

Regression and Survey Scales

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

Muhtasim Munif Fahim, MSc

Stat.BD

Data Science Research Lab, Department of Statistics & Data Science,
University of Rajshahi

Session 6 of 14 • Theory

Session 5 related two variables. Today, one outcome and several at once

Session 5 gave us

- ▶ chi-square for two categorical variables
- ▶ r for two numeric variables
- ▶ four explanations for any association

Each looked at two variables and nothing else.

Real questions have an outcome and several things that might explain it.

Today, in five acts

- | | |
|-----|---|
| I | A line through the cloud |
| II | Holding other things fixed |
| III | Yes or no outcomes: logistic regression |
| IV | Survey scales and Cronbach's alpha |
| V | Reading a published results table |

The question behind today

How much does the outcome change when one thing changes and the others stay put?

Class data throughout: **simulated** stand-in for the class survey (data/class-data.R, fixed seed). Not real students.



What you will be able to do by the end of today

The test of this session

Given a regression table from a paper, say in plain words what each number means, what was held fixed, and what the table cannot show.

- ▶ Interpret regression coefficients in plain language
- ▶ Explain what “controlling for” a variable means
- ▶ Say when logistic regression is needed instead of linear regression
- ▶ Build a scale score from Likert items and interpret Cronbach’s alpha
- ▶ Read and explain a results table from a published paper



The question Session 5 left: where did the rest go?

	Stress points per extra screen hour
Screen time alone	0.88
Screen time, sleep held fixed	0.64

Same 115 students in both rows.

Session 5 could hold residence fixed by splitting the class into three groups. Sleep is a number, with dozens of values. Splitting by it leaves a handful of students in each slice.

Regression holds a number fixed without splitting anyone off.

Where today ends up

By the middle of Act II, the drop from 0.88 to 0.64 will have a picture and a reason.

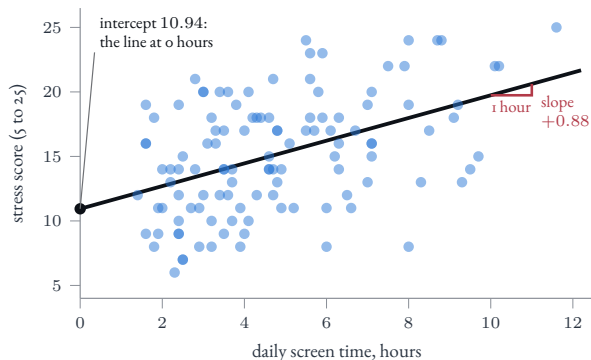
Simulated class data.



Act I

A line through the cloud: simple linear regression

A line through the scatter: stress on screen time



Session 5 measured how closely these points follow a line: $r = 0.45$. Regression draws the line.

$$\widehat{\text{stress}} = 10.94 + 0.88 \times \text{screen hours}$$

The hat means “predicted by the line”.

Regression has a direction

Stress is the **outcome**, on the up axis. Screen time is the **predictor**, across. Correlation had no such roles.

Simulated class data, $n = 115$ students with both answers. Line fitted by least squares (software).



Intercept and slope, in words

Slope, $b = 0.88$

For each extra hour of daily screen time, the predicted stress score is 0.88 points higher.

Units: stress points per hour. Change the units and the number changes.

Intercept, $a = 10.94$

The predicted stress score at zero hours of screen time.

Nobody in the class reports zero; the smallest is 1.4 hours.
The intercept anchors the line; it describes no real student.

“Screen time raises stress by 0.88”?

The slope compares students who differ by an hour. It does not say what would happen if one student added an hour. That needs an experiment.

Simulated class data.



Using the line to predict

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.

A prediction is an average for students like that, not a forecast for one person. Two students with 4 hours each can sit far apart.

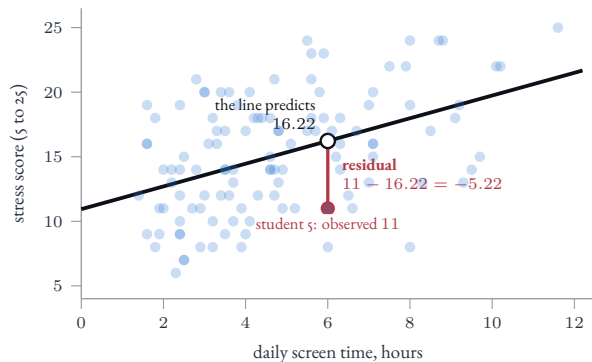
Simulated class data. The 3.52 and the 28.5 are our arithmetic from the rounded coefficients.

