 1.2 • One student's residual

Student 5 reports 6.0 hours of screen time. The line predicts

$$10.94 + 0.88 \times 6 = 16.22.$$

Example 1.2 • continued

Their stress score is 11, so the residual is $11 - 16.22 = -5.22$. Their stress is lower than their screen time would suggest, for reasons the line knows nothing about. Figure 2 draws it.

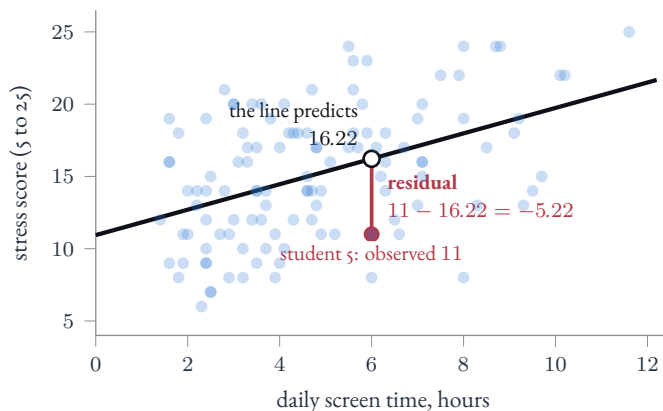
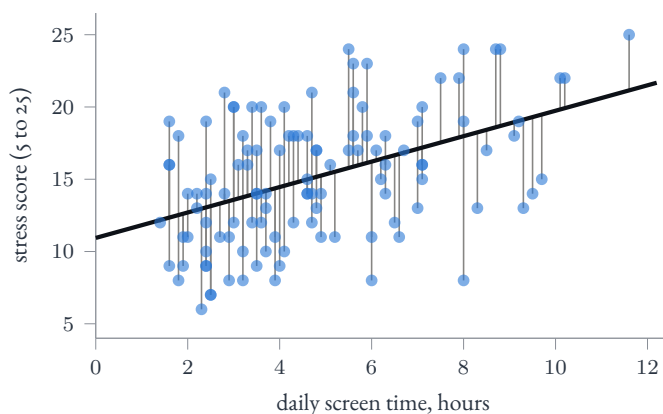


Figure 2: One student's residual: the gap between the observed score (red) and the line's prediction (open circle). **Simulated** class data.

A negative residual means less stressed than the line expects; a positive one, more. Residuals return three times in this chapter: to choose the line, to measure how well it fits, and to check its assumptions.

1.5 Why this line? Least squares

Many lines could be drawn through the cloud. Figure 3 shows every residual of the regression line as a grey stick. Tilt or shift the line and some sticks shrink while others grow. The regression line is the one that makes the sum of the squared residuals, $\sum e^2$, as small as possible. Hence its name: the *least-squares* line.



Least squares

the method that chooses the line making the sum of squared residuals as small as possible

Figure 3: Every residual as a vertical stick. No other straight line makes the sum of their squares smaller. **Simulated** class data.

Why square?

For the same reason as the standard deviation in Chapter 2: squares stop positive and negative gaps from cancelling, and they punish big misses more than small ones. The second property has a cost. One point far from the rest has a huge squared residual, so it can pull the line towards itself. That is Chapter 5's outlier problem again.

Software finds the least-squares line with a formula, and nobody tries lines by hand. The formula needs only the summaries of Chapters 2 and 5:

$$b = r \times \frac{s_y}{s_x}, \quad a = \bar{y} - b\bar{x} \quad (1.3)$$

r the correlation between x and y
 s_x, s_y the standard deviations of x and y
 $\bar{x}, \bar{y}$ their means

Example 1.3 • The slope and intercept from five summaries

For the 115 students in the model, screen time has mean 4.78 hours and SD 2.32; stress has mean 15.15 and SD 4.53; and $r = 0.45$.

$$b = 0.45 \times \frac{4.53}{2.32} = 0.88, \quad a = 15.15 - 0.88 \times 4.78 = 15.15 - 4.21 = 10.94.$$

Both agree with the software's line. The second formula also says the line always passes through the point of means, $(\bar{x}, \bar{y})$.

Equation (1.3) shows how a slope and a correlation differ. The correlation has no units and runs from -1 to $+1$. The slope is the correlation rescaled into the units of the outcome per unit of the predictor.

1.6 R^2 : how much of the spread the line accounts for

Chapter 5 promised that r^2 would reappear. The share of the variation in the outcome that the model accounts for is called R^2 :

$$R^2 = 1 - \frac{\sum e^2}{\sum (y - \bar{y})^2} \quad (1.4)$$

$\sum e^2$ the sum of squared residuals: the misses left after using the line

$\sum (y - \bar{y})^2$ the sum of squared distances from the mean: the misses if every student were predicted to have the mean score

R^2
 the share of the variation in the outcome that the model accounts for; with one predictor it equals r^2

Example 1.4 • R^2 for stress on screen time

The squared residuals add to 1858.9. The squared distances from the mean add to 2334.5.

$$R^2 = 1 - \frac{1858.9}{2334.5} = 1 - 0.796 = 0.204.$$

With one predictor, $R^2 = r^2$: $0.45^2 \approx 0.20$. Chapter 5's r and this R^2 are the same information on two scales.

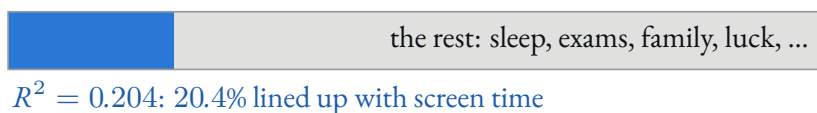


Figure 4: R^2 as a share of the spread in stress. **Simulated** class data.

Figure 4 draws the share. About a fifth of the spread in stress moves in step with screen time. For one question about something as tangled as stress, that is not small. A second way to judge the fit is often more honest. The residuals have a standard deviation of 4.06 points, against 4.57 for stress around its own mean. Knowing a student's screen time narrows the guess, but not by much.

What R^2 does not say

R^2 says nothing about whether a straight line is the right shape, or whether the relationship is causal. Anscombe's four datasets in Chapter 5 all have the same r , and so the same R^2 .

1.7 Is the slope more than luck?

The slope is an estimate from one class. Another class of the same size would give a different line. The machinery of Chapters 3 and 4 applies to b exactly as it applied to a mean. The standard error measures how much b would vary from sample to sample. The null hypothesis for a slope is that, among students like these, it is zero: screen time predicts nothing.

$$t = \frac{b}{SE(b)}, \quad 95\% \text{ CI} = b \pm t^* \times SE(b), \quad \text{df} = n - 2 \quad (1.5)$$

$SE(b)$ the standard error of the slope, from software

t^* the t value that leaves 2.5% in each tail, on $n - 2$ degrees of freedom

df two are used up, one by the intercept and one by the slope

Example 1.5 • Testing the screen-time slope

$b = 0.88$, $SE = 0.164$, $n = 115$, so $\text{df} = 113$ and $t^* = 1.98$.

$$t = \frac{0.88}{0.164} \approx 5.38, \quad 0.88 \pm 1.98 \times 0.164 = 0.88 \pm 0.32 = 0.56 \text{ to } 1.20.$$

The interval leaves out 0 and $p < 0.001$. (t is from the unrounded values.)

Say both

Between 0.56 and 1.20 stress points per hour, with $p < 0.001$. The interval says how big the slope plausibly is; p says only that it is probably not zero. This is the reporting habit of Chapter 4: the interval first, then p .

1.8 Reading the coefficient table

Software prints a regression as a table of coefficients. Table 1 is the one for stress on screen time, in the layout most packages use. Every regression table in every paper is a version of it.

Table 1: Stress on screen time. $n = 115$; $R^2 = 0.204$; residual SD 4.06. **Simulated** class data.

	b	SE	t	p	95% CI
(Constant)	10.94				
Screen hours	0.88	0.164	5.38	< 0.001	0.56 to 1.20

SPSS calls the intercept “(Constant)” and the coefficient “B”.

Read it in this order. First, find the predictor’s row, not the constant’s. Second, put b in a sentence with its units. Third, read the interval: how big, plausibly. Fourth, and last, read p . The constant’s row usually carries its own test as well. That test asks whether the line passes through zero at zero hours, which almost nobody needs to know; Table 1 leaves it blank.

In a report

Each extra hour of daily screen time went with 0.88 more stress points (95% CI 0.56 to 1.20; $n = 115$).

1.9 Three ways a regression line misleads

Reading the slope as strength. The slope depends on units. Per minute of screen time the slope is 0.015; per hour it is 0.88. The line, the data and r are the same. Strength is measured by r or R^2 , which have no units; size is measured by b , in the outcome’s units. Never compare slopes of predictors measured in different units.

Swapping the roles. Least squares minimises the misses in the outcome, which are vertical. Regress screen time on stress instead and it minimises horizontal misses, giving a different line: 0.23 hours of screen time per stress point. That is not $1 \div 0.88$. The two slopes multiply to r^2 : $0.88 \times 0.23 \approx 0.202$. The two lines coincide only when r is $+1$ or -1 . Decide the outcome before fitting.

Reading a curve as a line. A straight line fitted to a curved relationship gives a slope for a shape that is not there. Anscombe’s second panel in Chapter 5 is the example. Plot first, and check the residuals (section 2.10).

And the one from Chapter 5

A slope from a survey is still an association. Chance, a common cause and reverse causation all remain possible, however small the p -value.

2 Multiple regression

2.1 Adding a second predictor

Stress may go with screen time partly because students who spend longer on screens sleep less. To separate the two, both go into one model:

$$\widehat{\text{stress}} = 22.87 + 0.64 \times \text{screen hours} - 1.60 \times \text{sleep hours} \quad (2.1)$$

Table 2 sets it beside the line of section 1. Both models use the same 115 students, who also answered the questions used in the next section. If the two models used different students, their slopes could differ for that reason alone.

Table 2: Stress on screen time, without and with weeknight sleep. **Simulated** class data, the same 115 students in both.

	Screen alone	Screen and sleep
(Constant)	10.94	22.87
Screen hours	0.88	0.64
Weeknight sleep, hours		−1.60

The equation has grown a term, and the meaning of every coefficient has changed. In a multiple regression, each coefficient carries a condition:

Coefficients with their conditions

Screen time: each extra hour goes with 0.64 more stress points, *among students who sleep the same number of hours.*

Sleep: each extra hour goes with 1.60 fewer points, *among students with the same screen time.*

Multiple regression

a regression with more than one predictor; each coefficient is the change in the outcome per unit of its predictor, with the others held fixed

That condition is what “controlling for”, “adjusting for” and “holding fixed” mean. A coefficient in a multiple regression answers one question: what if this predictor differed and the others in the model did not?

2.2 What “holding sleep fixed” looks like

Equation (2.1) is one line for each amount of sleep. Put in 7 hours and it becomes $22.87 - 1.60 \times 7 + 0.64 \times \text{screen}$, a line in screen time alone. Figure 5 draws three of them.

Moving along one coloured line changes screen time with sleep fixed: 0.64 points an hour. Moving from one line to the next changes sleep with screen time fixed: 1.60 points an hour. At 5 hours of screen time, a student who sleeps 6 hours and one who sleeps 8 are predicted to differ by $2 \times 1.60 = 3.20$ points.

