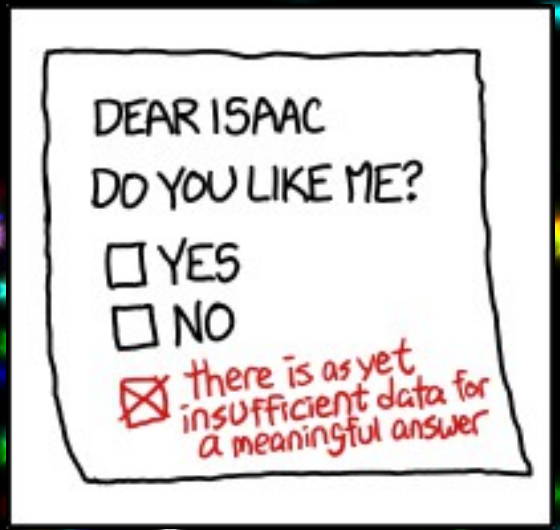
Stats - Introduction to multiple regression Part 1

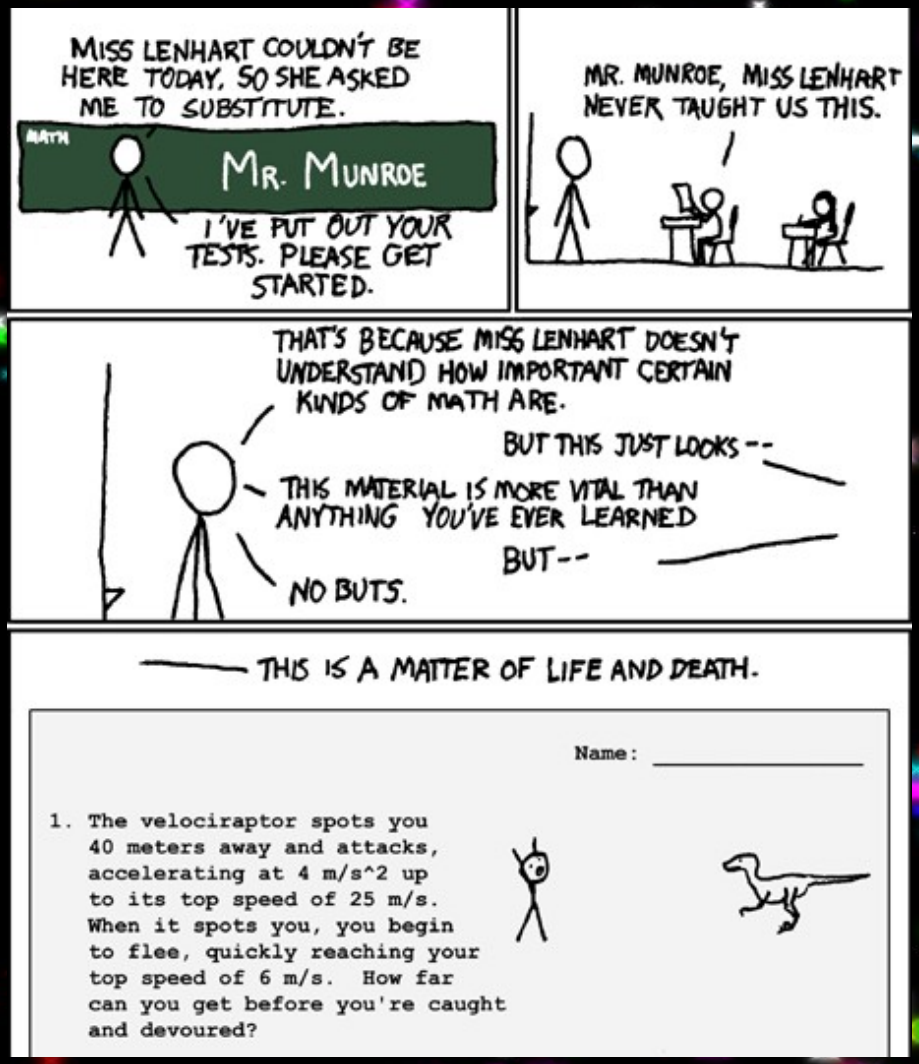