Stay inside the data

The class runs from 1.4 to 11.6 hours. At 20 hours the line predicts 28.5, on a scale that stops at 25.



Residuals: how far each student sits from the line



$$\text{residual} = \text{observed} - \text{predicted}$$

Student 5 reports 6.0 hours of screen time.

The line predicts
 $10.94 + 0.88 \times 6 = 16.22$.

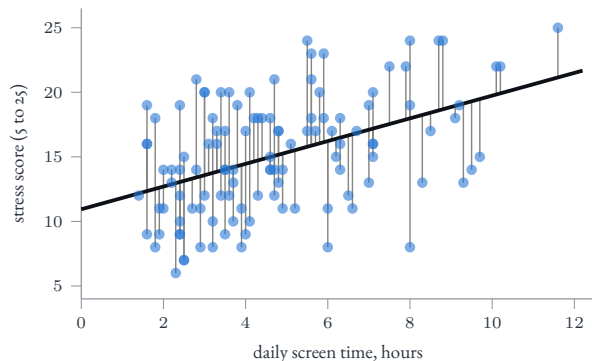
Their score is 11, so the residual is
-5.22.

Negative: less stressed than the line expects.
Positive: more.

Simulated class data. Student 5 is the fifth row of the dataset with both answers.



Why this line? Least squares



Every grey stick is one residual. Tilt or shift the line and some sticks shrink while others grow.

The regression line is the one that makes

$$\sum (\text{residual})^2$$

as small as possible. Hence **least squares**.

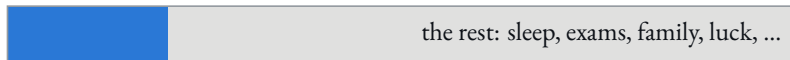
Why square?

For the same reason as the SD in Session 2: squares stop positive and negative gaps from cancelling, and punish big misses more than small ones.

Simulated class data.



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



$R^2 = 0.204$: 20.4% lined up with screen time

With one predictor, $R^2 = r^2$:
 $0.45^2 \approx 0.20$.

Session 5's r and today's R^2 are the same information on two scales.

Typical miss: the residuals have an SD of 4.06 points, against 4.57 for stress around its mean.

Knowing screen time narrows the guess, but not by much.

Simulated class data. The stress SD is over the 118 students with a score; the model uses 115.



Is the slope more than luck? Standard error, t , interval, p

The same machinery as Sessions 3 and 4, applied to b :

Slope	$b = 0.88$
Standard error	0.164
$t = b \div \text{SE}$	$0.88 \div 0.164 \approx 5.38$
95% interval	0.56 to 1.20
p	< 0.001

H_0 : in the population the slope is 0. Screen time predicts nothing.

Interval: $b \pm 1.98 \times \text{SE}$, with 1.98 the t value for $115 - 2 = 113$ degrees of freedom.

The interval leaves out 0, and p is tiny.

Say both

Between 0.56 and 1.20 stress points per hour, with $p < 0.001$. The interval says how big; p only says “not zero”.

Simulated class data. SE, t , p and the interval by software; t from the unrounded values.



Reading the output: the coefficient table

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

$n = 115$. $R^2 = 0.204$. Residual SD 4.06.

Read it in this order

1. Which row? The predictor, not the constant
2. b , with its units, in a sentence
3. The interval: how big, plausibly
4. p last

In a report

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

Simulated class data. Layout as most software prints it; SPSS calls the constant “(Constant)” and b “B”.



Three ways a regression line misleads

Mistake	What goes wrong	Instead
Reading the slope as strength	The slope depends on units: per minute of screen time it is $0.88 \div 60 = 0.015$. Same data, same r .	Strength is r or R^2 ; size is b
Swapping the roles	Stress on screen time and screen time on stress are two different lines, each minimising misses in its own outcome.	Decide the outcome first
Reading a curve as a line	A straight line fitted to a curve gives a slope for a shape that is not there.	Plot first; check the residuals

And the one from Session 5

A slope from a survey is still an association. Everything from Session 5's Act III applies.

