PSY 5939: Longitudinal Data Analysis

Introduction to longitudinal data

Longitudinal data

What is longitudinal data?

Longitudinal data is

- *repeated* observations
- of *multiple units*
- of the *same variable*
- from the *same unit or participant*
- over *time*

What isn't longitudinal data?

Prospective data

- Variable A at time 1 predicts variable B at time 2
- Not measuring the same variable repeatedly

Repeated surveys over time of different people

- e.g., political polls that ask a different group each time
- Not measuring the same unit on each occasion

Time series

- 1 unit measured over time
- Not measuring multiple units

Advantages of longitudinal data

We can answer research questions about **change over time**

- Change over time, such as development or aging over time
- Is change **homogeneous** across all individuals?
- **Predictors of change** at the group or individual level
- **Change as a predictor** of other outcomes

We can also use longitudinal data to clarify time sequencing to strengthen *causal inference*

Challenges of longitudinal data

Collecting, maintaining, and analyzing longitudinal data has particular challenges

- Requires at least 2, preferably 3 or more waves of data
- Outcome of interest may change over time
- Need a sensible metric for time that is measured
- Data management and subject tracking
- Missing data and attrition are more common than in cross-sectional data

Approaches

Statistical models of change

There are *many* types of statistical analyses for longitudinal data

Vary in terms of how they assess change:

- Individual versus group
- Difference versus trajectory
- Discrete versus continuous change
- Software used

This course will cover a variety of methods that are **commonly used in psychology** and other social and behavioral sciences

What we will cover

- Regression-based methods for 2 waves of data
  - Differences, controlling for baseline (ANCOVA), partial change
- Mixed (or multilevel or hierarchical linear or random effects) models
  - Focus on individual change and variability
- Latent growth models
  - Focus on individual change and variability
- Statistical mediation
  - How a causal process unfolds over time

What we won't cover

- Repeated-measures ANOVA
  - Limited utility. Assumptions typically violated
- Auto-regressive and cross-lag models
  - Often interpreted incorrectly. Better alternatives
- Survival analysis
  - Not commonly used in psychology
- Time series analysis
  - We will focus on models for multiple units

Measurement

Measurement issues for longitudinal data

Newsom, Jones, and Hofer, 2012, Chapter 4

> The study of change rests on the assumption that observed differences in measurements over time reflect **true change in the construct** being measured.

Why are you seeing change over time?

- Is the construct actually changing?
- Is how you're measuring the construct changing?
  - In other words, is it changing in a meaningful way?

Measurement issues: Reliability change

- Reliability of the construct may change over time
  - Become more or less reliable over time
- May distort results significantly
  - How much depends on method
- Also, reliability of change itself

Measurement issues: Construct change

- Many constructs, by their nature, change over time
  - e.g., smoking as a measure of deviant behavior (10 vs 20)
- Change over time may be masked by change in the construct
- Construct may be stable but measure may change

Measurement issues: Measuring time

How we measure **time** is absolutely key

- **Actual time** between observations versus **waves**
- Time measurement **varies** across individuals
- **When** we measure impacts whether we can detect change

Data issues

Dataset issues for longitudinal data

There are two ways that your data may be organized

- Person-level structure
  - also called multivariate or wide
  - Typical for cross-sectional data
- Person-period structure
  - also called univariate, tall, or stacked
  - Typical for longitudinal data

Person-level (wide)

- Rows for participants (1 row per **person**)
- Columns for variables
- Different timepoints are different variables / columns
- Typical structure for cross-sectional data
- Preferred structure for latent growth models in Mplus

Wide data

ID	extern1	extern2	extern3	extern4	extern5
1	4	3	4	4	5
2	3	5	4	4	3
3	2	3	3	5	3
4	4	5	3	5	3
5	5	3	4	4	3

Person-period (tall or stacked)

- Multiple rows per participant
- 1 row per participant, **per timepoint**
- Preferred format for most programs for mixed models
- Also easier to include time-varying variables

Tall data

ID	time	extern
1	1	4
1	2	3
1	3	4
1	4	4
1	5	5
2	1	3
2	2	5
2	3	4
2	4	4
2	5	3
3	1	2
3	2	3
3	3	3
3	4	5
3	5	3
4	1	4
4	2	5
4	3	3
4	4	5
4	5	3
5	1	5
5	2	3
5	3	4
5	4	4
5	5	3

