

In-person session 2

August 30, 2021

PMAP 8521: Program evaluation
Andrew Young School of Policy Studies

Plan for today

Regression FAQs

Miscellaneous R stuff

Visualizing data with ggplot2

Transforming data with dplyr

Regression FAQs

**Why use `tidy()`
from the broom package?**

How was the 0.05 significance threshold determined?

Could we say something is significant if $p > 0.05$, but just note that it is at a higher p-value?

Or does it have to fall under 0.05?

**Why all this convoluted
logic of null worlds?**

Different "dialects" of statistics

Frequentist

$$P(\text{data} \mid H_0)$$

"Regular" statistics;
what you've learned
(and are learning here)

Bayesian

$$P(H \mid \text{data})$$

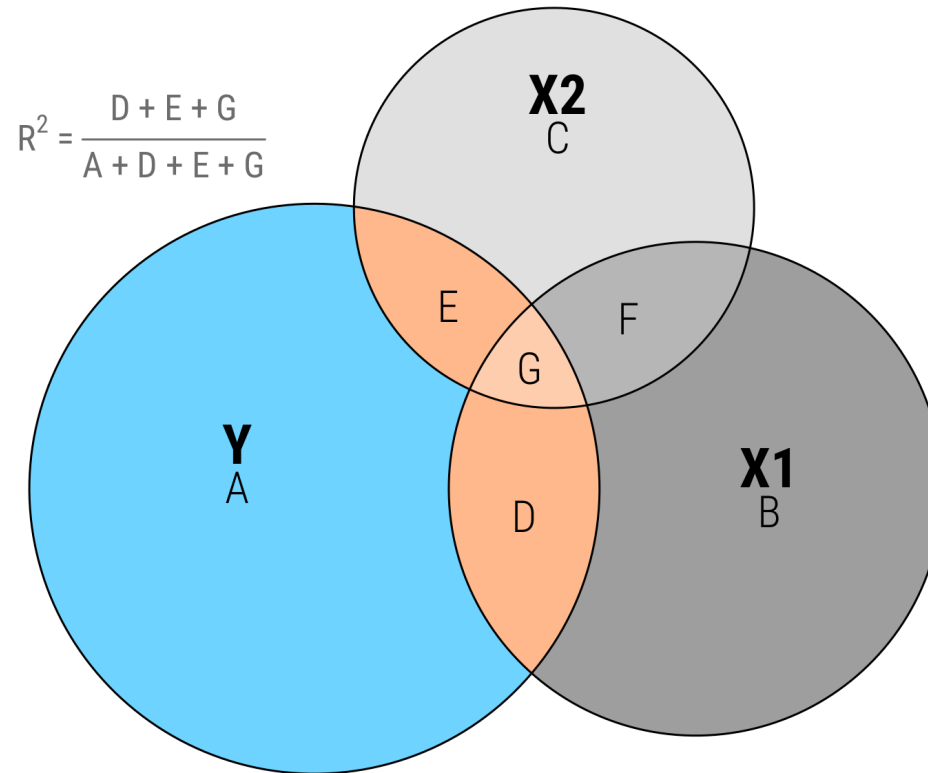
Requires lots of
computational power

**Do we care about the actual coefficients
or just whether or not they're significant?**

How is R^2 related to multiple regression?