Simulated class data. The per-minute slope is our arithmetic.



Act II

Holding other things fixed: multiple regression

Add sleep: the screen-time slope drops from 0.88 to 0.64

Model 1: screen time alone

$$\widehat{\text{stress}} = 10.94 + 0.88 \text{ screen}$$

Model 2: screen time and weeknight sleep

$$\widehat{\text{stress}} = 22.87 + 0.64 \text{ screen} - 1.60 \text{ sleep}$$

Each slope now has a condition attached:

0.64: per extra screen hour, **among students who sleep the same.**

−1.60: per extra hour of sleep, **among students with the same screen time.**

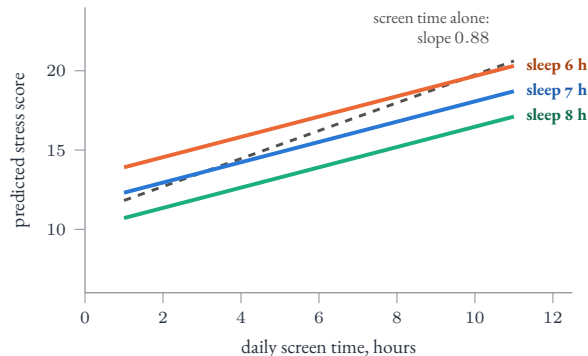
This is “controlling for”

A coefficient in a multiple regression answers: what if this variable differed and the others in the model did not?

Simulated class data, the same 115 students in both models.



What “controlling for sleep” means, drawn



Model 2 is one line for each amount of sleep. The lines are parallel: the slope is 0.64 whatever the sleep.

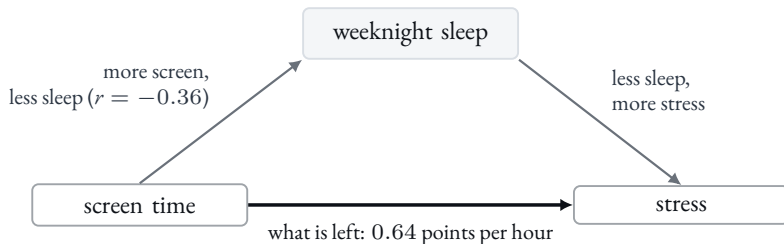
Moving along one coloured line changes screen time with sleep fixed. Moving between lines changes sleep with screen time fixed: -1.60 a line.

The dashed grey line lets sleep vary too, so it is steeper.

Simulated class data. Lines drawn from the Model 2 coefficients at 6, 7 and 8 hours of sleep.



Why the slope shrank



Screen time goes with less sleep, and less sleep goes with more stress. Alone, screen time gets the credit for both paths.

With sleep in the model, the path through sleep belongs to sleep. The $0.88 - 0.64 = 0.24$ that went is the part that travelled with it.

If the arrows point the other way

If stress keeps people up and on their phones, the same numbers arise. Regression cannot tell which way the arrows run.

Simulated class data. The 0.24 is our arithmetic.



Four predictors at once

Predictor	<i>b</i>	SE	<i>t</i>	<i>p</i>	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

Outcome: stress score. $n = 115$. $R^2 = 0.296$, adjusted $R^2 = 0.271$.

Same table, more rows

Every row is read exactly like Act 1's, with one addition: "holding the other three fixed".

Simulated class data. Fitted by least squares (software).



Each coefficient in one sentence

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 points, that is 1.51 fewer (95% CI -2.39 to -0.62), 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.

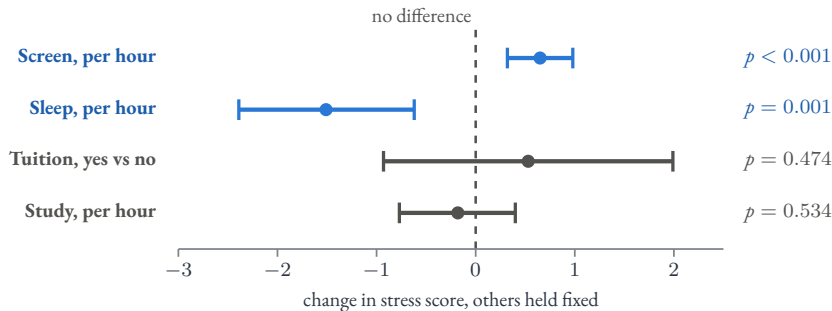
“Tuition has no effect on stress”

The interval allows tutors to be 0.93 points lower or 1.99 higher. “We could not tell” is the honest reading, as in Session 4.