anthony.smith@scu.edu.au

Record Attendance

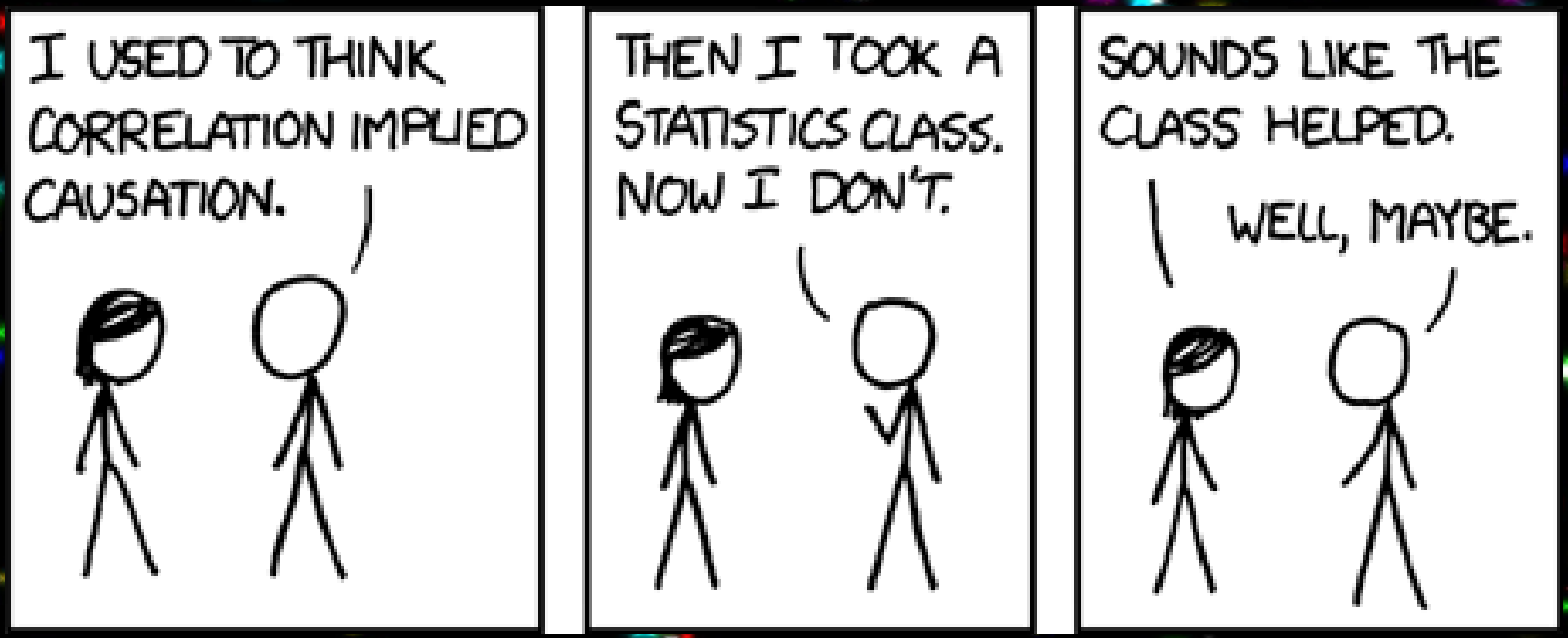
Any Questions About Any Maths At All?