The grey line is steeper because, as screen time rises, students also tend to sleep less. Moving along it, a student drifts from the 8-hour line towards the 6-hour line, so its slope mixes both movements. That the coloured lines are parallel is an assumption of this model, not a finding.

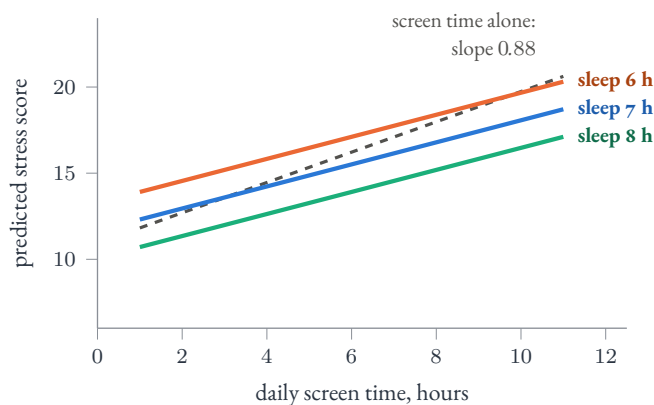


Figure 5: The two-predictor model drawn as one line for 6, 7 and 8 hours of sleep. The lines are parallel, each with slope 0.64. The dashed grey line is the screen-only model, which lets sleep vary too. **Simulated** class data.

A model can let the slope of screen time differ with sleep; that is called an interaction and is beyond this course.

2.3 Why the slope shrank

Chapter 5 described a common cause: a third variable related to both of the others. Sleep plays that part here, although with the arrows arranged differently (Figure 6). Screen time goes with less sleep, and less sleep goes with more stress. With screen time alone in the model, it takes the credit for both paths.

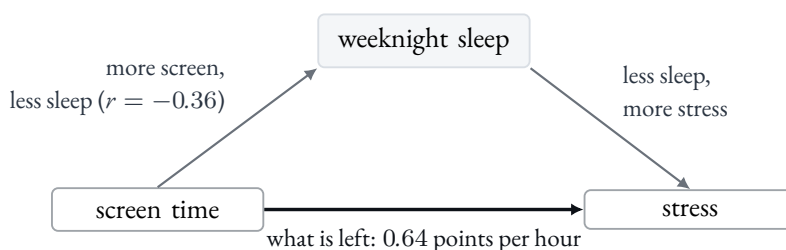


Figure 6: One story that fits the numbers: screen time goes with less sleep, and less sleep with more stress. **Simulated** class data.

The borrowing can be measured. For the same students, the simple slope equals the adjusted slope plus a borrowed part:

$$b_{\text{alone}} = b_{\text{adjusted}} + b_{\text{sleep}} \times \delta \quad (2.2)$$

b_{alone} the screen-time slope with no other predictor

b_{adjusted} the screen-time slope with sleep in the model

b_{sleep} the sleep coefficient in that model

δ the slope of sleep on screen time: how much sleep changes per extra hour on screens

Example 2.1 • Where the 0.24 went

Among the 115 students, each extra hour of screen time goes with -0.147 hours of sleep, so $\delta = -0.147$.

$$b_{\text{sleep}} \times \delta = (-1.60) \times (-0.147) \approx 0.24, \quad 0.64 + 0.24 = 0.88.$$

The simple slope is back. The part screen time borrowed is the path through sleep: screen time's link with sleep, times sleep's link with stress.

If the arrows point the other way

Suppose stress keeps students awake and on their phones. The same three correlations arise, and so do the same slopes. Regression measures how the variables move together; it cannot tell which way the arrows run. Figure 6 is a story brought to the data, and the data fit other stories.

2.4 Four predictors at once

Chapter 1 promised to adjust for private tuition: students who tutor may have more screen time and more stress, for reasons of money and time. Study hours is an obvious further candidate. Table 3 adds both.

Table 3: Stress on four predictors. $n = 115$; $R^2 = 0.296$, adjusted $R^2 = 0.271$; residual SD 3.86. **Simulated** class data.

Predictor	b	SE	t	p	95% CI
(Constant)	22.44	3.541	6.34	< 0.001	15.42 to 29.46
Screen hours	0.65	0.169	3.85	< 0.001	0.32 to 0.98
Weeknight sleep, hours	-1.51	0.446	-3.38	0.001	-2.39 to -0.62
Gives tuition (1 = yes)	0.53	0.739	0.72	0.474	-0.93 to 1.99
Study hours	-0.18	0.295	-0.62	0.534	-0.77 to 0.40

Every row is read like Table 1, with one addition: “holding the other three fixed”. Two rows have intervals that include 0, tuition and study hours. The screen-time slope is now 0.65, close to the 0.64 it had with sleep alone. Adding tuition and study hours barely moved it, so they are doing little of the explaining.

2.5 Each coefficient in one sentence

A coefficient is reported as a sentence with four parts: the number, its units, its interval and its condition.

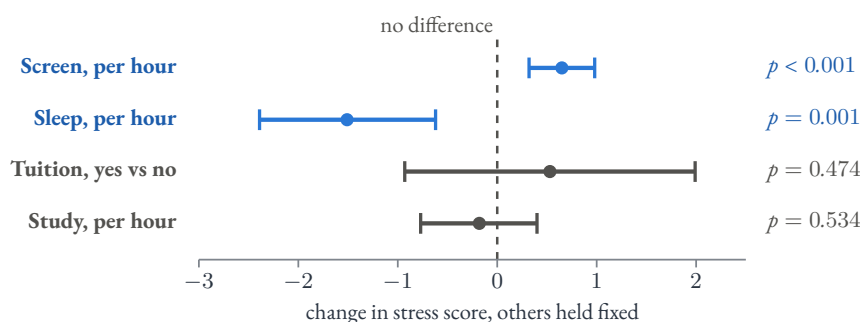
Screen time	Each extra hour goes with 0.65 more stress points (95% CI 0.32 to 0.98), with sleep, tuition and study hours held fixed.
Weeknight sleep	Each extra hour goes with 1.51 fewer points (95% CI -2.39 to -0.62), the others fixed.
Tuition	Tutors score 0.53 points higher, but the interval runs from -0.93 to 1.99: no clear difference either way.
Study hours	-0.18 points per hour, interval -0.77 to 0.40: no clear relationship.
Constant	A student with no screen time, no sleep, no tuition and no study. Nobody.

Tuition is coded 0 or 1, so its coefficient is a difference between two groups rather than a slope per unit. The arithmetic is the same; the sentence changes.

“Tuition has no effect on stress”

The interval allows tutors to be 0.93 points lower or 1.99 higher. “The data could not tell” is the honest reading. Chapter 4 made the point: a wide interval around zero is not evidence of zero.

Figure 7 draws Table 3 as a coefficient plot, which many papers now use instead of a table.



Coefficient plot

a chart of each coefficient as a dot with its confidence interval as a bar

Figure 7: The four coefficients of Table 3 with their 95% intervals. Blue intervals leave out 0; grey ones include it. **Simulated** class data.

Compare each dot with zero, not with the others

Each row is in its own units: per hour of sleep, for tutors against non-tutors. Saying that sleep “matters more” than tuition because its dot is further from zero compares unlike things. Measure sleep in minutes and its dot would sit almost on the zero line.

2.6 Predicting for one student

A prediction from a multiple regression is a weighted sum: multiply each value by its coefficient and add.

Example 2.2 • A student who tutors

A student with 5 hours of screen time and 6 of sleep, who tutors and studies 2 hours a day:

Example 2.2 • continued

Constant	22.44
Screen: 0.65×5	+3.25
Sleep: -1.51×6	-9.06
Tuition: 0.53×1	+0.53
Study: -0.18×2	-0.36
Predicted stress	16.80

The unrounded coefficients give the same 16.80. Change only the sleep, to 8 hours, and the prediction falls by $2 \times 1.51 = 3.02$ points, whatever the other values are. That is “holding fixed” in arithmetic: change one value and only one term moves.

A 0/1 predictor switches its term on or off. For a student who does not tutor, the tuition line becomes $0.53 \times 0 = 0$, and the prediction is $16.80 - 0.53 = 16.27$.

2.7 R^2 , adjusted R^2 , and the F test

With four predictors, R^2 rises from 0.204 to 0.296. But R^2 never falls when a predictor is added, even a column of random numbers. Least squares can always use a new variable to shave a little off the residuals. Adjusted R^2 charges for each predictor:

$$R^2_{\text{adj}} = 1 - (1 - R^2) \times \frac{n - 1}{n - k - 1} \quad (2.3)$$

n the number of people in the model

k the number of predictors, not counting the constant

Adjusted R^2

R^2 with a charge for each predictor, so that a useless predictor lowers it

Example 2.3 • Adjusted R^2 for the four-predictor model

$n = 115, k = 4, R^2 = 0.296$:

$$1 - (1 - 0.296) \times \frac{114}{110} = 1 - 0.704 \times 1.036 \approx 0.271.$$

Software, from the unrounded R^2 , prints 0.271. For screen time alone, adjusted R^2 is 0.197. The three extra predictors earned more than their charge, though almost all of the gain is sleep's.

The F test asks one question of the whole model: could all four slopes be zero together?

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}, \quad \text{on } k \text{ and } n - k - 1 \text{ df} \quad (2.4)$$

 F test

a single test of whether all the slopes in a model could be zero together

From the rounded R^2 , $F = (0.296/4) \div (0.704/110) \approx 11.56$. Software prints $F = 11.59$ on 4 and 110 df, $p < 0.001$. The null hypothesis is rejected, but the test does not say *which* predictors matter; tuition and study hours on their own do not. This is the regression version of Chapter 4's ANOVA F : one test for the whole model. It often comes first in software output, and it is rarely the interesting line.

2.8 A categorical predictor: dummy variables

Residence has three groups: hall, mess and family home. It cannot enter a regression as one column coded 1, 2, 3, because those codes are names, and a slope for them would mean nothing. Chapter 5 listed that as its first wrong choice. Instead, residence becomes two 0/1 columns, called *dummy variables* (Table 4).

Table 4: Residence coded as two dummy variables, with hall as the reference group, and their coefficients in a model of stress on screen time and residence. **Simulated** class data.

	Mess	Home		<i>b</i>	<i>p</i>
Hall (reference)	0	0	(Constant)	10.47	
Mess	1	0	Screen hours	0.87	
Family home	0	1	Mess vs hall	0.81	0.356
			Home vs hall	0.90	0.361

Dummy variable

a 0/1 column marking membership of one group; a variable with g groups needs $g - 1$ of them

Three groups need two columns. The group with no column of its own is the *reference group*, and each dummy coefficient is a gap from it, at the same screen time.

Reference group

the group every other group is compared with; its dummy columns are all 0

Example 2.4 • Three students with 5 hours of screen time

Hall	$10.47 + 0.87 \times 5 = 10.47 + 4.35 = 14.82$
Mess	$14.82 + 0.81 = 15.63$
Home	$14.82 + 0.90 = 15.72$

Mess students are predicted to score 0.81 points above hall students with the same screen time, and home students 0.90 above. Neither gap is distinguishable from zero. Once screen time is held fixed, residence adds little in these data. That was Chapter 1's other promised adjustment.