Simulated class data.



The coefficients with their intervals



Blue intervals leave out 0; grey ones include it. Each row is in its own units, so compare each with 0, not with each other.

Simulated class data. 95% intervals.



One student's predicted score

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

Constant	22.44
0.65×5	+3.25
-1.51×6	-9.06
0.53×1	+0.53
-0.18×2	-0.36
Predicted stress	16.80

Multiply each value by its coefficient and add.
That is all a linear model does.

Change only the sleep to 8 hours:
 $2 \times 1.51 = 3.02$ points lower, whatever the other values are.

A 0/1 variable just switches a term on

For a student who does not tutor, the 0.53 line becomes $0.53 \times 0 = 0$.

Simulated class data. Our arithmetic from the rounded coefficients; the unrounded ones also give 16.80.



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

	Screen alone	Four predictors
R^2	0.204	0.296
Adjusted R^2		0.271

R^2 **never falls** when a predictor is added, even a useless one. Adjusted R^2 charges for each predictor:

$$1 - (1 - R^2) \frac{n - 1}{n - k - 1}$$

k = number of predictors.

Simulated class data. Software. With $n = 115$, $k = 4$: $1 - 0.704 \times 114/110 = 1 - 0.704 \times 1.036 \approx 0.271$, our arithmetic from the rounded R^2 .

The F test asks whether all four slopes could be zero together.

$F = 11.59$ on 4 and 110 df, $p < 0.001$.

Rejected. But it does not say *which* predictors matter: tuition and study hours on their own do not.



A categorical predictor: residence as dummy variables

Residence has three groups. Code it as **two** 0/1 columns:

	Mess	Home
Hall (reference)	0	0
Mess	1	0
Family home	0	1

Three groups need two columns. The group with no column is the reference.

Stress on screen time and residence:

	<i>b</i>	<i>p</i>
Mess vs hall	0.81	0.356
Home vs hall	0.90	0.361

Each is a gap from hall students, at the same screen time.

Never residence = 1, 2, 3 as one number

Session 5's first wrong choice. The codes are names; a slope for them means nothing.

Simulated class data. Screen-time row not shown.



What a linear regression assumes

-
- L Linear:** the average outcome follows a straight line in each predictor
 - I Independent:** one student's answers do not drive another's
 - N Normal residuals:** matters most in small samples
 - E Equal spread:** the residuals scatter as widely at every fitted value
-

Normal what?

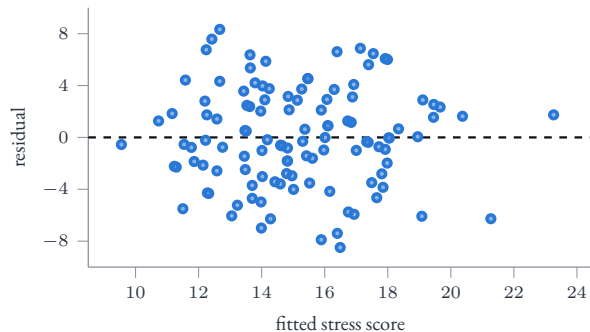
The assumption is about the **residuals**, not the predictors. Skewed screen time is fine as a predictor.

Independence is decided by the design, not by a plot: friends filling in the survey together break it.

The four conditions are the standard textbook set; the LINE initials are a memory aid.



Checking the assumptions: the residual plot



Residuals from the four-predictor model against the scores it predicts.

What we hope for

A shapeless band around 0, about as wide on the left as on the right.

What would worry us

A curve (L fails), a funnel (E fails), or a few points far out on their own.

Ours

No curve, no funnel, no point far from the rest. The model's assumptions look reasonable here.

Simulated class data, $n = 115$.



When predictors overlap: the variance inflation factor

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

$$VIF = \frac{1}{1 - R^2_{\text{this predictor on the others}}}$$

If the other predictors already say most of what one predictor says, the model cannot tell their effects apart, and the standard errors grow.

A VIF of 1 means no overlap. The standard error is inflated by $\sqrt{VIF}$: for sleep, $\sqrt{1.23} \approx 1.11$, about 11% wider.