Review the Lecture



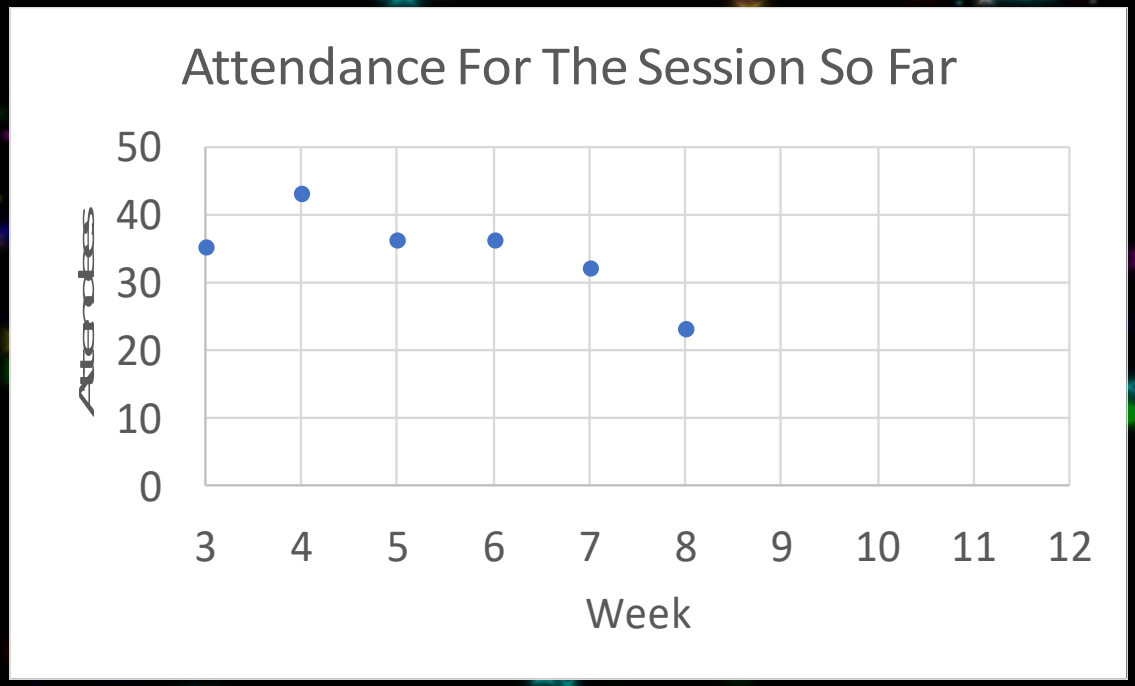
Correlation vs Causation



Research Risks

Attendance Attrition: The Plot Thickens

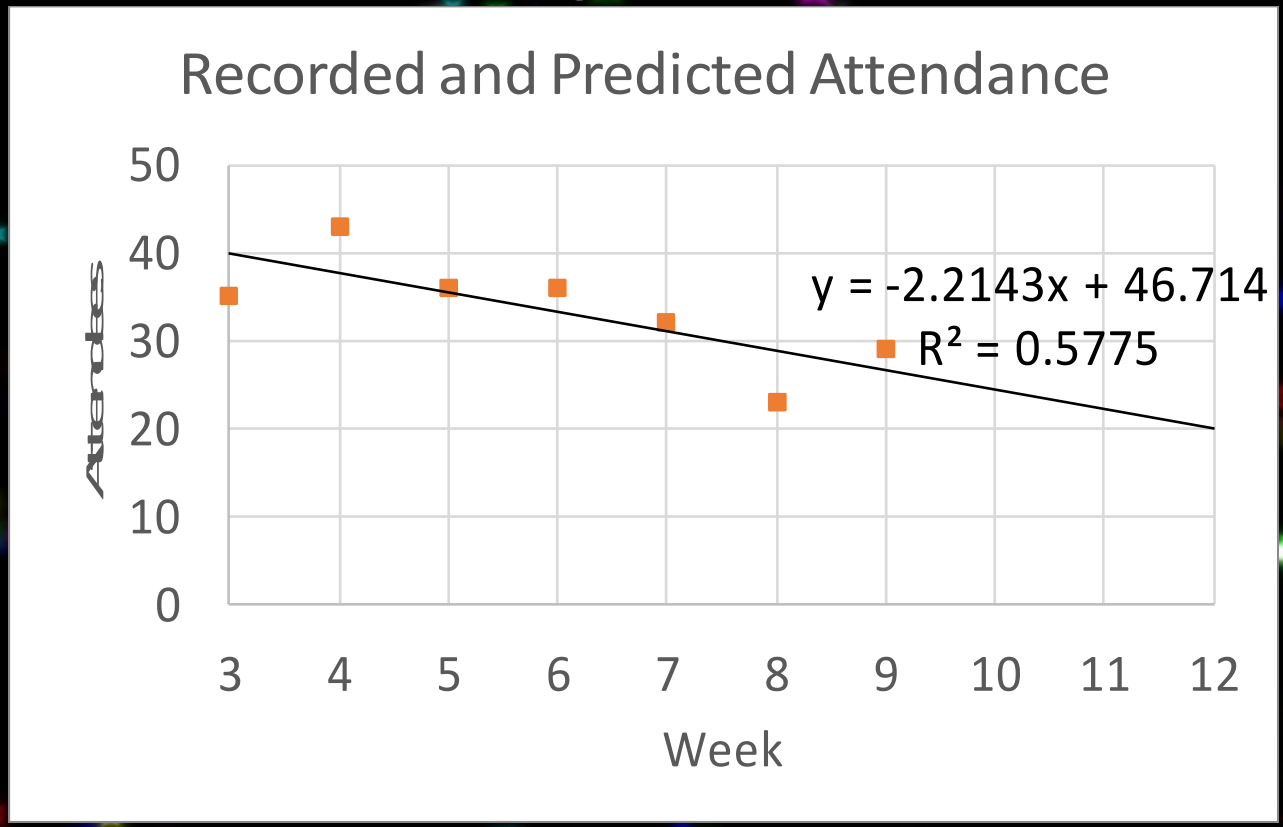
- Last time: evidence of declining attendance throughout the session so far
  - Week 3: 35
  - Week 4: 43
  - Week 5: 36
  - Week 6: 36
  - Week 7: 32
  - Week 8: 23



- BUT attendance in week 9 was 29: an *increase*!

Attendance Attrition: Data

- Updated plot of attendance against time
  - Week 3: 35
  - Week 4: 43
  - Week 5: 36
  - Week 6: 36
  - Week 7: 32
  - Week 8: 23
  - Week 9: 29



- Might there be more going on here than a simple decline with time?
- What else might be happening?

Something Else Going On?