Software builds the dummies itself when told that a variable is categorical, and lets the analyst choose the reference group. Choose one that is large and makes a natural comparison. A variable with five faculty groups needs four dummy columns.

2.9 What a linear regression assumes

The table and its intervals are trustworthy only if four conditions hold roughly. Their initials spell LINE (Table 5).

Normal what?

The assumption is about the *residuals*, not the predictors or the outcome. Screen time is right-skewed, as Chapter 2 showed, and that is no obstacle to using it as a predictor. Running a normality test on every variable before a regression, finding one skewed and abandoning the regression is a common mistake.

Independence comes from how the data were collected. Friends filling in the survey together, or twins in the same class, would break it, and no plot of the data would show it.

Table 5: The four conditions of a linear regression. The LINE initials are a memory aid.

Condition	How it is checked
L Linear : the average outcome follows a straight line in each predictor	residual plot: no curve
I Independent : one person's answers do not drive another's	from the design, not from the data
N Normal residuals ; matters most in small samples	a histogram of the residuals
E Equal spread : residuals scatter as widely at every fitted value	residual plot: no funnel

2.10 Checking the assumptions: the residual plot

Plot each person's residual against the score the model predicts for them. If the model is sound, the result is a shapeless band around zero, about as wide on the left as on the right. Figure 8 is the plot for the four-predictor model.

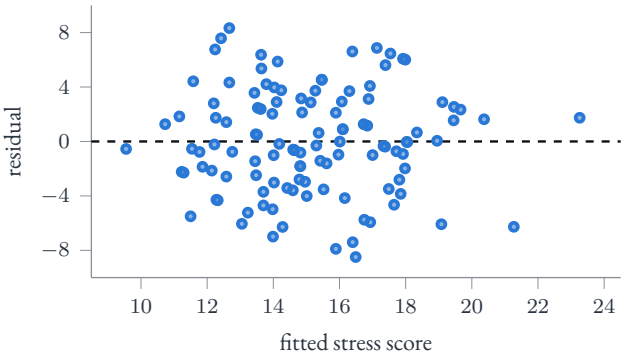


Figure 8: Residuals from the four-predictor model against the scores it predicts. A shapeless band: no curve, no funnel. **Simulated** class data, $n = 115$.

Three patterns would worry an analyst. A curve means the relationship is not a straight line, so L fails. A funnel, with residuals spreading wider as the fitted values rise, means E fails. And a few points far from the rest may be pulling the line. Figure 9 draws the first two.

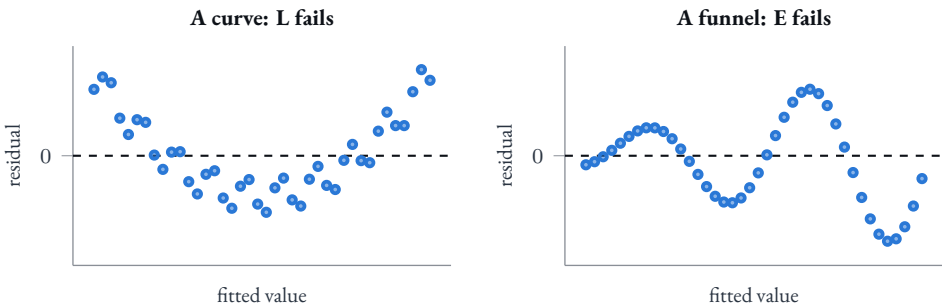


Figure 9: What failure looks like. **Constructed** for teaching, not from any data.

For the class model there is no curve, no funnel, and no point far from the rest, and a histogram of the residuals is roughly symmetric. The stress score runs only from 5 to 25, which

limits how large a residual can be; with real data, look before trusting the table.

2.11 When predictors overlap: the variance inflation factor

If the other predictors already say most of what one predictor says, the model cannot share the credit between them. The coefficients become unstable and their standard errors grow. This is called *multicollinearity*, and it is measured for each predictor by the *variance inflation factor*:

$$\text{VIF} = \frac{1}{1 - R_j^2} \quad (2.5)$$

R_j^2 the R^2 from regressing predictor j on all the other predictors

A VIF of 1 means no overlap. The standard error of that coefficient is inflated by $\sqrt{\text{VIF}}$ compared with no overlap at all. Table 6 gives the class model's values.

Table 6: Variance inflation factors for the four-predictor model. **Simulated** class data.

Predictor	VIF
Screen hours	1.17
Weeknight sleep	1.23
Gives tuition	1.05
Study hours	1.01

All four are close to 1. For sleep, $\sqrt{1.23} \approx 1.11$: its standard error is about 11% wider than it would be with no overlap. Textbooks quote cut-offs such as 5 or 10. They are conventions, and none has been checked against a source for these notes, so the safe reading is “the larger, the worse”.

Two versions of one thing

Weekend and weeknight sleep correlate at $r = 0.74$. In a model with just those two, each would have $\text{VIF} = 1/(1 - 0.74^2) \approx 2.2$, and standard errors about 1.5 times as wide. Put two measures of the same thing in one model and neither coefficient means much. Pick one.

2.12 Controlling for is not magic

Adjusting for a variable is useful, and it is regularly asked to do more than it can.

1. **It works only for what is in the model.** Anything unmeasured, or measured badly, is still free to confound. Stress from family pressure is not in the class model.
2. **It can remove what the question wanted.** If screen time raises stress *through* lost sleep, holding sleep fixed hides that route. The 0.65 is then only the part that does not go through sleep. A variable on the path between a predictor and the outcome is called a *mediator*.
3. **It is still an association.** Adjusting does not turn a survey into an experiment.

Point 2 means the question decides the model, not the data. If the question is “does screen

Multicollinearity

overlap among predictors, which makes their separate coefficients unstable and their standard errors large

Mediator

a variable on the path from a predictor to the outcome; adjusting for it removes that route



time go with stress for reasons other than sleep?”, sleep belongs in the model. If it is “how much stress goes with screen time, all routes included?”, sleep stays out. This is one reason papers often show their estimates with and without adjustment.

Before fitting

Draw the arrows you believe in, then choose what to adjust for. Choosing after seeing which predictors are significant is Chapter 5’s problem of many tests in another form.

2.13 Reporting a multiple regression

A report names the outcome, the predictors and n , gives each coefficient of interest with its interval and condition, and adds R^2 .

In a report

The stress score was regressed on screen time, weeknight sleep, private tuition and study hours ($n = 115$). With the other predictors held fixed, each extra hour of screen time was associated with 0.65 more stress points (95% CI 0.32 to 0.98). Each extra hour of sleep was associated with 1.51 fewer points (95% CI 0.62 to 2.39 fewer). Tuition and study hours showed no clear association. The model accounted for 29.6% of the variation in stress.

3 Logistic regression

3.1 When the outcome is yes or no

Many outcomes have two values: passed or failed, depressed or not on a questionnaire cut-off, recovered or not. Most tables in medical papers model an outcome of this kind. The class example is a CGPA of 3.00 or above, coded 1 for yes and 0 for no, as Chapter 2 set up. Of the 105 students with a result, 57 reached it: 54.3%.

The mean of a 0/1 variable is a proportion

Average a column of 1s and 0s and the result is the share of 1s, here 0.543. So predicting a yes or no outcome means predicting a probability.

The question is a natural one for students: does studying longer go with a better chance of a CGPA of 3.00 or above, with screen time held fixed?

3.2 Why a straight line fails

Nothing stops a least-squares line being fitted to 0s and 1s. Figure 10 does so, for study hours:

$$\hat{p} = 0.13 + 0.152 \times \text{study hours}.$$

At 6.5 hours, the busiest student in the data, the line predicts a probability of 1.12: more than a certainty, for a real student. The deeper problem is in the middle. The line says each

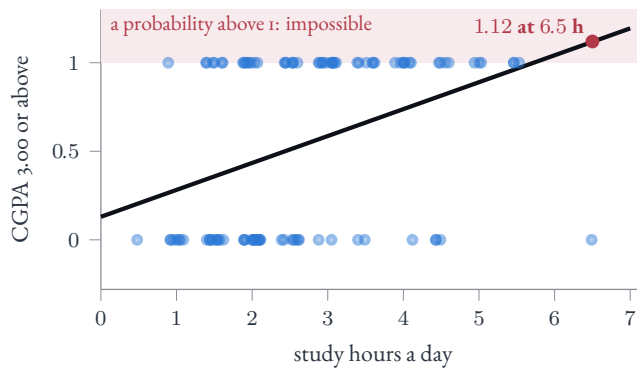


Figure 10: Each student is a dot at 0 or at 1. A straight line through them passes 1 inside the range of the data. **Simulated** class data.

extra hour adds the same 0.152 to the probability, whether a student starts at 1 hour or at 5. Near 0 and 1 that cannot be true. Going from a 50% chance to 65% is possible; going from 90% to 105% is not. The effect of an extra hour has to shrink near the top, so the model needs a curve that bends to stay between 0 and 1.

3.3 Odds and the odds ratio

The curve is built from odds, which Chapter 5 used for two-by-two tables:

$$\text{odds} = \frac{\text{number who did}}{\text{number who did not}} = \frac{p}{1-p} \quad (3.1)$$

p the probability, or the proportion, who did

Odds

the probability that something happens divided by the probability that it does not

Odds are a betting word. Odds of 3 mean “3 to 1”: three yes for every no, a probability of 3 in 4. Odds of 1 mean an even chance.

Example 3.1 • Tuition by residence, from Chapter 5

	Tutor	Do not	Odds
Mess	32	11	$32 \div 11 = 2.91$
Hall	17	30	$17 \div 30 = 0.57$

The odds ratio, mess against hall, is $2.91 \div 0.57 = 5.13$. Mess students’ odds of tutoring are about five times hall students’.

An odds ratio of 1 means the same odds in both groups: no association. Above 1, the odds are higher in the first group; below 1, lower. A logistic regression of tuition on residence alone, with hall as the reference, would give exactly this 5.13 for mess. What the regression adds is the ability to hold other variables fixed, including numeric ones.

Odds ratio

the odds in one group divided by the odds in another; 1 means no association

3.4 Probability, odds and log odds

Table 7 sets the three scales side by side.

Table 7: Five probabilities as odds and log odds. Log odds are the natural logarithm of the odds.

Probability	Odds	Log odds
0.10	0.11	-2.20
0.25	0.33	-1.10
0.50	1	0
0.75	3	1.10
0.90	9	2.20

A probability is stuck between 0 and 1. Odds remove the ceiling but are still stuck above 0. Taking the logarithm removes the floor: log odds run from minus infinity to plus infinity, and are symmetric, so 0.25 and 0.75 become -1.10 and +1.10. On that scale a straight line can go anywhere without producing nonsense. Logistic regression puts the line there:

$$\ln(\text{odds}) = a + b_1x_1 + b_2x_2 + \dots \quad (3.2)$$

$\ln$ the natural logarithm

a, b_1, b_2 the constant and the coefficients, each on the log-odds scale

To turn a value of the right-hand side, call it L , back into a probability:

$$\text{odds} = e^L, \quad p = \frac{\text{odds}}{1 + \text{odds}} = \frac{1}{1 + e^{-L}} \quad (3.3)$$

Nobody computes these by hand in practice, but a calculator with an e^x key does it in two steps.