Ours are all close to 1.

Two versions of one thing

Weekend and weeknight sleep together ($r = 0.74$) would overlap badly. Pick one.

Simulated class data. VIFs by software; the 1.11 is our arithmetic.



1. **Only for what is in the model.** Anything unmeasured, or measured badly, is still free to confound. Stress from family pressure is not in ours.
2. **It can remove what you wanted to see.** 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.
3. **It is still an association.** Adjusting does not turn a survey into an experiment.

Before fitting

Draw the arrows you believe in, then choose what to adjust for. Choosing after seeing which predictors are significant is Session 5's many-tests problem again.

Session 5, Act III, for the four explanations of an association.



Act III

Yes or no outcomes: logistic regression

When the outcome is yes or no

Outcome: a CGPA of 3.00 or above.

Coded 1 for yes, 0 for no, as Session 2 set up.

57 of the 105 students with a result: 54.3%.

Question: do study hours and screen time predict who gets there?

Yes or no outcomes are everywhere:

- ▶ passed or failed
- ▶ depressed or not, on a cut-off
- ▶ recovered or not
- ▶ bought or did not

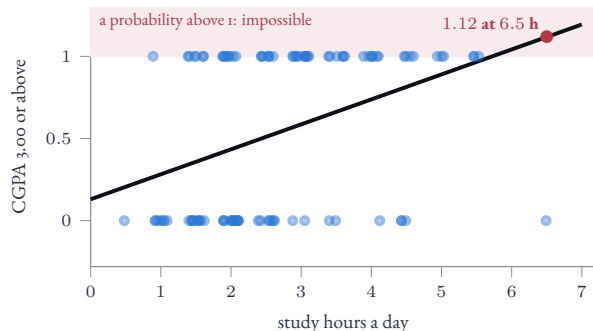
The mean of a 0/1 variable is a proportion

Average the 1s and 0s and you get 0.543, the share with 3.00 or above. So “predict the outcome” now means “predict a probability”.

Simulated class data. Students without a CGPA yet (first years) are left out.



Why a straight line fails



Every student is a dot at 0 or at 1. A least-squares line through them:

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

At 6.5 hours, the busiest student in the data, it predicts a probability of 1.12.

A line has no ceiling. A probability does.

Not just the ends

A line also says every extra hour adds the same 0.152, whether you start at 1 hour or at 5. Near 0 and 1, that cannot be true.

Simulated class data. Dots nudged sideways by up to 0.12 hours so tied answers stay visible.



$$\text{odds} = \frac{\text{number who did}}{\text{number who did not}} = \frac{p}{1-p}$$

Residence and tuition, from Session 5:

	Tutor	Not	Odds
Mess	32	11	2.91
Hall	17	30	0.57

Odds ratio, mess against hall:

$$\frac{2.91}{0.57} = 5.13$$

Mess students' odds of tutoring are about five times hall students'.

1 means the same odds in both groups: no association.

Odds are a betting word

Odds of 3 means “3 to 1”: three yes for every no, a probability of 3 in 4.

Simulated class data. The odds ratio is from the unrounded odds; $2.91 \div 0.57$ gives 5.1.



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

Log odds = $\ln(\text{odds})$, the natural log.

Probability is stuck between 0 and 1.

Odds are stuck above 0.

Log odds run **from minus to plus infinity**, and are symmetric: 0.25 and 0.75 are -1.10 and +1.10.

So put the straight line on the log odds:

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

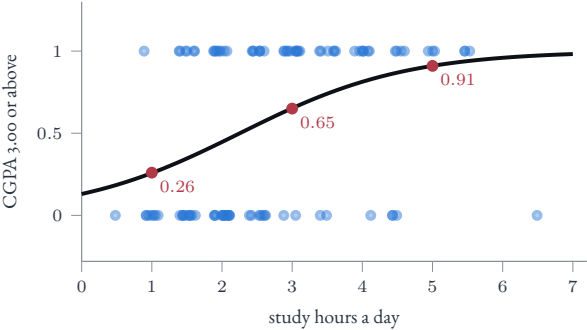
That is logistic regression

A linear model for the log odds. The curve on the next frame is that line, turned back into probabilities.

Arithmetic, not data.



The logistic curve



The fitted model, drawn for a student with the median screen time, 4.4 hours.