Missing data

Missing data in longitudinal studies

"Missingness" refers to whether someone has a value on a particular variable or not

- Missing values on *items* or entire *scales*

Longitudinal studies have an additional issue of **attrition**

- Participants are involved in some waves but then drop out
- These participants may or may not return for later waves

Oversimplified review of missing data mechanisms

Imagine that you have 2 variables: X and Y
Y has some missing values

MCAR: Missingness on Y is unrelated to values of X and unrelated to the (missing) values of Y itself

MAR: Missingness on Y is unrelated the (missing) values of Y, but it is related to values of X

MNAR: Missingness on Y is related to the (missing) values of Y

See Baraldi & Enders (2010), Enders (2010) for more information

- Or take Tim Hayes' Missing Data course

Regression review

Terminology

Terminology for regression

Dependent variable, also called

- Y
- criterion
- outcome
- response variable
- regressand
- explained variable
- effect

Terminology for regression

Independent variable, also called

- X
- predictor
- explanatory variable
- regressor
- manipulated variable
- cause

General linear model

General linear model

General method to predict a continuous outcome

Continuous and/or categorical predictors and regression weights

ANOVA and linear regression are examples of GLM

General linear model

$Y_i = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p + e_i$

or

$\hat{Y}_i = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p$

- Y_i is the **observed** outcome for individual i
- $\hat{Y}_i$ is the **predicted** outcome for individual i
- The Xs are the predictor variables
- The bs are the regression coefficients or weights
- e_i is the error or residual term for individual i

Assumptions

Assumptions 1

Linear regression and ANOVA models are estimated using ordinary least squares (OLS) estimation

In order for OLS estimates of regression coefficients to be the **best linear unbiased estimates (BLUE)** of the population regression coefficients, we make three assumptions about the residuals (e_i):

1. Average or expected value of the residuals is 0
2. Variance of the residuals is constant across values of the predictor and greater than 0
3. **Errors for individual observations are uncorrelated**

Assumptions 2

We make a 4th assumption to allow for statistical inference (i.e., significance tests)

4. Errors have a (conditionally) normal distribution

This 4th assumption means that we can replace assumption 3 with the stronger assumption that individual errors are **independent** (not just uncorrelated)

Independence of observations

One observation's values do not depend on another observations values

Independence is a stronger assumption than uncorrelated

- Independence **implies** NO correlation
- No correlation **does not imply** independence

Non-independence occurs because of clustering of observations in groups (e.g., families, classrooms) or **repeated observations on the same person over time**