So *logistic regression* is a linear model for the log odds. The curve in the next section is that line, turned back into probabilities.

Log odds

the natural logarithm of the odds; they run from minus to plus infinity, with 0 at an even chance

Logistic regression

a regression for a yes or no outcome, linear on the log-odds scale

3.5 The logistic curve

The class model has two predictors, study hours and screen time. Its fitted equation is

$$\ln(\text{odds of 3.00 or above}) = 0.300 + 0.843 \times \text{study hours} - 0.500 \times \text{screen hours.} \quad (3.4)$$

Figure 11 draws it for a student with the median screen time, 4.4 hours.

The curve is steepest in the middle, where an extra hour makes the most difference, and flattens at the top, where there is little room left. At 6.5 hours it predicts 0.97, where the straight line predicted 1.12. More screen time shifts the whole S to the right.

The line of section 1 was chosen by least squares. A logistic curve is chosen by *maximum likelihood*: software finds the coefficients that make the is and os actually observed as probable as possible. Islam and colleagues describe their model the same way.

Maximum likelihood

the method that chooses the coefficients making the observed outcomes most probable

Example 3.2 • The probability for one student

A student who studies 3 hours a day and has 4.4 hours of screen time.

$$1. \text{ Log odds: } 0.300 + 0.843 \times 3 - 0.500 \times 4.4 = 0.300 + 2.529 - 2.200 = 0.629.$$



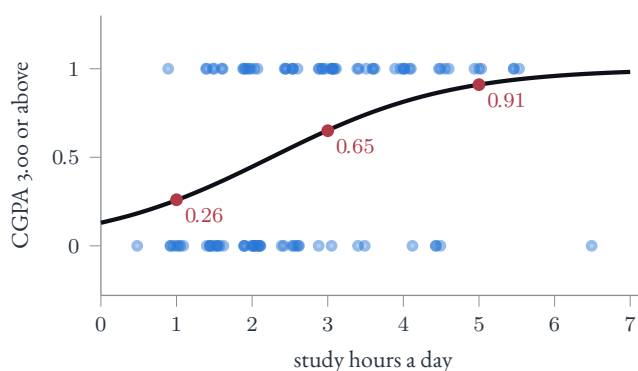


Figure 11: The fitted logistic model for a student with 4.4 hours of screen time. An S-shape that flattens near 0 and near 1 and never crosses either. **Simulated** class data.

Example 3.2 • continued

2. Odds: $e^{0.629} = 1.88$.
3. Probability: $1.88 \div (1 + 1.88) = 0.65$.

About two in three students like this one reach 3.00. Software, from the unrounded coefficients, gives odds of 1.87 and the same probability.

3.6 Reading the logistic output

Table 8 is the output. The b column is on the log-odds scale, which nobody can picture. Raise e to the power b and the result is an odds ratio, which people can.

Table 8: Logistic regression of a CGPA of 3.00 or above on study hours and screen time. $n = 103$, of whom 56 reached 3.00. **Simulated** class data.

	b (log odds)	OR = e^b	95% CI for OR	p
(Constant)	0.300			
Study hours	0.843	2.32	1.50 to 3.59	< 0.001
Screen hours	-0.500	0.61	0.47 to 0.78	< 0.001

SPSS labels the columns B and Exp(B).

The two rows in words

Study hours: each extra hour multiplies the odds of 3.00 or above by 2.32 (95% CI 1.50 to 3.59), screen time held fixed.

Screen time: each extra hour multiplies the odds by 0.61 (95% CI 0.47 to 0.78), a fall of about 39%, study hours held fixed.

For an odds ratio the null value is 1, not 0. An interval for an odds ratio that includes 1 means no clear association; both intervals here exclude 1. An odds ratio below 1 is a fall: 0.61 means the odds are 61% of what they were, a drop of 39% in the *odds*.

3.7 An odds ratio multiplies the odds, not the probability

Table 9 follows the model from 1 to 5 hours of study, with screen time at 4.4 hours.

Table 9: Odds and probability of a CGPA of 3.00 or above, by study hours, at 4.4 hours of screen time. **Simulated** class data; the gains are our arithmetic.

Study hours	Odds	Probability	Gain
1	0.35	0.26	
2	0.81	0.45	+0.19
3	1.87	0.65	+0.20
4	4.36	0.81	+0.16
5	10.12	0.91	+0.10

Each row's odds are the row above times 2.32: the same multiplier every hour. The probability does not rise by the same amount each hour; the gain shrinks as it nears 1. One odds ratio corresponds to a large change in probability in the middle of the curve and a small one near the top. This is why reporting only an odds ratio can mislead, and why good papers also give probabilities for typical people.

“Each hour adds 2.32 to the chance”

Nothing is added, and the chance does not double. The odds are multiplied by 2.32.

3.8 An odds ratio is not “times as likely”

“Times as likely” is a ratio of probabilities, the risk ratio of Chapter 5. The odds ratio is a different number, and how different depends on how common the outcome is. Table 10 compares them in two cases.

Risk ratio

the probability of the outcome in one group divided by the probability in another; “times as likely”

Table 10: Risk ratio and odds ratio for a common and a rare outcome. The first row is Chapter 5's **simulated** table of tuition by year; the second is **constructed**.

	Group 1	Group 2	Risk ratio	Odds ratio
Tutoring, seniors against juniors	61.8%	32.7%	1.89	3.33
A rare outcome	2%	1%	2.00	2.02

For a rare outcome, the odds and the probability are nearly the same number: a probability of 2% is odds of 0.0204. So their ratios agree. For a common outcome, like tutoring among seniors at 61.8%, the odds run away from the probability, and the odds ratio runs away from the risk ratio. The two are linked by

$$OR = RR \times \frac{1 - p_0}{1 - p_1} \quad (3.5)$$

p_1, p_0 the probabilities in the group of interest and in the comparison group

When both probabilities are small, the fraction is close to 1 and $OR \approx RR$.

A very common misreading

“OR = 3.33, so seniors are 3.33 times as likely to tutor.” They are 1.89 times as likely. Section 5 finds a published paper making this slip, with an outcome that about 82% of its students had.

3.9 How many predictors can a logistic model carry?

A logistic model learns from the contrast between the two outcome groups, so what limits it is the size of the *smaller* group. The usual measure is *events per variable* (EPV): the smaller group’s size divided by the number of coefficients estimated.

Peduzzi and colleagues (reference 2) tested this by simulation. They took a real study of 673 patients, 252 of whom died, with seven predictors, and drew 500 samples at each of several smaller EPVs, from 2 to 25. Their abstract reports that “for EPV values of 10 or greater, no major problems occurred”. Below 10, the coefficients were biased in both directions, the confidence intervals did not have their stated coverage, and associations in the wrong direction became more common.

Events per variable
the size of the smaller outcome group divided by the number of coefficients estimated, apart from the constant

Example 3.3 • Two models against the rule of ten

	Smaller group	Coefficients	EPV
The class model	47 below 3.00	2	23.5
Islam 2020, depression	84 not depressed	8	10.5

Islam’s model has seven variables, but age has three groups and so takes two coefficients: eight in all. The class model is comfortable. Islam’s is at the edge.

It is the smaller group that counts

476 students sounds large. With 84 in the smaller group, eight coefficients is about the limit. The rule of ten comes from one set of simulations; it is a rule of thumb, not a law, and later work has argued about it.

3.10 Reporting a logistic regression

A report names the outcome and what counts as a yes, and gives n and the number of events. Each odds ratio comes with its interval and its condition, and is stated as a change in *odds*.

In a report

A CGPA of 3.00 or above was modelled by logistic regression on daily study hours and screen time ($n = 103$; 56 reached 3.00). With screen time held fixed, each extra hour of study was associated with 2.32 times the odds (95% CI 1.50 to 3.59). For a student with the median screen time, the predicted probability rose from 0.26 at one hour of study to 0.65 at three.

4 Survey scales and Cronbach’s alpha

4.1 Five statements, one score

Every stress score in Chapters 2 to 6 came from five statements in the class survey, each answered from strongly disagree (1) to strongly agree (5).

s1	I feel overwhelmed by my academic workload.
s2	I worry about my exam results.
s3	I find it hard to relax because of my studies.
s4	I feel confident I can handle my coursework.
s5	Academic pressure affects my sleep.

Each statement is an *item*, and each is ordinal: five ordered answers. Together they make a *scale*, one score for one idea, study stress. Answer formats of this kind are called Likert items. Item s4 points the other way: agreeing with it means *less* stress. Questionnaires include reverse-worded items like this to catch people who agree with everything without reading. Before the items are added, s4 is reversed, so that 5 becomes 1, 4 becomes 2, and so on:

reversed answer = (highest + lowest) – answer = 6 – answer

(4.1)

Written for teaching

These five items were written for this course. They are not a published, validated stress scale, and the class results say nothing about stress in any wider population. The method is what transfers.

Item, scale
one question or statement in a questionnaire; several items combined into one score for one idea

Reverse scoring
turning a reverse-worded item round so that a high score means the same as on the other items; on a 1 to 5 scale, 6 – answer

4.2 How the class answered

Figure 12 shows each item’s answers as one stacked bar. It is the standard picture for Likert items. s4 is shown after reversing, so dark blue always means most stressed.

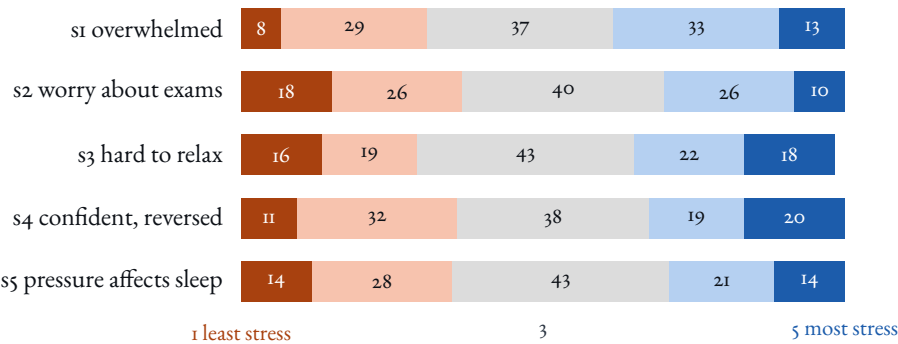


Figure 12: Counts of students giving each answer to the five stress items, s4 after reversing. **Simulated** class data.

Every item uses the whole range, and the middle answer is the most common on each. For a single item, this distribution and the median are the right summary, as Chapter 2 advised.

Nothing in the figure says whether the items agree with each other, though. Five items could have identical bars and measure five unrelated things. That needs the correlations of section 4.5.

4.3 Scoring one student

Example 4.1 • Student 2

	s1	s2	s3	s4	s5
Answer	4	3	4	4	4
Scored	4	3	4	2	4

s4 reversed: $6 - 4 = 2$. Total: $4 + 3 + 4 + 2 + 4 = 17$, on a scale from 5 to 25.

Without the reversal the total would be 19. This student agreed that they feel confident, and the unreversed score counts that as stress.