Study hours	Probability
1	0.26
3	0.65
5	0.91

An S-shape that flattens near 0 and near 1, and never crosses either.

Simulated class data, $n = 103$ students with a CGPA and both hour answers. Fitted by maximum likelihood (software).



	b (log odds)	Odds ratio 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

Outcome: CGPA 3.00 or above. $n = 103$, of whom 56 reached it. SPSS labels the columns B and Exp(B).

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: a fall of about 39%, study hours held fixed.

Simulated class data. The 39% is our arithmetic, $1 - 0.61$.



An odds ratio multiplies the odds, not the probability

Screen time at 4.4 hours:

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, many different changes in probability.

“Each hour adds 2.32 to the chance”

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

Simulated class data. Odds and probabilities by software; the gains are our arithmetic from the rounded probabilities.



An odds ratio is not “times as likely”

Session 5's table: seniors and juniors who tutor

Seniors, juniors	61.8%, 32.7%
Ratio of percentages	1.89
Odds ratio	3.33

A rare outcome: 2% against 1%

Ratio of percentages	2.00
Odds ratio	2.02

“Times as likely” is the ratio of the percentages, the **risk ratio**.

When the outcome is rare, the odds ratio is close to it.

When the outcome is common, the odds ratio is further from 1, sometimes much further.

A very common misreading

“OR = 3.33, so seniors are 3.33 times as likely to tutor.” They are 1.89 times as likely.

Simulated class data for Session 5's table. The 2% and 1% example is our arithmetic.



How many predictors can a logistic model carry?

Peduzzi and colleagues simulated logistic models with fewer and fewer **events per variable** (EPV).

At 10 or more, “no major problems occurred”. Below 10, coefficients were biased and intervals had the wrong coverage.

Count the **smaller** outcome group:

Our model	$47 \div 2 = 23.5$
-----------	--------------------

Islam et al., depression	$84 \div 8 \approx 10.5$
--------------------------	--------------------------

Islam: 84 students were not depressed; age takes two coefficients, so eight in all.

It is the smaller group that counts

476 students sounds large. With 84 in the smaller group, eight coefficients is about the limit.

Peduzzi et al. 1996, abstract. Our model: **simulated** class data. Islam et al. 2020, Table 1. The EPV figures are our arithmetic.



Act IV

Survey scales: one score from five statements



Stat.BD

Five statements, one score

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 item is **ordinal**: five ordered answers. Together they make a **scale**, one score for one idea, study stress.

s4 points the other way: agreeing means *less* stress. It is reversed before adding:
6 — answer.

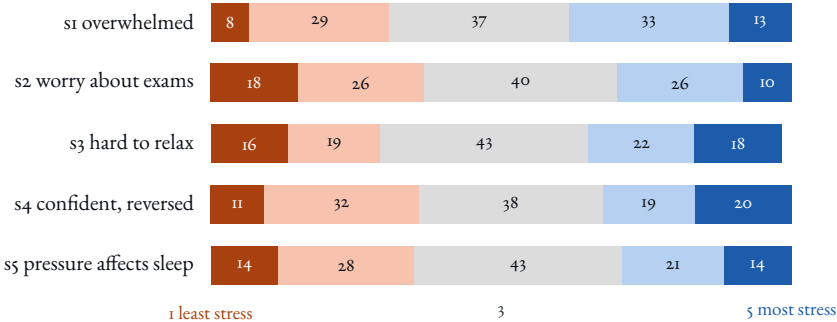
Written for teaching

These five items are not a published, validated stress scale. The method is what transfers.

The class survey, questions 10 to 14, from Session 1.



How the class answered



Counts of students giving each answer. s4 is shown *after* reversing, so dark blue always means most stressed.

Every item uses the whole range, and the middle answer is the most common on each. For one item, report this distribution and the median.

Simulated class data. s3 has two missing answers.



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.

Simulated class data, the second row of the dataset.

Without the reversal

The total would be 19. This student said they feel *confident*, and the unreversed score counts that as stress.

A missing answer on any item leaves the total missing. That is our rule; the class notes discuss others.



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

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

Same information: the average is the sum divided by 5. The average reads on the answer scale, so 3.40 is “between neutral and agree”.

One item has five ordered values: ordinal, median.

The total of five has 21 possible values and a roughly symmetric spread. Treating it as numeric, with means, SDs and regression, is common practice and usually reasonable.