R² represented as an Euler diagram

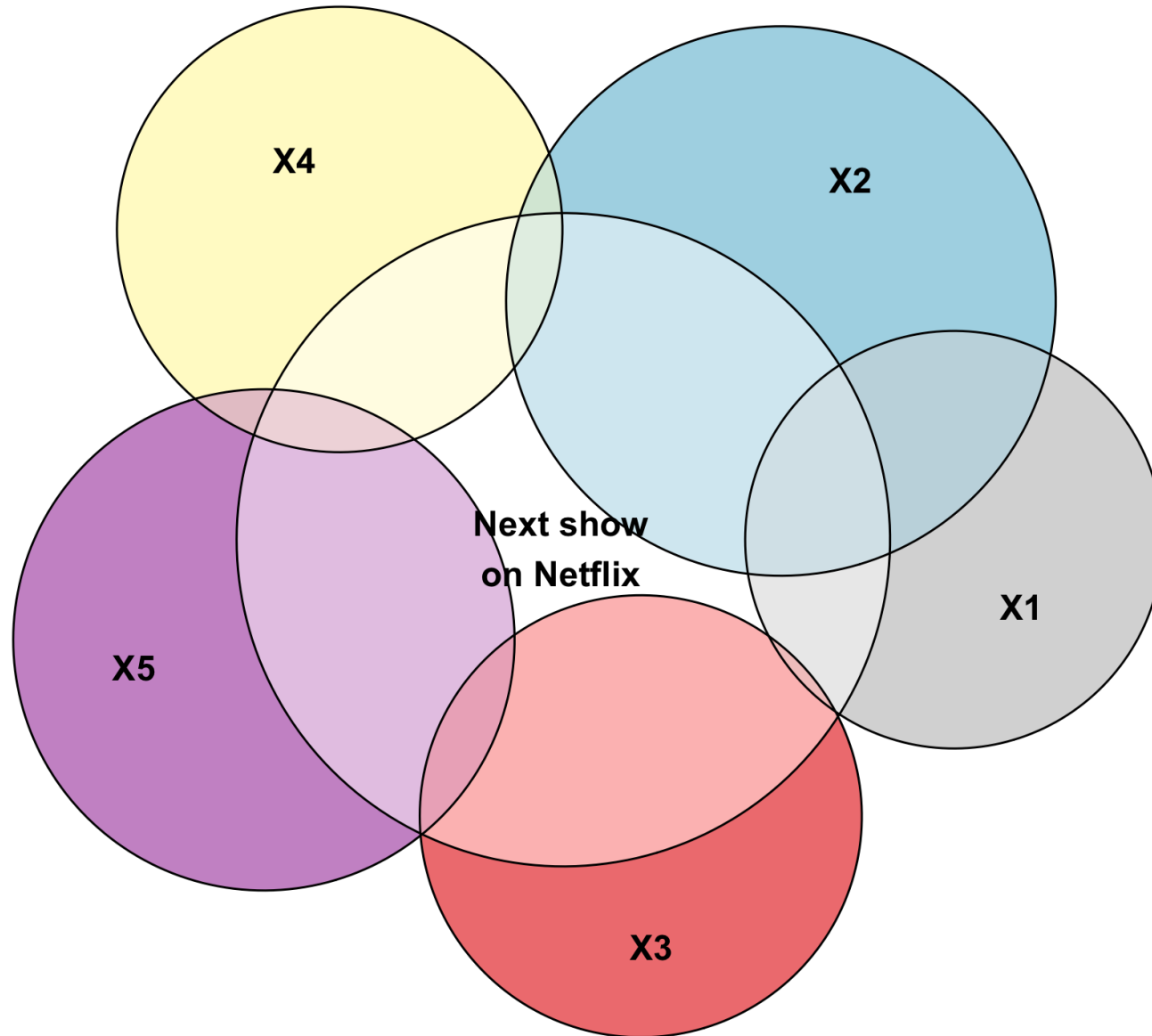
Orange area (D + E + G) shows the total variance in outcome Y that is jointly explained by X1 and X2