Interpreting effects

Interpreting linear regression effects

$\hat{Y}_i = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p$

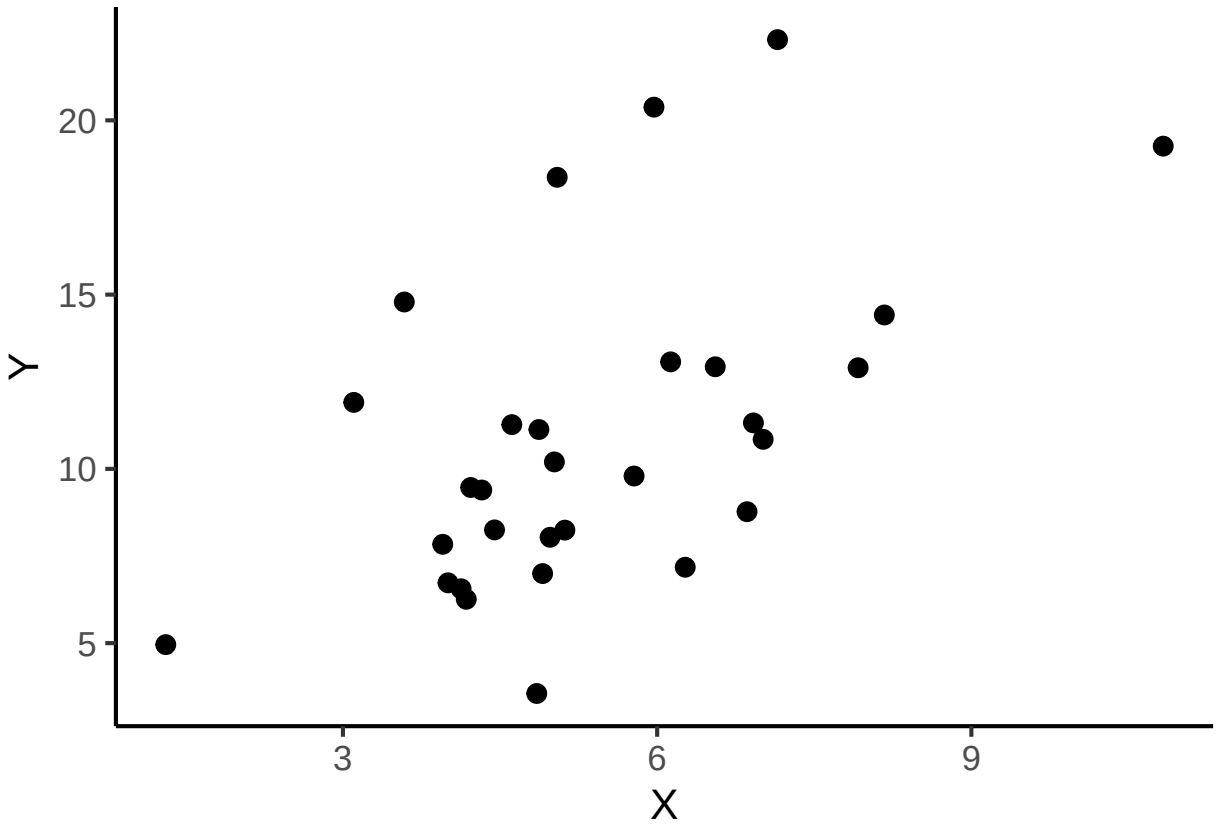
The effect of each predictor is a *partial* effect

- **Controlling for** or **holding constant** the effects of all other predictors