A scale also needs a rule for missing answers. Adding whatever items someone answered gives a lower total to anyone who skipped one, which is wrong. The class rule is strict: a missing answer on any item leaves the total missing. Of the 120 simulated students, 118 answered all five. A common alternative is to average the items answered when no more than one is missing. Published scales usually state their own rule in their manual, and it should be followed and reported.

4.4 Sum or average, and what the score can be treated as

The total and the average carry the same information, since the average is the total divided by 5.

	Sum	Average
Range	5 to 25	1 to 5
Student 2	17	3.40
Class mean	15.04	3.01

The average reads on the answer scale, so 3.40 is “between neutral and agree”. That is its advantage. The sum is what this course has used since Chapter 2.

One item has five ordered values, so it is ordinal and its summary is a median. The total of five items has 21 possible values and, in the class, a roughly symmetric spread. Treating it as numeric, with means, standard deviations and regression, is common practice and usually reasonable. It remains a debated point, and some researchers use rank methods even for totals. This course treats the total as numeric, as most published papers do, and says so.

Why five items beat one

Each answer carries some noise: a bad day, a misread word. Across five items the noise partly cancels, and the total is steadier than any single answer. That works only if the items measure the same thing, which is the next question.

4.5 Do the items move together? Item–total correlations

The first check is to correlate each item with the total of the *other four*. This is the *item–total correlation*. The item is left out of the total first; otherwise it is partly correlated with itself.

Item–total correlation
the correlation between one item and the total of the remaining items

Table 11: Item–total correlations, each item against the total of the other four. **Simulated** class data, $n = 118$.

Item	r with the rest
s1 overwhelmed	0.61
s2 worry about exams	0.65
s3 hard to relax	0.67
s4 confident, reversed	0.64
s5 pressure affects sleep	0.63

All five are similar and clearly positive (Table 11). A student who scores high on one item tends to score high on the rest. An item with a correlation near zero would be measuring something else, and Tavakol and Dennick (reference 3) suggest exactly this check for finding weak items. An item with a *negative* correlation is usually scored the wrong way round, as section 4.7 shows.

4.6 Cronbach's alpha

Item–total correlations look at one item at a time. *Cronbach's alpha*, named after the psychologist who set it out in 1951 (reference 4), summarises the whole scale in one number. It measures *internal consistency*: how far the items rise and fall together.

Cronbach's alpha
a summary of how consistently a set of items varies together; it runs up to 1

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum s_{\text{item}}^2}{s_{\text{total}}^2} \right) \quad (4.2)$$

k the number of items

$\sum s_{\text{item}}^2$ the sum of the five item variances

s_{total}^2 the variance of the total score

Example 4.2 • Alpha for the five stress items

Item variances (s4 reversed): 1.19, 1.37, 1.51, 1.48, 1.33. Software sums the unrounded values to 6.89; the rounded ones add to 6.88. The variance of the total is 20.91, the square of Chapter 2's standard deviation of 4.57.

$$\alpha = \frac{5}{4} \times \left(1 - \frac{6.89}{20.91} \right) = 1.25 \times 0.670 = 0.838.$$

Why it works

The variance of a total is the sum of the item variances plus a term for every pair of items that move together. If the items were unrelated, that extra term would be zero, the ratio in equation (4.2) would be 1, and α would be 0. When the items rise and fall together, the total spreads much more widely than its parts, the ratio falls, and α rises towards 1.

The factor $k/(k-1)$ corrects for the number of items and needs no further thought. Alpha of 0.838 says the five items vary together strongly: high internal consistency.

4.7 Forget to reverse s4

Chapter 1 said that forgetting the reversal would make the scale fall apart. Table 12 shows by how much.

Table 12: The scale with s4 reversed, and with s4 as typed. **Simulated** class data.

	s4 reversed	s4 as typed
s4's correlation with the rest	0.64	−0.64
Cronbach's α	0.838	0.327
Student 2's total	17	19

One wrongly scored item out of five, and α falls from 0.838 to 0.327. Unreversed, s4 pulls against the other four: confident students score higher on it and lower on the rest. Every stress score in the course would be wrong, and every result built on it, with no error message anywhere.

A negative item—total correlation is a fingerprint

Whenever one appears, the first suspect is a reverse-worded item that was not reversed. Only after that has been checked is it worth asking whether the question itself is bad.

4.8 Alpha if an item is deleted, and alpha with fewer items

Reliability output usually includes one more column: what alpha would be with each item removed.

Table 13: Cronbach's alpha with each item removed. **Simulated** class data.

Item removed	α without it
s1	0.813
s2	0.804
s3	0.798
s4	0.805
s5	0.809
None (all five)	0.838

Removing any item *lowers* alpha (Table 13), so every item earns its place. An item whose removal raised alpha would be a candidate to rewrite or drop.

Items s1 to s3 alone give $\alpha = 0.761$. Fewer items, lower alpha, even though those three agree with each other well. Alpha depends on the number of items as well as on how strongly they agree. Tavakol and Dennick make the point that a longer test raises alpha whether or not its items measure one thing. A high alpha from a 40-item questionnaire therefore says less than it seems. Exercise 30 shows how quickly alpha climbs as items are added.

Do not prune just to push alpha up

Deleting items to chase a number, then reporting the new alpha on the same data, overstates how well the scale works. Chapter 5 met the same problem as many tests: choose after looking, and the result flatters itself.

4.9 What alpha does not tell you

Chapter 1 asked whether the five statements measure one thing. Alpha cannot settle that, and it is often read as if it could. Tavakol and Dennick's short paper on alpha lists the misreadings. Four matter here.

1. **Not that the items measure one idea.** They note that a test measuring several ideas at once need not have a lower alpha than one measuring a single idea.
2. **Not that the scale measures stress.** Consistent is not the same as valid. Five items could agree with each other perfectly and measure tiredness. A scale must be reliable to be valid, but reliability does not make it valid.
3. **Not a property of the questionnaire.** Alpha belongs to the scores from one sample. The same items given to a different group can give a different alpha, so it should be computed and reported every time.
4. **Not "higher is always better".** The published advice they review puts acceptable values between 0.70 and 0.95. A maximum of 0.90 has also been recommended, because very high values suggest that items repeat each other.

Benchmarks are conventions

Treat 0.70, 0.90 and 0.95 as conventions, not tests. A short, broad scale can be useful with a modest alpha; a long one can reach 0.95 by repetition.

Showing that a set of items measures a single idea takes a different method, usually *factor analysis*. It looks at the whole pattern of correlations among the items and asks how many underlying ideas are needed to explain it. Factor analysis is beyond this course. Tavakol and Dennick's paper is a good first step, and any textbook chapter on exploratory factor analysis is the next.

Factor analysis

a family of methods for finding how many underlying ideas a set of items measures, and which items go with which

4.10 Reporting a scale

A report says where the items came from and how they were scored, including any reversal and the rule for missing answers. It gives the range of the score and alpha for this sample.

In a report

Study stress was measured with five items written for the survey, each scored 1 to 5, with one reverse-worded item reversed before summing (range 5 to 25). Students missing any item were excluded (118 of 120 retained). Cronbach's α in this sample was 0.838, and removing any item lowered it. The items have not been validated against another measure of stress.

5 Reading a published regression table

5.1 Islam 2020, Table 3

Islam and colleagues surveyed 476 students at Bangladeshi universities online during the COVID-19 pandemic (reference 1). Chapters 4 and 5 used their Tables 2 and 3. Table 3 is a pair of logistic regressions, one for depressive symptoms (mild or worse on the PHQ-9 questionnaire) and one for anxiety. Table 14 reproduces the depression model as printed.

Table 14: Islam 2020, Table 3, the depression model, as printed. Each odds ratio compares the group named with the reference group in brackets, with the other variables held fixed. Reproduced under CC BY 4.0.

Variable (compared with)	<i>B</i>	<i>p</i>	OR	95% CI
Age 21–24 (17–20)	0.057	0.082 †	1.059	0.632 to 1.775
Age over 24 (17–20)	1.060	0.059	2.886	0.961 to 8.669
Gives private tuition (no)	0.181	0.046 †	1.199	0.736 to 1.952
Female (male)	0.032	0.902	1.033	0.616 to 1.732
Urban (rural)	0.022	0.933	1.022	0.618 to 1.690
Lagging academically (no)	0.585	0.020	1.795	1.098 to 2.935
Exercises (no)	−0.128	0.601	0.880	0.544 to 1.422
Living with family (no)	0.954	0.002	2.595	1.418 to 4.751

† The printed *p* disagrees with its own interval. Chapter 4 rebuilt them from the intervals: 0.828 and 0.466.

The columns are those of Table 8: *B* on the log-odds scale, its *p*-value, the odds ratio e^B , and the interval for the odds ratio. Every odds ratio compares one group with a reference group, shown in brackets, and means nothing until the reference is known. Age has three groups, so it takes two rows, exactly as residence took two dummies in section 2.8.

Example 5.1 • Reading the lagging row

1. $B = 0.585$, on the log-odds scale.
2. $e^{0.585} = 1.795$, the odds ratio.
3. In words: students who felt they were lagging academically had about 1.8 times the *odds* of depressive symptoms of students who did not (95% CI 1.098 to 2.935). The other seven variables are held fixed.

5.2 “1.8 times more likely to be depressed”?

The paper’s Results turn the lagging row into a statement that lagging students were “1.8 times ... more likely to be depressed”. Section 3.8 showed that an odds ratio is not “times as likely”. The paper’s own Table 2 gives the counts to check how far apart the two readings are.

Table 15: Depressive symptoms by lagging academically, from the counts in Islam 2020, Table 2. The percentages and ratios are our arithmetic.

Lagging?	<i>n</i>	Depressed	%
No	118	91	77.1
Yes	358	301	84.1

Example 5.2 • What “1.8” means in probabilities

Risk ratio: $84.1 \div 77.1 = 1.09$. Crude odds ratio: 1.57. Adjusted odds ratio, from Table 3: 1.795.

Apply the adjusted odds ratio to the non-lagging group’s odds, $91 \div 27 = 3.37$:

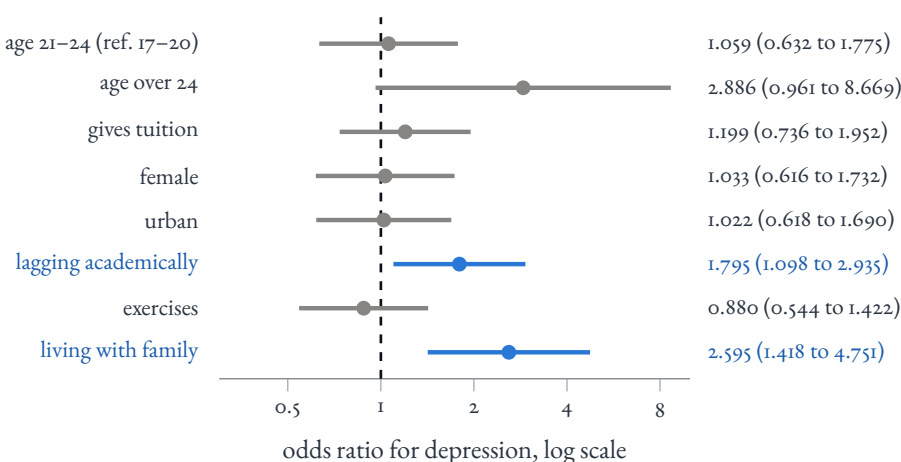
$$3.37 \times 1.795 = 6.05, \quad \frac{6.05}{1 + 6.05} = 85.8\%.$$

So “1.8 times the odds” is roughly 77.1% against 85.8%. Nothing like 1.8 times as likely.