- Theory / Hypothesis: there's a “holiday effect”, whereby students are slow to return from the Christmas/New Year holiday, causing a lower attendance in the first week back
- This isn't in our model, which says there is a linear decline in attendance with time
- How do we fix (or rather, improve) our model?

Use the Multiple Regression Model

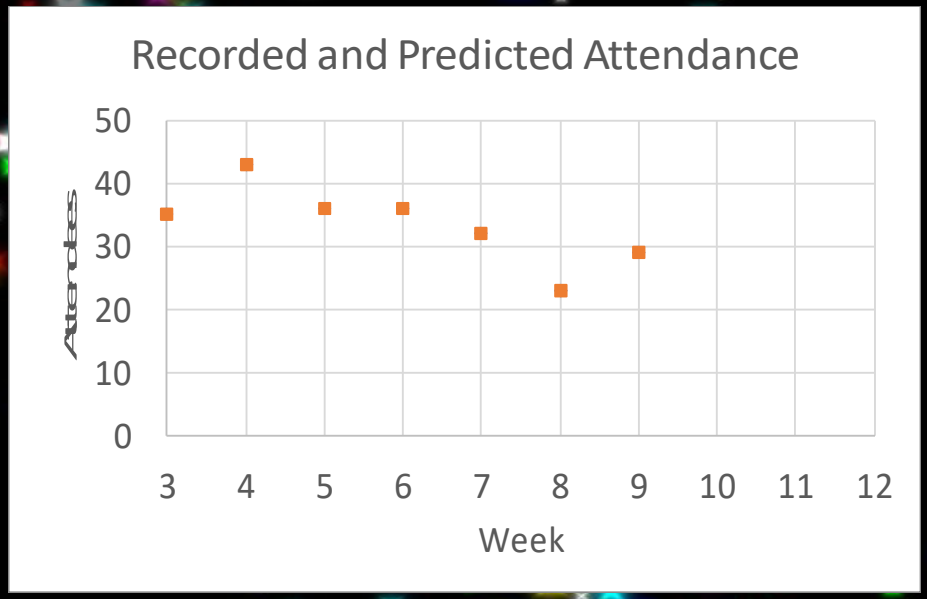
- $y = b_0 + b_1 x_1 + b_2 x_2$
- We know what some of these variables/coefficients are from last time
- y = number of attendees
- b_0 is the vertical intercept at (x_1, x_2, . . .) = 0, as before
- x_1 is the week number, a proxy for time
- b_1 is the regression coefficient corresponding to the week number (that tells us how sensitive y is to the number of weeks that have passed)
- What do we do with b_2 and x_2?

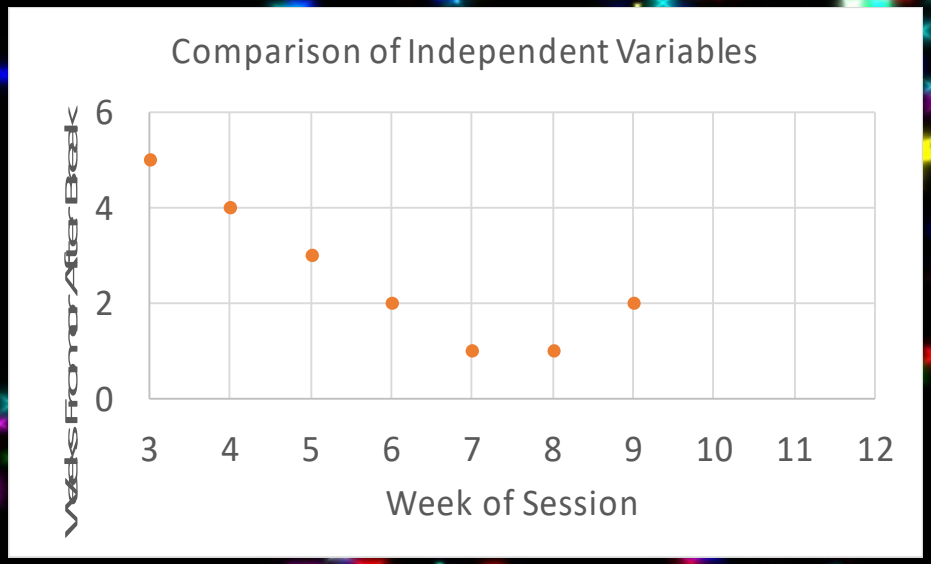
Defining x_2 and b_2

- Additionally, maybe people were leaving to go on break early, accounting for the decline before the break
- Let's make x_2 the number of weeks since or before the Christmas/New Year break
- Then b_2 is the regression coefficient that tells us how sensitive the attendance is to proximity to the break
- What does this look like?