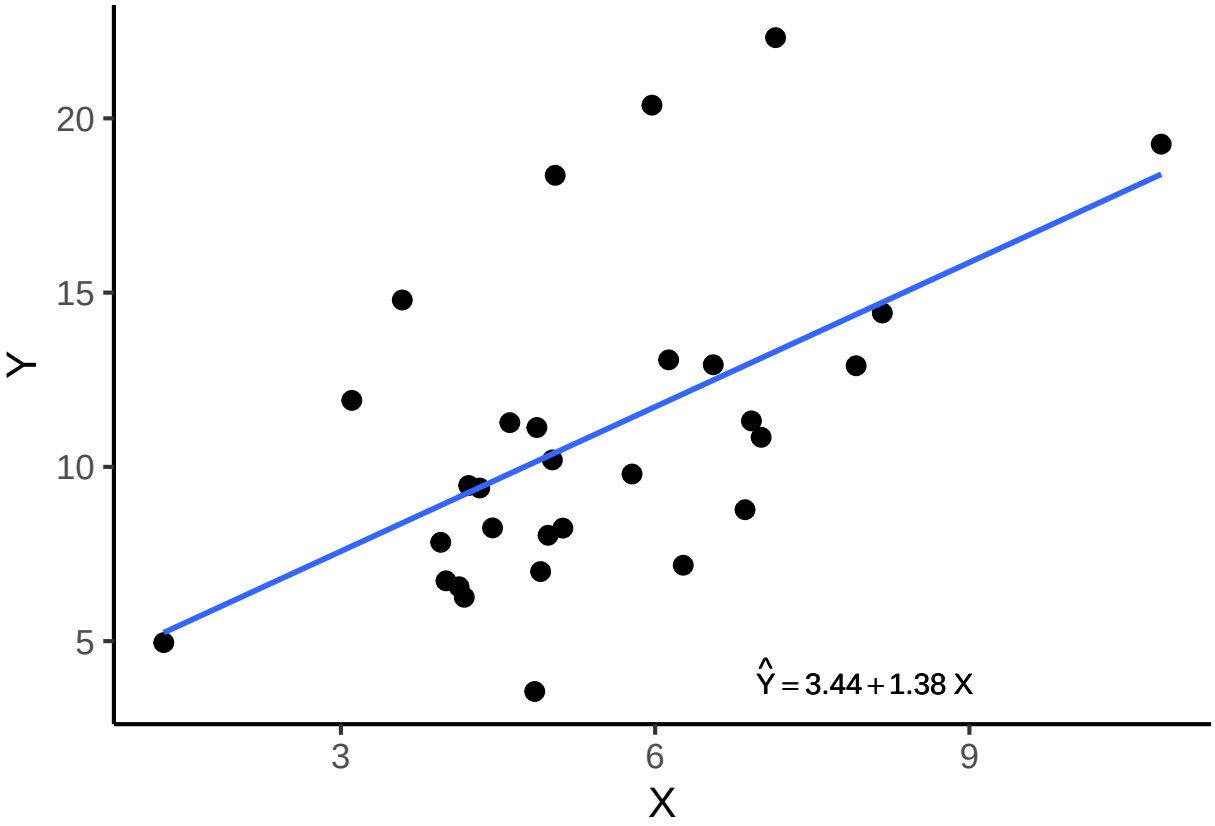
Linear, additive effects

- For a 1 unit increase in X_1, there is a b_1 unit increase in the predicted outcome (**holding all other predictors constant**)

Linear effect



Linear effect



Interactions

Interaction effects in linear regression

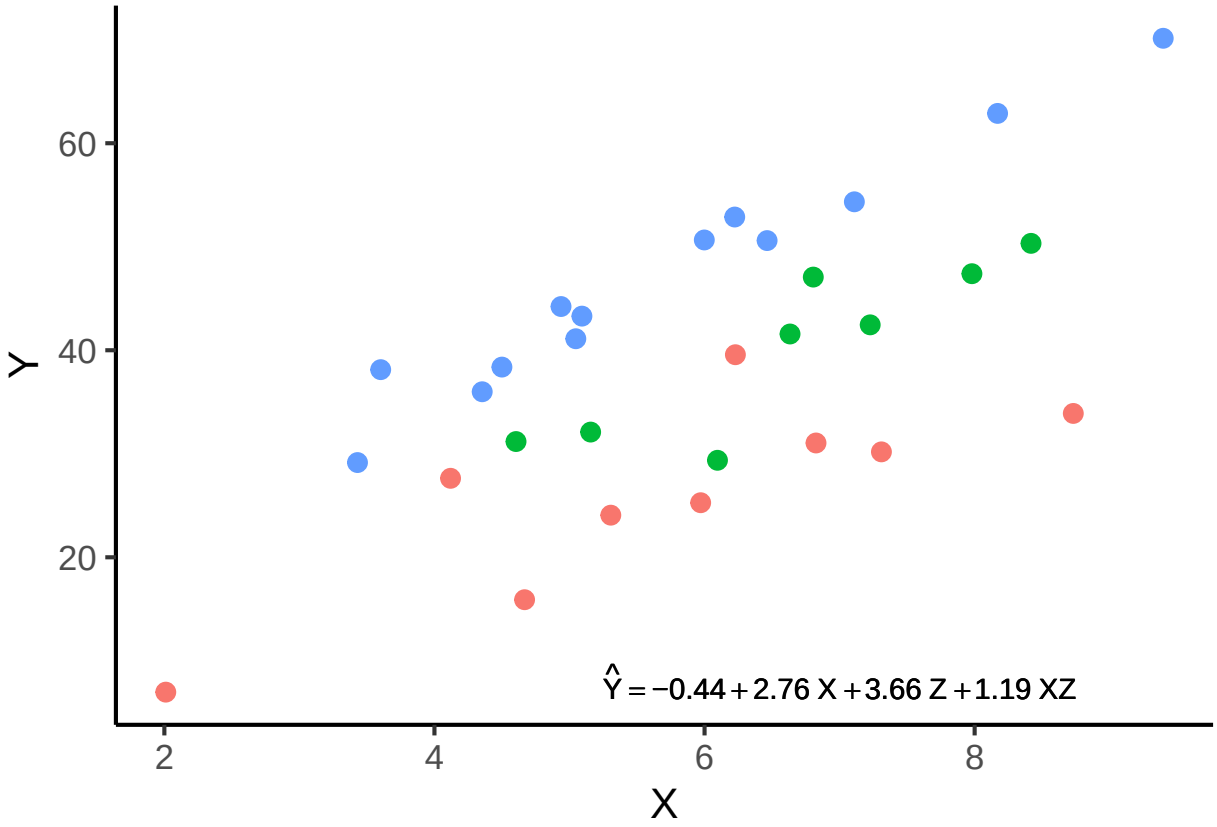
$\hat{Y}_i = b_0 + b_1 X + b_2 Z + b_3 X Z$