That is what Sessions 2 to 6 have done.

Why five items beat one

Each answer carries some noise: a bad day, a misread word. Across five items the noise partly cancels. Only if the items measure the same thing, which is the next question.

Simulated class data. Class mean of 118 scores, SD 4.57; the 3.01 is our arithmetic, $15.04 \div 5$.



Do the items move together? Item–total correlations

Correlate each item with the total of the *other four*:

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

Leave the item out of the total first; otherwise it is partly correlated with itself.

All five are similar and clearly positive. A student who scores high on one item tends to score high on the rest.

An item near 0, or negative, would be measuring something else, or scored the wrong way.

Simulated class data, $n = 118$ students who answered all five. SPSS calls this the “corrected item–total correlation”.



$$\alpha = \frac{k}{k-1} \left(1 - \frac{\text{sum of item variances}}{\text{variance of the total}} \right)$$

Item variances: 1.19, 1.37, 1.51, 1.48, 1.33. Sum 6.89.

Variance of the total: 20.91.

$$\alpha = \frac{5}{4} \times (1 - 6.89 \div 20.91) = 1.25 \times 0.670 = \mathbf{0.838}$$

α runs up to 1. Ours, 0.838, says the five items vary together strongly: high internal consistency.

Why it works

If the items were unrelated, the variance of the total would equal the sum of the item variances, and α would be 0. When items rise and fall together, the total spreads much more widely.

Simulated class data, $n = 118$. Variances by software; the rounded item variances add to 6.88. Named after Cronbach (1951).



Forget to reverse s4: what happens

	s4 reversed	s4 as typed
s4's r with the rest	0.64	-0.64
Cronbach's α	0.838	0.327

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 signal

Before anything else, check that the item was reversed.

Simulated class data, $n = 118$.



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

Drop	α without it
s1	0.813
s2	0.804
s3	0.798
s4	0.805
s5	0.809
All five	0.838

Dropping any item **lowers** α . Every item earns its place.

An item whose removal *raised* α would be a candidate to rewrite or drop.

s1 to s3 alone give $\alpha = 0.761$. Fewer items, lower α , even though those three agree with each other well.

Do not prune just to push α up

Deleting items to chase a number, then reporting the new α on the same data, overstates how well the scale works.

Simulated class data, $n = 118$. SPSS prints the left column as “Cronbach’s Alpha if Item Deleted”.



What alpha does not tell you

- ▶ **Not that the items measure one thing.** A scale mixing two ideas need not have a lower α .
- ▶ **Not that the scale measures stress.** Consistent is not the same as valid.
- ▶ **Not a property of the questionnaire.** It belongs to these scores, from this sample. Compute it again each time.
- ▶ **Not “higher is always better”.** Above about 0.90, items may be repeating each other.

Benchmarks

Published advice on acceptable values ranges from 0.70 to 0.95. Treat these as conventions, not tests.

So: do our five items measure one thing? Alpha says they hang together. Showing they measure a single idea takes more, usually factor analysis.

Tavakol & Dennick 2011, *Int J Med Educ*; PMID 28029643. Our α : **simulated** class data.



Act V

Reading a published regression table

Islam 2020, Table 3: the depression model

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

Read one row. $B = 0.585$ is on the log-odds scale;
 $e^{0.585} = 1.795$. Holding the other seven fixed, students
 who felt they were lagging had about 1.8 times the *odds* of
 depressive symptoms.

† The printed p disagrees with its own interval.
 Rebuilt in Session 4: 0.828 and 0.466.

Islam MA et al. *PLoS One* 2020, Table 3, depression columns, as printed; reference groups in brackets. PMID 32845928. CC BY 4.0. Daggers and rebuilt p : ours.



“1.8 times more likely to be depressed”?

The paper's Results turn the lagging row into “1.8 times ... more likely to be depressed”. Its own Table 2 counts:

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

Risk ratio: $84.1 \div 77.1 = 1.09$. Crude OR: 1.57.
Adjusted OR: 1.795.

Apply the adjusted OR to the non-lagging odds:
 $3.37 \times 1.795 = 6.05$, a probability of 85.8%.
So “1.8 times the odds” is about 77.1% against 85.8%. Nothing like 1.8 times as likely.