The 85.8% is ours and only illustrates what an odds ratio of 1.8 does to a probability that is already 77%. Depression was common in both groups, at about 82% overall, so the odds ratio sits far from the risk ratio, as equation (3.5) predicts. A reader told “1.8 times more likely” would picture something far larger than the difference in the counts. Adjusting for the other seven variables moved the odds ratio only a little, from 1.57 to 1.795.

5.3 Eight intervals, and what they can and cannot say

Figure 13 draws Table 14 as a *forest plot*. For ratios the axis is logarithmic, which puts 0.5 and 2 the same distance from 1 and is fair to effects in both directions.



Forest plot

a chart of several estimates, each a dot with its interval as a line, against a vertical line at no effect

Figure 13: Islam 2020, Table 3, depression: each odds ratio and its 95% interval as printed, on a log scale. Intervals that clear 1 are blue.

Two lessons stand out. The first is width. The age-over-24 interval runs from 0.961 to 8.669, because only 42 students were over 24. The data fit no difference and a large one. Chapter 4’s lesson applies: $p = 0.059$ is not evidence of no effect, and it is not evidence of an effect either. The paper’s abstract nevertheless reports that older students had greater depression, quoting

this odds ratio and interval, and presents the tuition odds ratio of 1.199 (0.736 to 1.952) in the same way. Both intervals include 1.

The second lesson is the direction of time. Feeling one is lagging could raise the risk of depression; depression could make a student feel they are lagging. A survey taken once cannot separate the two. The authors say so in their limitations: the factors they found could be either causes or results of depression. That is to their credit, and it is the reading every row of Table 3 needs.

5.4 Check a row against the counts: living with family

In Figure 13 the living-with-family row looks like the strongest finding in the table: 2.595, $p = 0.002$. Its counts are in Table 2.

Table 16: Depressive symptoms by living with family, from the counts in Islam 2020, Table 2. The percentages and the crude odds ratio are our arithmetic.

With family?	<i>n</i>	Depressed	Not	%
No	14	12	2	85.7
Yes	462	380	82	82.3

Example 5.3 • The crude odds ratio from the counts

Odds with family: $380 \div 82 = 4.63$. Odds without: $12 \div 2 = 6$. Odds ratio, with family against without: $4.63 \div 6 = 0.77$.

Its standard error on the log scale is $\sqrt{1/380 + 1/82 + 1/12 + 1/2} \approx 0.77$, which gives a 95% interval of 0.17 to 3.52. Table 3's printed interval, 1.418 to 4.751, implies a standard error of 0.308.

Three things disagree.

1. **Direction.** The counts show slightly *less* depression among students living with family, 82.3% against 85.7%. Table 3 prints a large *increase*.
2. **Precision.** With only 2 students in one cell, the crude standard error is 0.77. Adjusting for seven other variables can move an estimate, even reverse it, but it does not normally make it far more precise than the smallest cell allows. A standard error near 0.31 is hard to believe.
3. **Anxiety.** All 14 students living apart had anxiety symptoms; none was in the minimal group. When one group contains no one without the outcome, its odds are infinite, and a logistic model cannot estimate a finite odds ratio. Software either warns or prints an enormous coefficient. Table 3 prints 1.838 (1.021 to 3.308), $p = 0.042$.

The paper's Discussion adds a fourth. It argues that living with family gives students reassurance and so reduces depression and anxiety, which is the opposite of what its Table 3 prints for depression.

It cannot be said which table is wrong

Perhaps the reference group was coded the other way round, perhaps something else happened. The counts cannot say. Until the authors explain, this row should not be cited as a finding. Writing to the corresponding author with the arithmetic is the right next step.

This check is the most transferable skill in the chapter. Before citing a row of a regression table, check it against the paper's counts, if it gives them. Check the direction, the rough size, and whether the interval is plausible for the smallest cell.

5.5 Writing a finding from a regression table

Most of the errors in this section and in Chapter 4's reading of the same table would be caught by a short checklist. A sentence reporting one row of a logistic regression should:

1. name the comparison group;
2. say *odds*, not "likely", unless a risk ratio has been computed;
3. give the interval, not just the odds ratio;
4. say "associated with", not "causes" or "leads to", unless the design earns more;
5. pass one check: does the printed p agree with the interval? An interval that includes 1 should come with $p > 0.05$, and one that excludes 1 with $p < 0.05$;
6. where the paper gives counts, agree with them in direction.

The lagging row, written properly

Students who felt they were lagging academically had higher odds of depressive symptoms than those who did not (adjusted OR 1.80, 95% CI 1.10 to 2.94). The model held age, tuition, sex, urban or rural home, exercise and living arrangement fixed. In the raw counts, 84% and 77% had symptoms. The survey cannot say whether feeling behind preceded the symptoms or followed them.

Exercise 24 applies the checklist to four more rows.

6 Fitting a regression: the procedure

1. **State the question** before seeing the data: the outcome, the predictor of interest, and whether the aim is prediction or a comparison with other things held fixed.
2. **Draw the arrows** you believe in and choose what to adjust for from them, not from which predictors turn out significant. Leave out mediators if the question is about all routes (section 2.12).
3. **Classify the outcome**: a number calls for linear regression; yes or no calls for logistic regression.
4. **Look first**: a scatterplot of the outcome against each numeric predictor. Check for codes, typos, curves and single points that dominate.

5. **Count:** the n actually in the model, after missing answers. For a logistic model, count the smaller outcome group and divide by the number of coefficients.
6. **Enter categorical predictors as dummies**, with a named reference group.
7. **Check the fit:** the residual plot for a curve, a funnel or lone points; the VIFs for overlap.
8. **Read each coefficient in one sentence:** the number, its units, its interval and its condition. For logistic regression, say odds.
9. **Report n ,** the estimates with their intervals, and R^2 or the number of events, in words that claim an association, not a cause, unless the design earns more.

For a scale: reverse the reverse-worded items and apply a stated rule for missing answers. Then look at the item–total correlations and compute alpha on this sample. Alpha is not proof of one idea, nor of validity.

7 Chapter summary

1. A regression line $\hat{y} = a + bx$ predicts an outcome from a predictor. The slope is the change in the predicted outcome per unit of the predictor, in the outcome's units; the intercept is the prediction at zero, often outside the data.
2. A residual is observed minus predicted. The least-squares line makes the sum of squared residuals as small as possible; its slope is $r \times s_y/s_x$.
3. R^2 is the share of the variation in the outcome the model accounts for. With one predictor it equals r^2 . It says nothing about shape or cause.
4. A slope has a standard error, a t statistic on $n - 2$ df, an interval and a p -value, read as in Chapter 4.
5. In a multiple regression each coefficient is the change in the outcome per unit of its predictor with the others held fixed. The screen-time slope fell from 0.88 to 0.64 with sleep in the model because, alone, it borrowed the path through sleep.
6. Adjusted R^2 charges for each predictor; the F test asks whether all the slopes could be zero together.
7. A categorical predictor with g groups enters as $g - 1$ dummy variables, each a gap from the reference group.
8. Linear regression assumes linearity, independence, normal residuals and equal spread. The residual plot checks two of them; the design decides independence. Normality concerns the residuals, not the predictors.
9. The VIF measures overlap among predictors; the standard error grows by $\sqrt{\text{VIF}}$.
10. Adjusting works only for what is measured, can remove a route the question wanted, and does not turn a survey into an experiment.
11. Logistic regression fits a straight line to the log odds of a yes or no outcome. e^b is an odds ratio: it multiplies the odds, not the probability, and is not "times as likely" unless the outcome is rare.
12. A logistic model needs about ten members of the smaller outcome group for each coefficient.
13. A scale adds items after reversing the reverse-worded ones. Cronbach's alpha measures how consistently the items vary together. It depends on the number of items and the sample, and it does not show that the items measure one idea or that they measure what they claim to.
14. Before citing a row of a regression table, check it against the paper's counts. One row of Islam 2020, Table 3, disagrees with its own Table 2 in direction and precision.

8 Key terms

Adjusted R^2

R^2 with a charge for each predictor, equation (2.3).

Coefficient plot

A chart of each coefficient as a dot with its interval as a bar.

Cronbach's alpha

A summary of how consistently a set of items varies together, equation (4.2).

Dummy variable

A 0/1 column marking one group of a categorical predictor.

Events per variable (EPV)

The smaller outcome group divided by the number of coefficients in a logistic model.

F test

A test of whether all the slopes in a model could be zero together, equation (2.4).

Factor analysis

Methods for finding how many underlying ideas a set of items measures.

Forest plot

Several estimates, each a dot with its interval, against a line at no effect.

Intercept

The predicted outcome when every predictor is zero.

Item–total correlation

The correlation of one item with the total of the others.

Least squares

Choosing the line that makes the sum of squared residuals smallest.

Log odds

The natural logarithm of the odds.

Logistic regression

A regression for a yes or no outcome, linear on the log-odds scale, equation (3.2).

Maximum likelihood

Choosing the coefficients that make the observed outcomes most probable.

Mediator

A variable on the path from a predictor to the outcome.

Multicollinearity

Overlap among predictors, measured by the VIF, equation (2.5).

Multiple regression

A regression with several predictors, each coefficient read with the others held fixed.

Odds, odds ratio

$p/(1 - p)$, equation (3.1); the odds in one group divided by the odds in another.

Outcome, predictor

The variable a model predicts, and a variable used to predict it.

R^2

The share of the variation in the outcome that the model accounts for, equation (1.4).

Reference group

The group every other group is compared with.

Residual

Observed minus predicted, equation (1.2).

Reverse scoring

Turning a reverse-worded item round, equation (4.1).

Risk ratio

The ratio of two probabilities; “times as likely”.

Scale, item

One score built from several items; one question or statement.

Slope

The change in the predicted outcome per unit of the predictor.

9 Exercises

Exercises in Part A need no calculation. Parts B to E need a calculator with an e^x key; where a critical value or a tail area is needed, it is given. Numbers not attributed to a paper or to the class data are constructed for teaching.

Part A. Ideas.

1. The correlation of stress with screen time equals the correlation of screen time with stress. Is the slope of stress on screen time the same as the slope of screen time on stress? Explain in two sentences.
2. A regression of adult weight (kg) on height (cm) has an intercept of -100 . A student writes that “a person of zero height weighs -100 kg, so the model is broken”. What is wrong with the reasoning?
3. In a survey, each extra hour of daily phone use goes with 0.9 points lower on a wellbeing score (95% CI 0.5 to 1.3 lower). A news article reports that “cutting phone use by an hour will raise wellbeing by 0.9 points”. Give two reasons the survey cannot support that sentence.
4. A study of 3,000 people reports a slope with $p < 0.001$ and $R^2 = 0.01$. Is the relationship strong, weak, or both significant and weak? What does each number tell the reader?
5. A supervisor tells a student that screen time is skewed, so it cannot be used as a predictor in a linear regression. Is the supervisor right? What should be checked instead?
6. A student puts weekend sleep and weeknight sleep in the same regression of stress. Both coefficients have wide intervals that include 0. Does that show sleep is unrelated to stress? What would help?
7. In a group where 60% pass, a factor has an odds ratio of 2 for passing. Explain why “twice as likely to pass” cannot be right, without calculating anything precisely.
8. A 40-item questionnaire reports Cronbach’s $\alpha = 0.95$. Give three things this does not show.