What this Definition of x_2 Looks Like

- A “V” shape to account for the shape of the attendance data
- Maybe we can obtain a better fit to the data by modelling two effects instead of just one: decline with time into the session, and the “holiday effect”





Aside: With Code

- For those of us doing IT
- You don't have to do it in this unit, but for situations in which readily-available tools aren't up to the task, you have to write your own code (potentially for statistics that you have to devise yourself)
- How might we do it?

Aside: With Code

- For those of us doing IT
- You don't have to do it in this unit, but for situations in which readily-available tools aren't up to the task, you have to write your own code (potentially for statistics that you have to devise yourself)

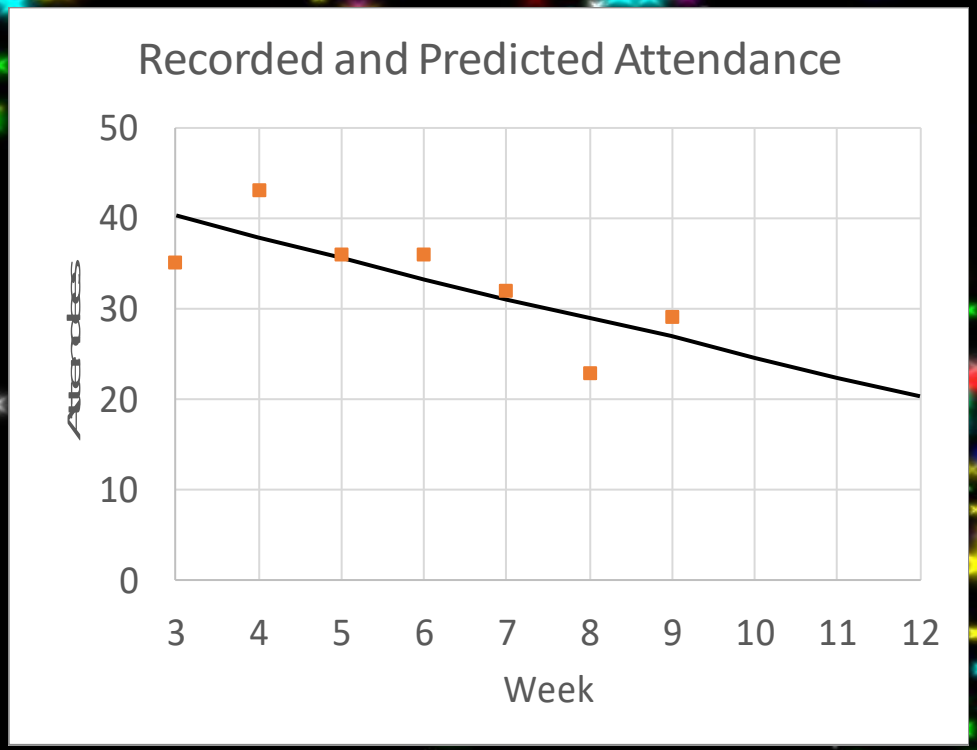
```
weeks=[3,4,5,6,7,8,9];
weekssincebreak=[5,4,3,2,1,1,2];
data=[35,43,36,36,32,23,29];
results=[];
for b0=44:0.1:46 % intercept
        for b1=-2.5:0.01:-1.5 % weeks
                for b2=0.1:0.01:0.3 % weeks since break
                        prediction=b0+b1*weeks+b2*weekssincebreak;
                        squares=(data-prediction).^2;
                        sumofsquares=sum(squares);
                        result=[b0,b1,b2,sumofsquares];
                        results=[results;result];
                end
        end
end
```

Then We Find

- $b_0 = 45.5$
- $b_1 = -2.10$
- $b_2 = 0.20$
- $y = 45.5 – 2.1x_1 + 0.2x_2$
- What might this look like?

What This Looks Like

- $y = 45.5 – 2.1x_1 + 0.2x_2$
- Not much has changed
- Why? How can we make this adaptation more meaningful?