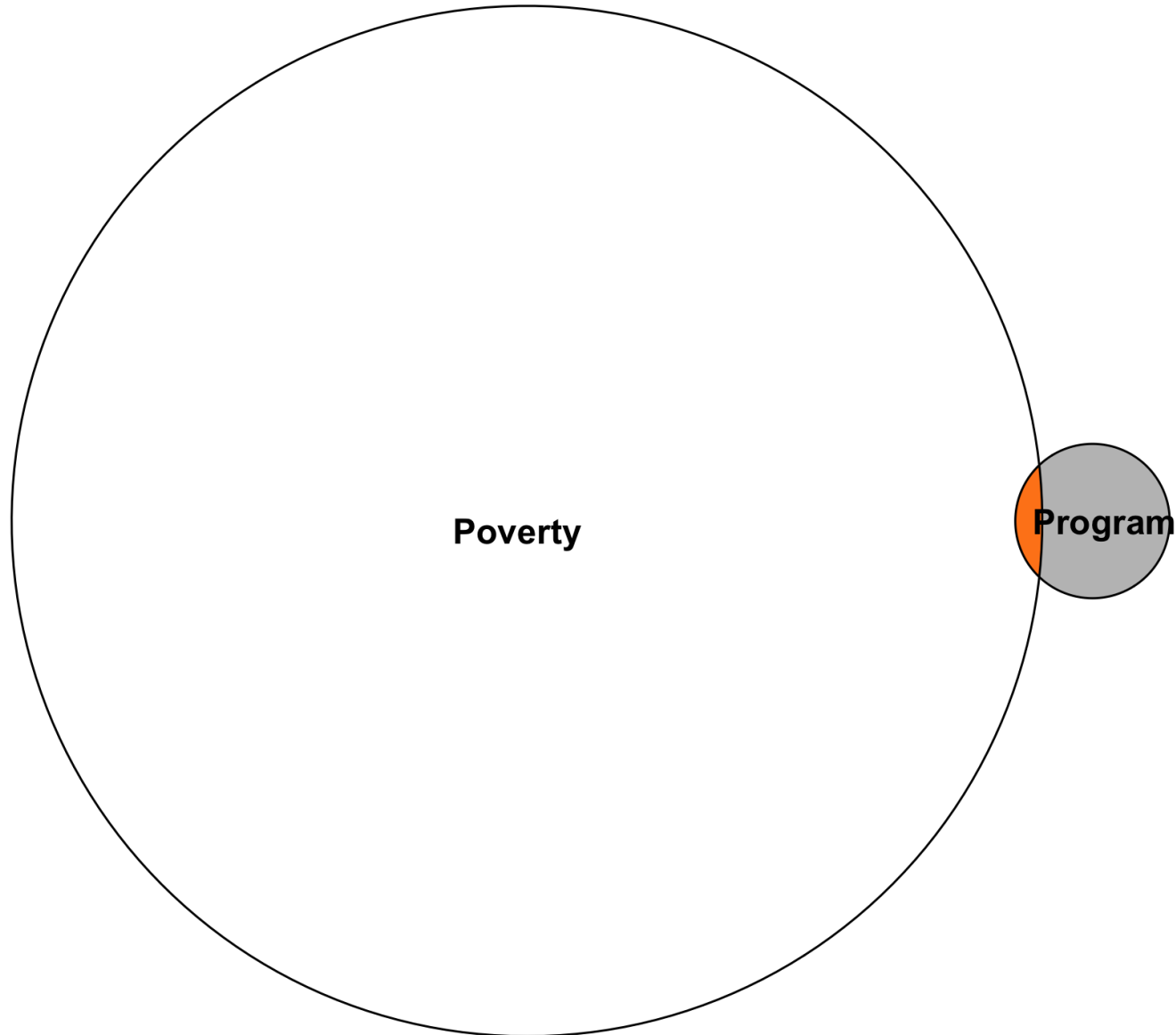
Circles sized according to each variable's sum of squares; size of overlapping areas is not 100% correct due to limitations in available geometric space

Is there a point at which multiple regression becomes inaccurate if you add more variables? Or does it become more accurate if you add more variables?

How many variables does a typical multiple regression usually contain? In published papers, for example.



What counts as a "good" R^2 ?



**If we can't use statistics to assert causation
how are we going to use this information
in program evaluation?**

What does logging do?

Why do we log things?

**What's the difference between
qualitative columns and categorical columns?
Both are text?**

Miscellaneous R stuff

Visualizing data with ggplot2

Transforming data with dplyr