Part B. Calculation.

9. In a class, daily study hours have mean 3 and SD 1.5, ranging from 0.5 to 7; exam marks have mean 62 and SD 12; and $r = 0.5$.

- (a) Find the slope and intercept of the line predicting the mark from study hours.
- (b) Predict the mark for 5 hours of study, and for 12 hours. Comment on the second.
- (c) What is R^2 ?

10. Using the line from Exercise 9, find the residuals of a student who studies 2 hours and scores 48, and of one who studies 6 hours and scores 80. Which one did better than the line predicted?

11. A slope is $b = 1.2$ with SE 0.5, from $n = 40$. On 38 df, $t^* = 2.024$, and the area beyond $t = 2.4$ in each tail is 0.0107. Find t , the 95% interval and the two-sided p -value.

12. Fifty students. Model A has one predictor and $R^2 = 0.25$. Model B has five predictors and $R^2 = 0.30$.

- (a) Compute adjusted R^2 for both. Which is better by this measure?
- (b) In Model B, one predictor regressed on the other four gives $R^2 = 0.75$. Find its VIF, and say by how much its standard error is inflated.

13. A model for exam marks is $\widehat{\text{mark}} = 40 + 5 \times \text{study hours} - 2 \times \text{screen hours} + 3 \times \text{tutorial}$, where tutorial is 1 for students who attend tutorials and 0 otherwise.

- (a) Predict the mark of a student who studies 4 hours, has 5 hours of screen time and attends.
- (b) Predict it for the same student if they did not attend.
- (c) By how much does the prediction change with two more hours of screen time, the rest fixed?
- (d) Residence (hall, mess, home) is to be added. How many columns, and what do their coefficients mean?

14.

- (a) Convert probabilities 0.20 and 0.80 to odds, and odds of 1.5 to a probability.
- (b) Group 1 has the outcome 80% of the time, group 2 20%. Find the risk ratio and the odds ratio.
- (c) Repeat for 4% against 2%. What do (b) and (c) show together?

15. A logistic coefficient is $b = 0.405$.

- (a) Find the odds ratio.
- (b) In people with a probability of 0.20 at baseline, what is the probability one unit higher? Find the risk ratio.
- (c) What odds ratio does $b = -0.693$ give, and what does it mean?

16.

- (a) A five-item scale is scored 1 to 5, and item 4 is reverse-worded. A student answers 2, 1, 2, 5, 1. Find the total correctly, and the total if item 4 is not reversed.
- (b) A four-item scale has item variances adding to 4.2, and the total has variance 10.5. Find Cronbach's α .
- (c) What would α be if the variance of the total were 4.2? What would that say about the items?

17. A hospital study of 300 patients, 45 of whom died, fits a logistic model with 12 coefficients. Find the events per variable. Roughly how many coefficients would the rule of ten allow?

Part C. The class data. All values are **simulated**.

18. Using $\widehat{\text{stress}} = 10.94 + 0.88 \times \text{screen hours}$, find the predicted stress and the residual of a student with 6 hours of screen time and a score of 20. Do the same for a student with 3 hours and a score of 20. Which is further above the line?

19. Screen time has SD 2.32 and stress SD 4.53, with $r = 0.45$. Find the slope of *screen time on stress*. Multiply it by the slope of stress on screen time, 0.88, and compare the product with r^2 .

20. From Table 3, write the sentence for a *two-hour* difference in weeknight sleep, with its interval.

21. Use Table 3 to predict the stress score of a student with 3 hours of screen time, 7 hours of sleep, no tuition and 3 hours of study.

22. Use equation (3.4) to find the probability of a CGPA of 3.00 or above for a student who studies 3 hours and has 6 hours of screen time. Compare it with Example 3.2.

23. Items s_1 , s_2 and s_3 have variances 1.19, 1.37 and 1.51, and their total has variance 8.26. Compute Cronbach's α for these three items.

Part D. Reading research.

24. **The activity.** Write one sentence for each row below from Islam 2020, Table 3, following the checklist of section 5.5. For the fifth point, rebuild the p -value from the interval as in Chapter 4 and compare it with the printed one.

	Row	OR (95% CI)	Printed p
1	Exercises, depression	0.880 (0.544 to 1.422)	0.601
2	Age over 24, depression	2.886 (0.961 to 8.669)	0.059
3	Lagging, anxiety	1.780 (1.099 to 2.883)	0.019
4	Female, anxiety	1.462 (0.866 to 2.468)	0.155

25. Each statement below misreads a regression. Say what is wrong with each.

- (a) "A student with no screen time would score 10.94 on stress."

- (b) “Study hours have no effect on stress ($p = 0.534$).”
- (c) “Sleep matters more than tuition, because its dot is further from zero.”
- (d) “The odds ratio of 0.61 means each hour of screen time lowers the chance of a 3.00 by 39%.”
- (e) “Residence was entered as 1 = hall, 2 = mess, 3 = home, and its slope was 0.4.”
- (f) “Cronbach’s α of 0.838 shows the scale measures study stress.”
- (g) “We began with twelve predictors and kept the three with $p < 0.05$. The final model shows these three predict stress, each with $p < 0.05$.”
- (h) “Our logistic model has 20 predictors, but 400 students is a large sample, so it is safe.”
Of the 400, 84 had the outcome.

Part E. Going further.

26. A simple slope of stress on a predictor X is 2.0. With a second predictor Z added, the slope of X is 1.2 and the coefficient of Z is -2.0 . Use equation (2.2) to find the slope of Z on X , and describe in words how X and Z are related.

27. A researcher asks whether a mentoring scheme raises CGPA. Mentored students study more, and study hours raise CGPA. Should study hours be in the model? Give the answer for two different questions the researcher might mean.

28. In the class data, study hours and screen time are almost uncorrelated ($r = 0.005$). With study hours alone, the odds ratio for a CGPA of 3.00 or above is 2.07; with screen time added, it is 2.32. In a linear regression, adding an uncorrelated predictor would leave a slope unchanged. Why can a logistic odds ratio change anyway?

29. Use equation (3.5) to find the odds ratio for (a) 20% against 10%, and (b) 60% against 30%. Both have a risk ratio of 2. (c) Check the crude odds ratio for lagging in Islam 2020 from the risk ratio 1.09 and the percentages 84.1% and 77.1%.

30. When all items have the same variance, alpha depends only on the number of items k and their average inter-item correlation $\bar{r}$:

$$\alpha = \frac{k \bar{r}}{1 + (k - 1) \bar{r}}.$$

The five class items have $\bar{r} \approx 0.51$. Find α for 3, 5, 10 and 40 items with the same $\bar{r}$. What does this say about Exercise 8?

31. In Islam 2020, all 14 students living apart from their families had anxiety symptoms. Explain why a logistic regression cannot estimate a finite odds ratio for this variable, and draft two sentences to the authors asking about the row.

10 Solutions

Part A.

1. No. Stress on screen time has slope 0.88 stress points per hour; screen time on stress has slope 0.23 hours per stress point. Least squares minimises the misses in whichever variable is the outcome, so the two lines differ unless r is ± 1 ; their slopes multiply to r^2 . (Exercise 1.)
2. The intercept is the prediction at zero height, far outside the heights in the data. A straight line fitted to adults says nothing about heights nobody has, so the intercept is not worth interpreting. The slope still describes how weight changes with height within the range of the data. (Exercise 2.)
3. First, the slope compares different people; it does not describe what happens when one person changes. Second, the association has other explanations: a common cause (loneliness might raise phone use and lower wellbeing) and reverse causation (low wellbeing might drive phone use). An experiment would be needed for “cutting ... will raise”. (Exercise 3.)
4. Both significant and weak. The p -value says the slope is unlikely to be exactly zero; with 3,000 people even a tiny slope is detected. $R^2 = 0.01$ says the predictor accounts for 1% of the variation in the outcome. (Exercise 4.)
5. No. The normality assumption concerns the residuals, not the predictors. Check the residual plot for a curve or a funnel, and a histogram of the residuals. It is worth looking at the few students in the long tail of screen time, in case one of them pulls the line. (Exercise 5.)
6. No. The two sleep measures overlap heavily (in the class, $r = 0.74$), so the model cannot share the credit between them and both standard errors inflate. Together they may still predict stress well. Use one of them, or combine them into a single measure of sleep. (Exercise 6.)
7. Twice 60% is 120%, which is impossible. An odds ratio of 2 doubles the odds, from 1.5 to 3, which is a probability of 75%: a risk ratio of 1.25. (Exercise 7.)
8. Any three of these:
 - that the items measure a single idea;
 - that the questionnaire measures what it claims to (validity);
 - that the same alpha will be found in another sample;
 - that the items are good: 40 items give a high alpha even when they correlate only modestly (Exercise 30);
 - that the items are not redundant: values above about 0.90 can mean repetition.

(Exercise 8.)

Part B.

9. (a) $b = 0.5 \times 12/1.5 = 4$ marks per hour; $a = 62 - 4 \times 3 = 50$. (b) At 5 hours: $50 + 4 \times 5 = 70$. At 12 hours: $50 + 48 = 98$. Twelve hours is far outside the data, which stop at 7, so the second prediction is extrapolation and should not be trusted. (c) $R^2 = 0.5^2 = 0.25$. (Exercise 9.)

10. 2 hours: predicted 58, residual $48 - 58 = -10$. 6 hours: predicted 74, residual $80 - 74 = +6$. The second student did better than the line predicted. (Exercise 10.)

11. $t = 1.2/0.5 = 2.4$. Interval: $1.2 \pm 2.024 \times 0.5 = 1.2 \pm 1.01$, from 0.19 to 2.21. $p = 2 \times 0.0107 \approx 0.021$. The interval leaves out 0, as $p < 0.05$ requires. (Exercise 11.)

12. (a) Model A: $1 - 0.75 \times 49/48 = 0.234$. Model B: $1 - 0.70 \times 49/44 = 0.220$. Model A is better by adjusted R^2 : the four extra predictors did not earn their charge. (b) $VIF = 1/(1 - 0.75) = 4$; the standard error is multiplied by $\sqrt{4} = 2$. (Exercise 12.)

13. (a) $40 + 20 - 10 + 3 = 53$. (b) 50. (c) $2 \times (-2) = -4$ marks. (d) Two dummy columns, for example mess and home, with hall as the reference group. Each coefficient is the gap in predicted mark from hall students with the same study hours, screen time and tutorial attendance. (Exercise 13.)