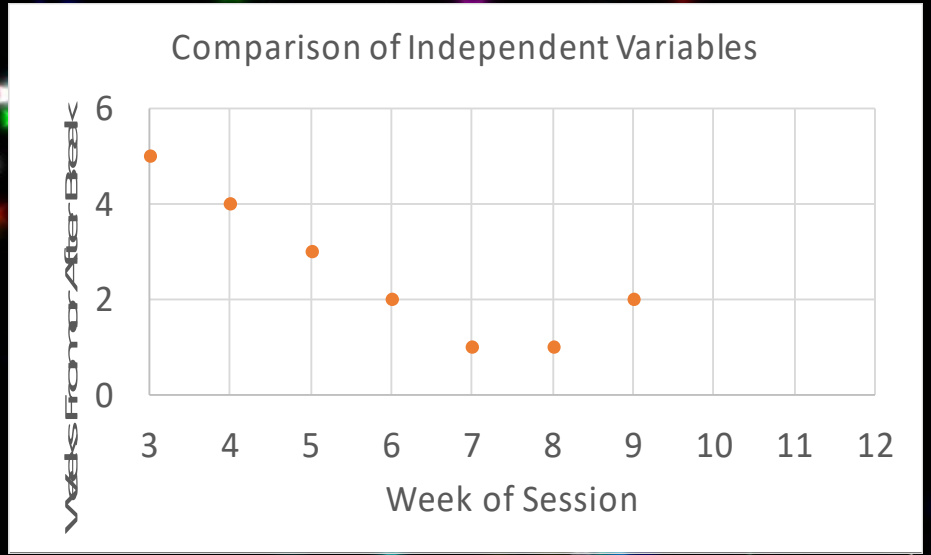


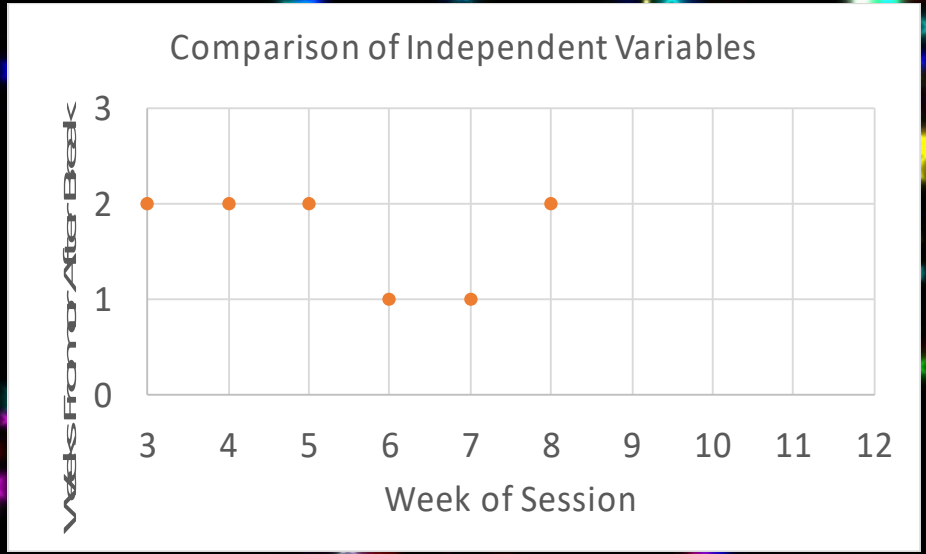
Improving the Model (Again)

- It's not realistic to model the holiday having an effect 5 weeks before it happens, close to the beginning of the session
- Let's try the same thing, but with no effect in weeks 3-5

What this New Definition of x_2 Looks Like

- Constrain the "V" shape to 1 week either side of the holiday
- Effectively models the holiday as a "blip" in the data
- There is probably some effect in the second week either side of the break as well! But we might need a nonlinear function to model it