Interaction or **moderation** is included as a product term

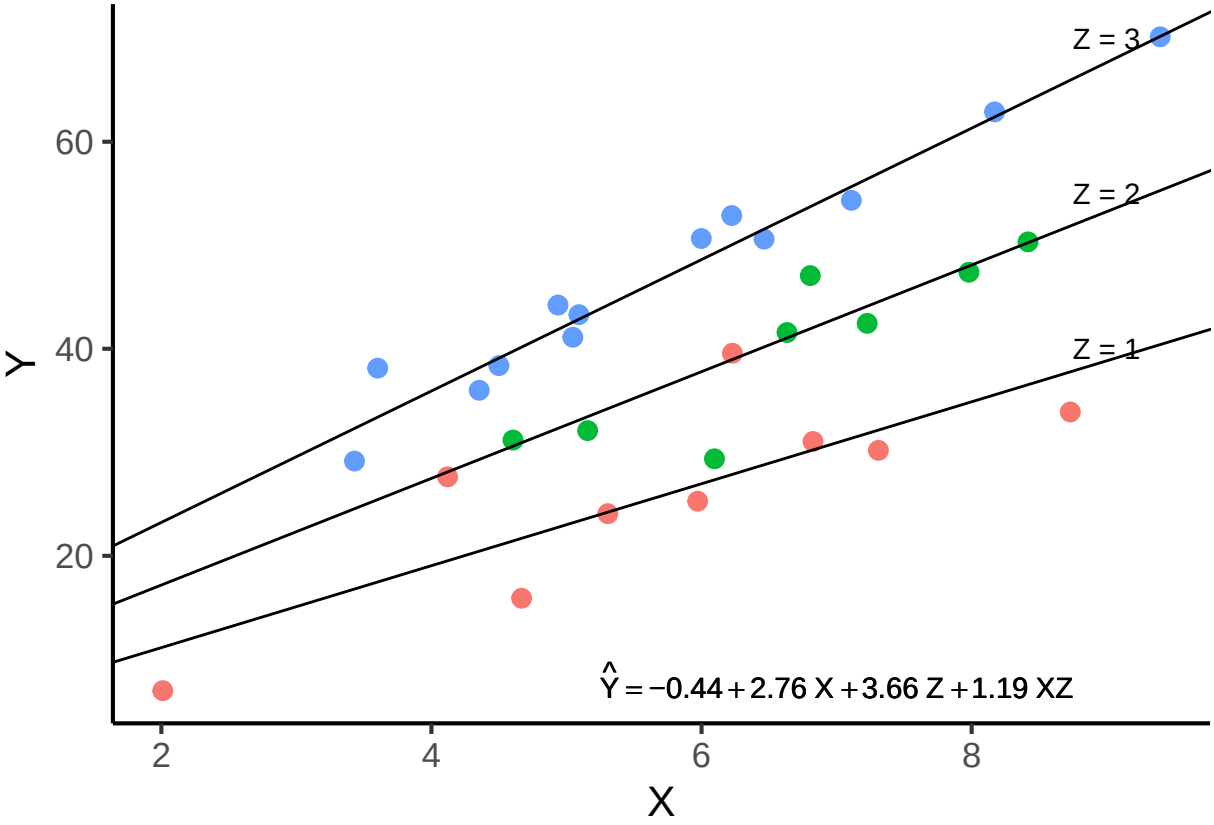
The effect of X on Y depends on the level of Z

- Equivalently, the effect of Z on Y depends on the level of X

Plot of Y versus X (colored by Z)



With simple slopes at the levels of Z



Interpretation

$\hat{Y}_i$ = -0.44 + 2.76X+ 3.66Z+ 1.19XZ

- -0.44 (b_0) is the expected value of Y **when both** X **and** Z **equal 0**
- 2.76 (b_1) is the effect of X on Y (slope) **when** $Z = 0$
- 3.66 (b_2) is the effect of Z on Y (slope) **when** $X = 0$

But both $X = 0$ and $Z = 0$ are outside of the range of data points

- X and Z effects are not meaningful
- We can **center** predictors to make interpretation more meaningful
- We often use the mean of the predictor, but can use other values

Centering at the mean

We often center at the *mean* of the variable (but not always!)