14. (a) $0.20/0.80 = 0.25$; $0.80/0.20 = 4$; $1.5/2.5 = 0.60$. (b) Risk ratio $80/20 = 4$; odds ratio $4/0.25 = 16$. (c) Risk ratio 2; odds ratio $(0.04/0.96)/(0.02/0.98) = 2.04$. For a common outcome the odds ratio is far from the risk ratio; for a rare one they almost agree. (Exercise 14.)

15. (a) $e^{0.405} = 1.50$. (b) Odds $0.20/0.80 = 0.25$, times 1.50 is 0.375, a probability of $0.375/1.375 = 0.273$. Risk ratio $0.273/0.20 = 1.36$, not 1.50. (c) $e^{-0.693} = 0.50$: each unit halves the odds. (Exercise 15.)

16. (a) Item 4 reversed: $6 - 5 = 1$. Total $2 + 1 + 2 + 1 + 1 = 7$; unreversed, $2 + 1 + 2 + 5 + 1 = 11$. (b) $\alpha = \frac{4}{3} \times (1 - 4.2/10.5) = \frac{4}{3} \times 0.6 = 0.80$. (c) $\frac{4}{3} \times (1 - 1) = 0$. The total would spread no more than its parts: the items would not move together at all. (Exercise 16.)

17. The smaller group is the 45 deaths (255 survived). $EPV = 45/12 = 3.75$, well below 10. The rule of ten allows about $45/10 = 4.5$, so four coefficients. (Exercise 17.)

Part C.

18. 6 hours: $10.94 + 0.88 \times 6 = 16.22$; residual $20 - 16.22 = 3.78$. 3 hours: $10.94 + 0.88 \times 3 = 13.58$; residual 6.42. The 3-hour student is further above the line. (Exercise 18.)

19. $b = 0.45 \times 2.32/4.53 \approx 0.23$ hours of screen time per stress point. The product $0.88 \times 0.23 \approx 0.202$, and $r^2 = 0.45^2 \approx 0.20$: the same, as section 1.9 said. (Exercise 19.)

20. Among students with the same screen time, tuition and study hours, two extra hours of weeknight sleep go with -3.02 stress points (95% CI -4.78 to -1.24). The coefficient and both ends of its interval are doubled. (Exercise 20.)



21. $22.44 + 1.95 + (-10.57) + 0.00 + (-0.54) = 13.28$, from 0.65×3 , -1.51×7 , 0.53×0 and -0.18×3 . (Exercise 21.)

22. Log odds $0.300 + 2.529 - 3.000 = -0.171$; odds $e^{-0.171} = 0.84$; probability $0.84/(1 + 0.84) = 0.46$. With 4.4 hours of screen time it was 0.65; 1.6 more hours of screen time, with study fixed, bring it below an even chance. (Exercise 22.)

23. Sum of variances 4.07. $\alpha = \frac{3}{2} \times (1 - 4.07/8.26) = 1.5 \times (1 - 0.493) \approx 0.760$. Software, from the unrounded variances, gives 0.761. Fewer items, lower alpha than the five-item 0.838. (Exercise 23.)

Part D.

24. Model sentences; others are possible.

1. Students who exercised had slightly lower odds of depressive symptoms than those who did not (OR 0.88, 95% CI 0.54 to 1.42), other factors held fixed, but the interval includes 1: no clear association. Rebuilt $p = 0.602$, against 0.601 printed: they agree.
2. Students over 24 had higher odds than those aged 17 to 20 (OR 2.89, 95% CI 0.96 to 8.67), but with only 42 students over 24 the interval is very wide and includes 1. The data fit no difference and a large one; this is not evidence of no association. Rebuilt $p = 0.059$: agrees.
3. Students who felt they were lagging had higher odds of anxiety symptoms (OR 1.78, 95% CI 1.10 to 2.88), other factors held fixed. Anxiety was common, so this is not “1.8 times as likely”. Rebuilt $p = 0.019$: agrees.
4. Female students had higher odds of anxiety symptoms than male students (OR 1.46, 95% CI 0.87 to 2.47), but the interval includes 1: no clear association. Rebuilt $p = 0.155$: agrees.

All four sentences say “associated” or “odds”, never “causes”. (Exercise 24.)

25. (a) Extrapolation: nobody in the class has zero screen time; the smallest answer is 1.4 hours. (b) Absence of evidence: the interval, -0.77 to 0.40 , allows a small relationship either way. “Effect” also claims a cause. (c) The two are in different units, per hour of sleep and tutors against non-tutors, so the distances cannot be compared. (d) The odds fall by 39% per hour. The chance falls by different amounts at different points on the curve. (e) The codes are names; residence needs two dummy variables. (f) Alpha measures consistency, not validity. (g) Choosing predictors by their p -values and then reporting those p -values on the same data makes them look better than they are; with twelve candidates, some pass by chance. (h) Count the smaller group: $84/20 = 4.2$ events per variable, far below 10. (Exercise 25.)

Part E.

26. $2.0 = 1.2 + (-2.0) \times \delta$, so $\delta = (2.0 - 1.2)/(-2.0) = -0.4$. Each unit of X goes with 0.4 units less Z , and less Z goes with more stress, so X alone borrowed 0.8 through Z . It is the sleep pattern of section 2.3. (Exercise 26.)

- 27.** Study hours are a mediator: on the path from mentoring to CGPA. If the question is “does mentoring raise CGPA at all?”, leave study hours out, or the model hides the main route. If it is “does mentoring raise CGPA by routes other than extra study?”, put them in. The data cannot choose; the question does. (Exercise 27.)
- 28.** Odds ratios are *non-collapsible*. With study hours alone, the model compares students of all screen times mixed together; screen time strongly affects the outcome, and mixing groups with different odds pulls the odds ratio towards 1. With screen time in the model, study hours are compared among students with the same screen time, and the odds ratio is further from 1. No confounding is needed: screen time is unrelated to study hours. So a change in an odds ratio on adjustment is not, on its own, evidence of confounding. (Exercise 28.)
- 29.** (a) $2 \times 0.90/0.80 = 2.25$. (b) $2 \times 0.70/0.40 = 3.5$. (c) $1.09 \times (1 - 0.771)/(1 - 0.841) = 1.09 \times 0.229/0.159 \approx 1.57$, the crude odds ratio from the counts. The same risk ratio gives a larger odds ratio as the outcome becomes more common. (Exercise 29.)
- 30.** $k = 3$: 0.757; $k = 5$: 0.839; $k = 10$: 0.912; $k = 40$: 0.977. The five-item value is close to the class’s actual 0.838. With 40 items, even an average correlation of 0.3 gives $12/12.7 \approx 0.94$. So an alpha of 0.95 from 40 items says more about the length of the questionnaire than about the quality of its items. (Exercise 30.)
- 31.** The odds of anxiety among students living apart are $14/0$, which is infinite. Any odds ratio comparing them with another group is infinite or zero, so the model’s estimate grows without limit as the fitting continues. Software warns, or prints a huge coefficient and standard error. A draft: “Your Table 2 shows that all 14 students living apart from their families had anxiety symptoms, so a finite odds ratio for this variable cannot be estimated. Could you explain how the value of 1.838 (1.021 to 3.308) in Table 3 was obtained, and confirm which group was the reference?” (Exercise 31.)

II Our arithmetic and constructed numbers

Every number below was computed for these notes. None is printed by the papers cited, although the inputs from real papers are.

Value	How	Inputs from
Risk ratio 1.09, crude OR 1.57, odds 3.37 and 6.05, and 85.8% (Example 5.2)	from the counts; equations (3.1) and (3.3)	Islam 2020, Tables 2 and 3
Crude OR 0.77, SE 0.77, interval 0.17 to 3.52; implied SE 0.308 (Example 5.3)	from the counts; the SE from the width of the printed interval	Islam 2020, Tables 2 and 3
EPV 10.5	$84 \div 8$	Islam 2020, Tables 1 and 3
Rebuilt p -values in Table 14 and Exercise 24	from the odds ratio and interval, as in Chapter 4	Islam 2020, Table 3
Gains in Table 9; the terms of Example 2.2; the 0.24 of Example 2.1; $F \approx 11.56$; adjusted $R^2 \approx 0.271$; $\sqrt{\text{VIF}} \approx 1.11$; VIF ≈ 2.2 for the two sleep measures	by hand from the printed class values	simulated class data
Table 7; the rare row of Table 10	from the definitions	none

Simulated, not real students: every figure from the class data, including every model, the residual plots, the student scored in Example 4.1, and the item statistics of section 4. All are generated by data/class-data.R from a fixed seed, and the class figures will be replaced by the real survey. The five stress items were written for the course and are not a validated scale.

Constructed for teaching, and not from any study: the two panels of Figure 9 and the rare outcome of Table 10. So is every scenario and number in Exercises 1 to 8, 9 to 17, 26 and 27, and in parts (a) and (b) of Exercise 29. Every answer to Part B was checked in R.

12 References and further reading

- Islam MA, Barna SD, Raihan H, Khan MNA, Hossain MT. Depression and anxiety among university students during the COVID-19 pandemic in Bangladesh: a web-based cross-sectional survey. *PLoS One* 2020;15(8):e0238162. PMID 32845928. PMC7449469. doi:10.1371/journal.pone.0238162. Licensed CC BY 4.0; free to read and share. Tables 1 to 3, the Results, the abstract, the Discussion and the limitations were read in the full text.
- Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996;49(12):1373–1379. PMID 8970487. doi:10.1016/s0895-4356(96)00236-3. The design and the quotation are from the abstract.
- Tavakol M, Dennick R. Making sense of Cronbach's alpha. *Int J Med Educ* 2011;2:53–55. PMID 28029643.

PMC4205511. doi:10.5116/ijme.4dfb.8dfd. Free to read; a short, clear starting point on alpha and on factor analysis.

4. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;16(3):297–334. doi:10.1007/BF02310555. Not indexed in PubMed; details checked against the Crossref record. Cited for the name and year only; not read for these notes.

5. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Used to generate the class data and fit every model.

6. Class data: a **simulated** stand-in for the Stat.BD class survey, generated by data/class-data.R from the fixed seed 20261005. Not real students. No DOI or PMID.

13 For students in the Stat.BD course

Before the next session. Read this chapter, and work the Part A exercises (Exercises 1 to 8) and the activity of Exercise 24. Exercise 25, spotting the wrong reading, is good preparation for reading papers. The rest are for practice at any point in the course.

What the next session does with this chapter. Session 7 moves to SPSS, from raw data to results. It opens with a check that SPSS is installed and working. It then imports the class survey, labels the variables and their codes, cleans the data, and computes the stress score with *s4* reversed, as in section 4.3. It ends with the descriptive tables and charts of Chapter 2. When SPSS computes the stress score, the reversal is the one step that silently breaks everything after it; section 4.7 showed what alpha looks like when it is missed. Session 8 then runs the tests of Chapters 4 and 5, the regressions of this chapter, and Cronbach’s alpha. It shows where each number in this chapter appears in SPSS output.

Questions this chapter left open, and where each is settled.

Question	Where it is settled
How is the stress score computed in software?	Session 7
Where do these tables come from in SPSS output?	Session 8
Do the five items measure one idea?	Factor analysis, beyond this course; start with reference 3
What if the slope of screen time differs with sleep?	Interactions, beyond this course
What do the real class answers show?	The real survey replaces the simulated numbers