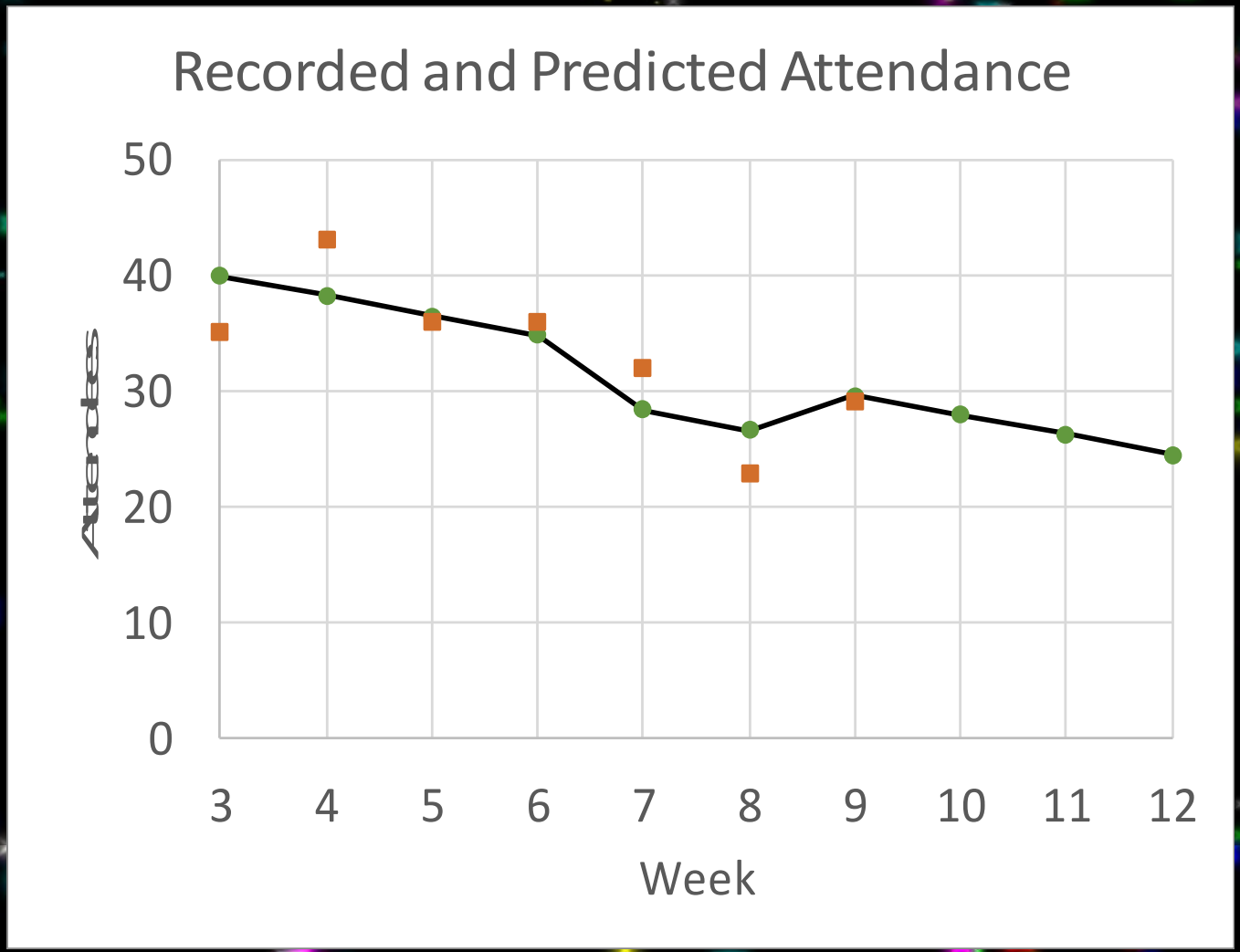


Then We Find

- $b_0 = 35.6$
- $b_1 = -1.71$
- $b_2 = 4.72$
- $y = 35.6 - 1.71x_1 + 4.72x_2$
- Think this will look closer?