The mean for X is 6.6
Center X by subtracting the mean from each observation

X	minus	Xmean	equals	Xcent
3	–	6.6	=	-3.6
4	–	6.6	=	-2.6
10	–	6.6	=	3.4
10	–	6.6	=	3.4
6	–	6.6	=	-0.6

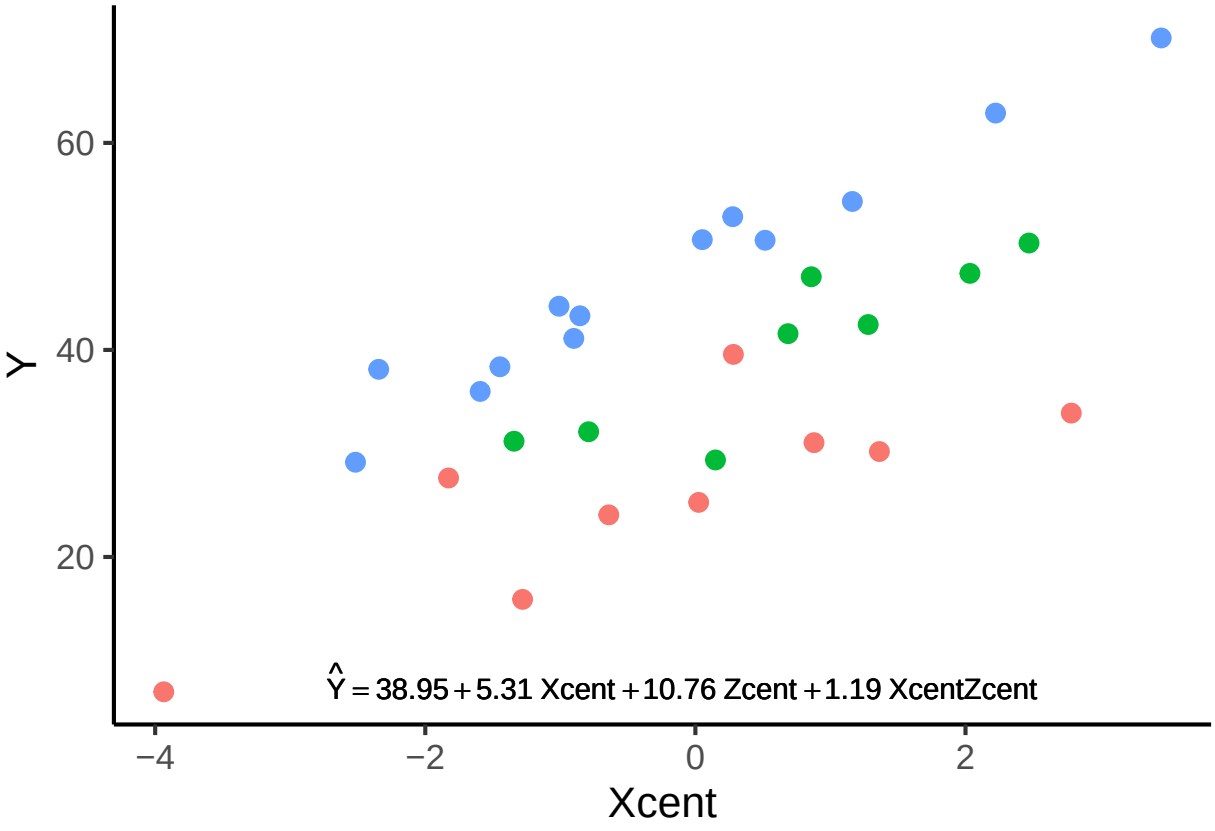
The mean of X is 6.6
The mean of Xcent is 0 (actual calc to 2 decimal places)