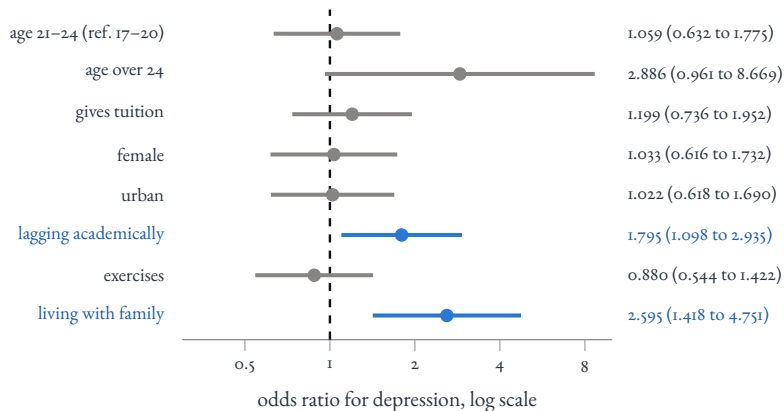
When the outcome is common

With 82% depressed, the OR is far from the risk ratio. “Times as likely” needs a risk ratio.

Islam 2020, Results, Tables 2 and 3. Every figure except 1.795 is **our arithmetic**; the 85.8% applies an adjusted OR to a crude baseline, as an illustration.



Eight intervals, and what they can and cannot say



Age over 24: 2.886, but 0.961 to 8.669. Only 42 students. The data fit no difference and a large one. Not significant is not “no effect”.

Islam 2020, Table 3, depression, as printed, on a log scale. Limitations section, paraphrased.

One survey, one moment. The authors call these associated factors, which could be causes or results of depression. Two intervals clear 1, in blue.



Check a row against the counts: living with family

Table 2 counts, depression

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

Crude OR, yes against no: 0.77 (0.17 to 3.52), SE 0.77.

Table 3 prints: 2.595 (1.418 to 4.751), $p = 0.002$; its interval implies SE 0.308.

- **Direction:** the counts show slightly *less* depression with family.
- **Precision:** with 2 students in one cell, an SE near 0.31 is hard to believe.
- **Anxiety:** all 14 living apart had symptoms. No finite OR exists; Table 3 prints 1.838.

We cannot tell which table is wrong

Until the authors explain, this row should not be cited.

Islam 2020, Tables 2 and 3, as printed. Crude OR, Woolf interval, SEs and the reading are **ours**.



Activity: one finding from Table 3, written properly

Groups of three. Each group takes one row:

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

Printed p : 0.601, 0.059, 0.019, 0.155.

Write one sentence that:

1. names the comparison group;
2. says *odds*, not “likely”;
3. gives the interval;
4. says “associated with”, not “causes”;
5. passes one check: does p agree with the interval?

4 minutes to write, 2 to read aloud.

Islam 2020, Table 3, as printed. Model answers in the class notes.



What today established

1. A regression line has an intercept and a slope. Residuals and R^2 say how well it fits.
2. With several predictors, each slope holds the others fixed. Screen time's 0.88 became 0.64 once sleep was held fixed.
3. Check the residual plot and the overlap between predictors. Controlling for something cannot fix what was never measured.
4. For a yes or no outcome, logistic regression gives odds ratios. They multiply the odds, not the probability.
5. A scale: reverse, add, then check α . Consistency is not validity. And check any published row against its counts.



Session 7: SPSS, from raw data to results

The class dataset in SPSS: importing it, labelling variables, cleaning, computing the stress score, then descriptive statistics and charts. The session opens with an installation check.

Before then

Read Chapter 6 of the class notes. Try the Part A exercises and the activity answers.

Course syllabus, Sessions 7 and 8.

One thing to watch for

When SPSS computes the stress score, s_4 must be reversed first. You now know what α looks like when it is not.



- ▶ 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. CC BY 4.0.
- ▶ 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.
- ▶ 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.
- ▶ Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;16(3):297–334. doi:10.1007/BF02310555. Cited for the name only; not read for this session.
- ▶ R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- ▶ Class data: **simulated** stand-in for the Stat.BD class survey, generated by data/class-data.R from a fixed seed. Not real students. Every regression, odds ratio and α computed from it is ours. So are the risk ratios, crude odds ratios, Woolf intervals and rebuilt *p*-values from Islam 2020, Tables 2 and 3.