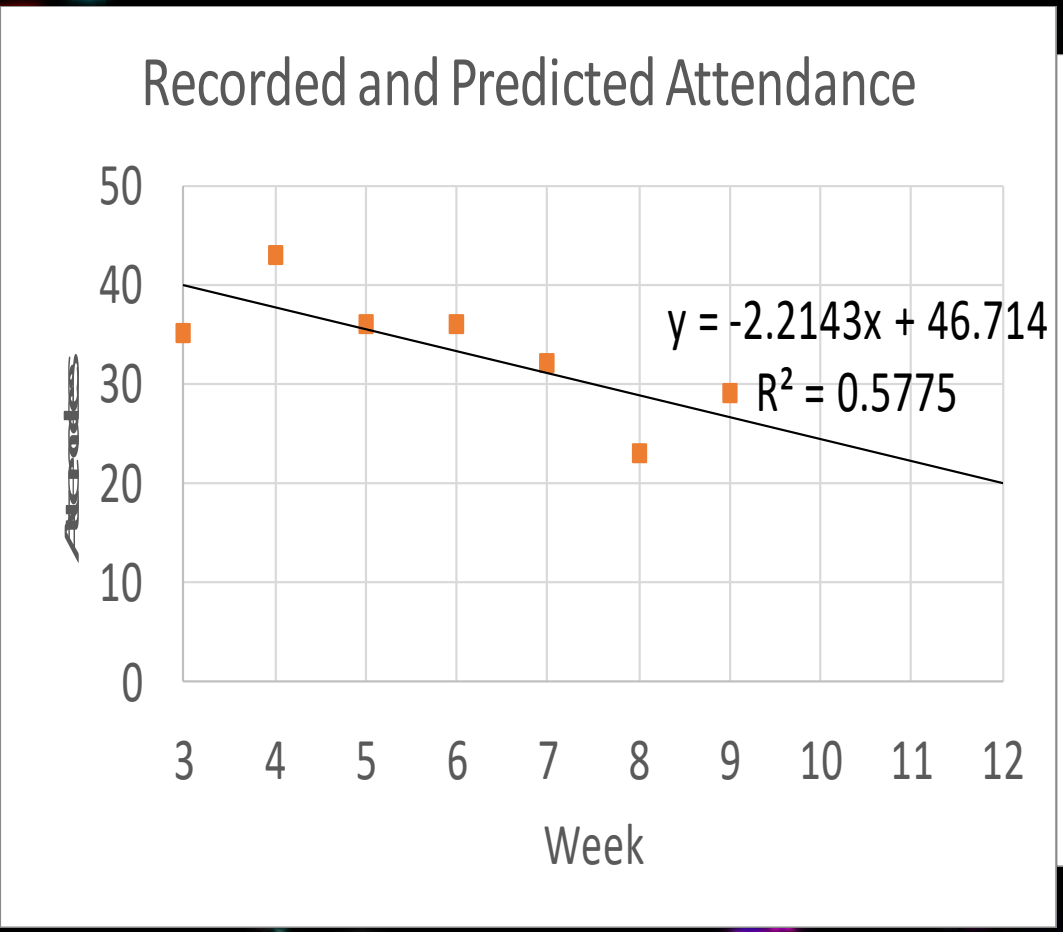
What This Looks Like

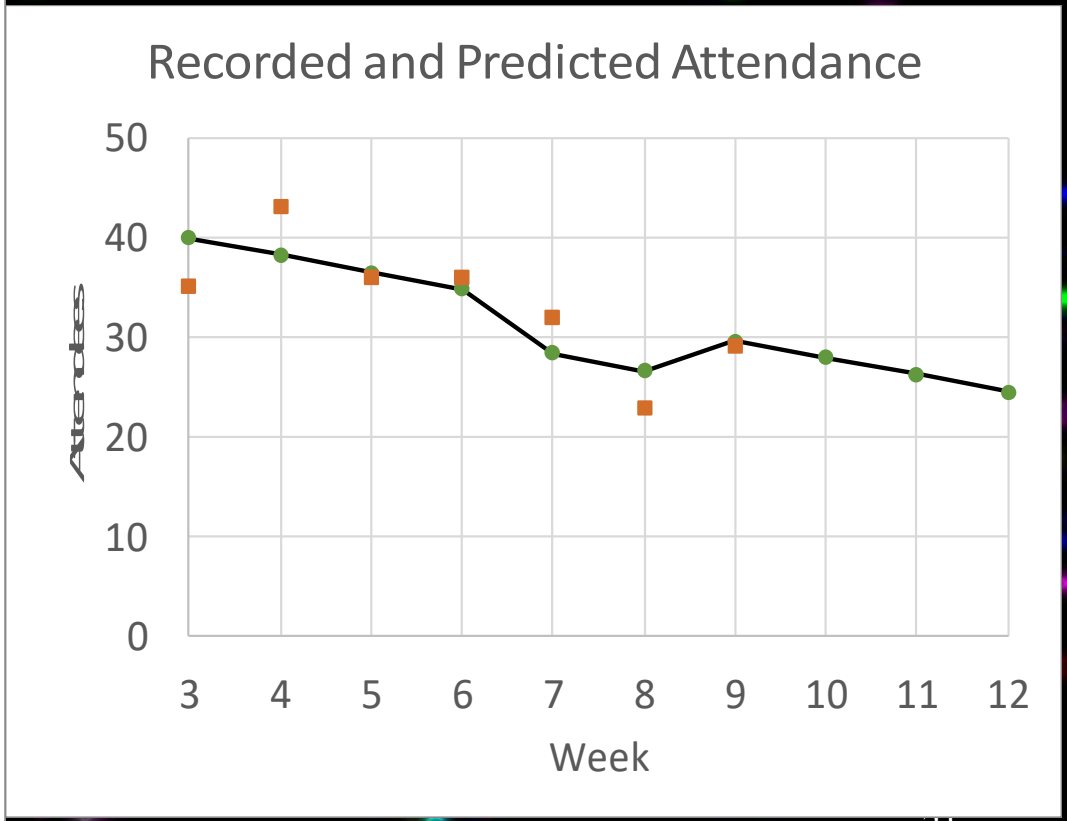
- $y = 35.6 - 1.71x_1 + 4.72x_2$



Is this an Improvement?

- The mere fact that x_2 is nonzero says this is, but side-by-side with the simple linear fit:





Jump They Say

- In scenarios analysed with multiple regression, a variable is influenced by two or more others
- This is like being told to do two different things by different people
- David Bowie wrote a song about his half-brother Terry Burns, who committed suicide by jumping onto train tracks
- Bowie imagined that Terry, a schizophrenic, could hear voices in his head telling him to jump ('*They say jump*')
- In the song, Bowie wishes he could have been there to tell his brother not to ('*I say he should watch his ass*')

Jump They Say

They say 'Hey that's really something'

They feel he should get some time

I say he should watch his ass

My friend don't listen to the crowd

They say 'Jump'

They say 'Jump'

Watch Out!

Watch Out!

- David Bowie, "Jump They Say" (*Black Tie White Noise*)

Questions, suggestions for next time:
anthony.smith@scu.edu.au

These slides available from the usual place

Any Questions About Any Maths At All?

Credits

- All comics from xkcd, the webcomic of romance, sarcasm, math and language
  - https://xkcd.com/
- Any songs were by David Bowie