Centered interaction

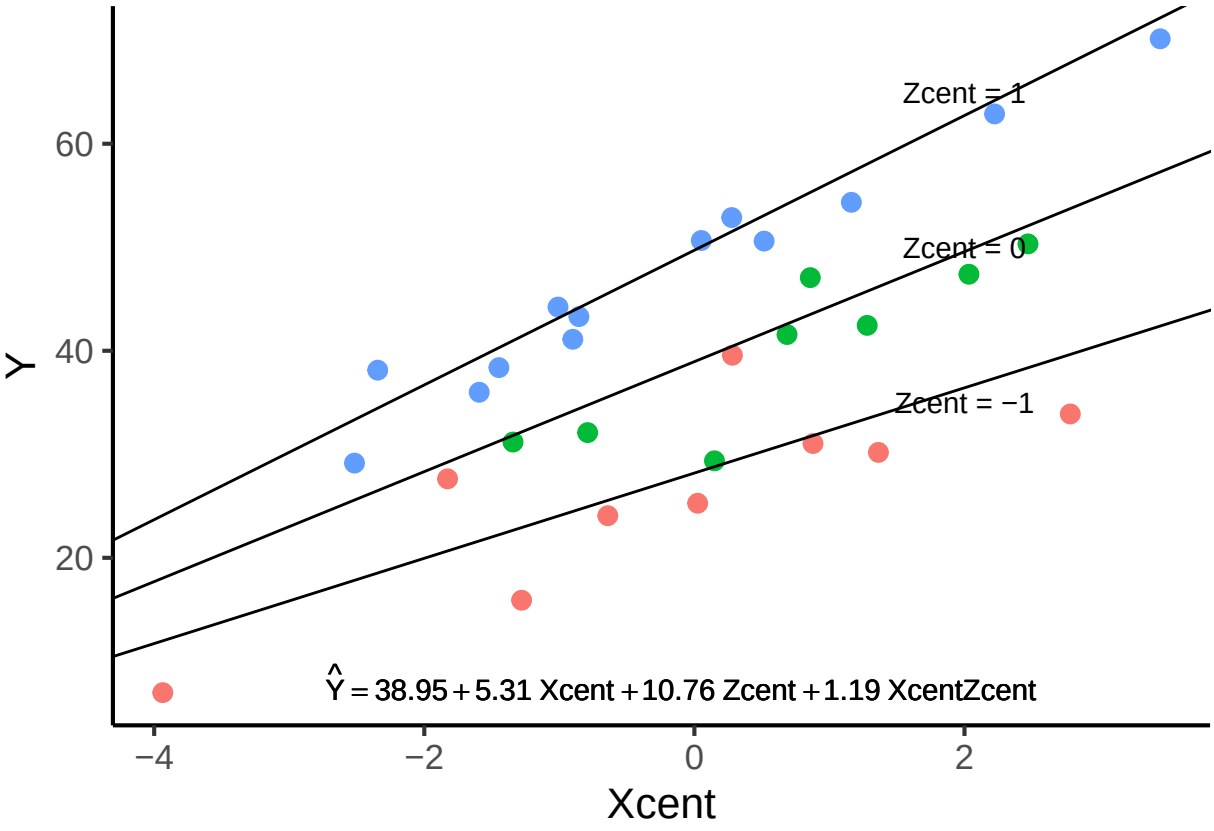
$\hat{Y}_i = b_0 + b_1 Xcent + b_2 Zcent + b_3 XcentZcent$

- Use the new centered predictors, $Xcent$ and $Zcent$
- Note: The interaction term uses the **product** of the centered terms

Plot of Y versus Xcent (colored by Zcent)



With simple slopes at the levels of Zcent



Centered interpretation

$\hat{Y}_i = 38.95 + 5.31 Xcent + 10.76 Zcent + 1.19 Xcent Zcent$

- 38.95 (b_0) is the expected value of Y **when both** $Xcent$ **and** $Zcent$ **equal 0**
  - Expected value of Y **for the middle** Z **group, at the mean of** X
- 5.31 (b_1) is the effect of $Xcent$ on Y (slope) **when** $Zcent = 0$
  - Expected increase in Y for a 1 unit increase in $Xcent$ **for the middle** Z **group**
- 10.76 (b_2) is the effect of $Zcent$ on Y (slope) **when** $Xcent = 0$
  - Expected increase in Y for a 1 unit increase in $Zcent$ **at the mean of** X

Why do we center?

To aid interpretation in any model with an *interaction* or an *intercept*

Centering will be especially important:

- Centering in mixed models to get accurate estimates
- Centering in growth models to get interpretable intercepts
  - Does it make sense to talk about effects at age = 0?

For more information on **centering**, see Aiken and West (1991) and Jaccard, Wan, and Turrisi (1990)

SEM review / crashcourse

SEM overview

For this course

I do NOT expect you to know structural equation modeling

- If you do: awesome!
  - Here is a refresher on some SEM topics relevant to this course
- If you don't: no problem!
  - I will cover the basics of SEM and some terminology I'll use in the next 20 minutes or so

Structural equation modeling

Structural equation modeling (SEM) is a general method of statistical data analysis that allows:

- Complex modeling
- Multiple outcome variables
- Unobserved (latent) variables
- and more

ANOVA, linear regression, path analysis, and factor analysis are all special cases of SEM

What is SEM?

Basically

- SEM lets you simultaneously solve several regression equations
- SEM also allows you to model measurement error

It is also really easy to represent SEMs using a graphical format (i.e., path models), rather than equations

- Matrices and equations underlie SEM, but the graphical format is a lot easier to start with

SEM graphical notation

Observed variables

Observed or manifest variables are actually measured

Observed variables are represented by **squares**

- In a study, we ask people how old they are (in years)
- "Age" is an observed variable; in a model, it would look like:

Latent variables

Latent or unobserved variables are not actually measured

Latent variables are represented by **circles**

- In a study, we are interested in intelligence, but we don't actually measure intelligence directly
- "Intelligence" is a latent variable; in a model, it would look like:

Means / intercepts

Means and intercepts are represented as a triangle with the number 1 in it

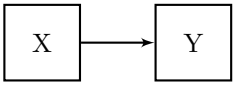
They are often omitted in figures unless they are specifically relevant to the model

They are important in certain models, so keep them in mind

Prediction

Prediction is represented by an **arrow** from the predictor variable to the outcome variable

X predicts Y looks like:

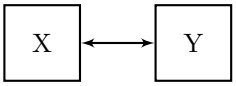


Correlation

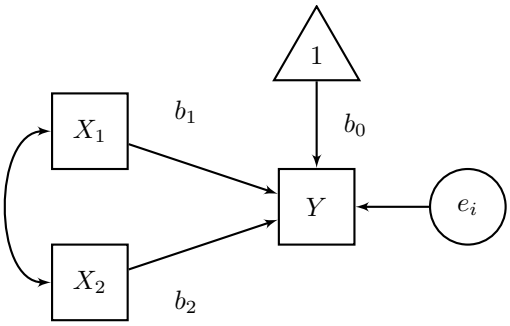
Correlation is represented by a **double-headed arrow** between the correlated variables

These arrows are often *curved* but they don't have to be

X is correlated with Y looks like:



Y predicted by X1 and X2



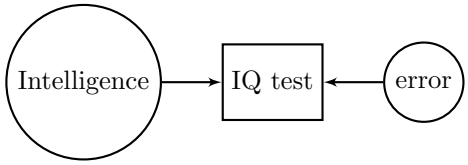
Applications

Measurement error

Measurement error reflects the discrepency between a measured score and unobserved "true score"

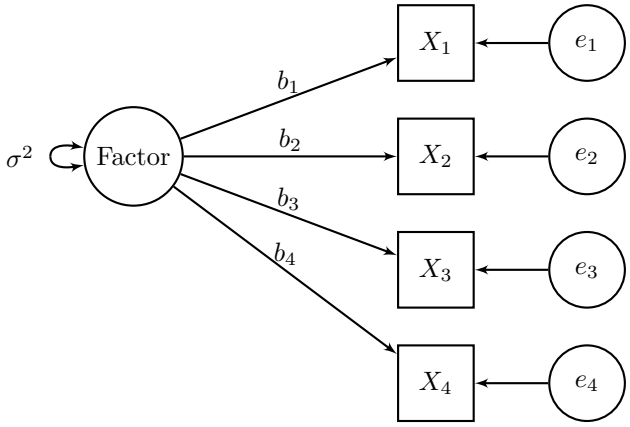
We don't directly measure intelligence, but we can try to measure it indirectly, with an IQ test

But there will be some **error** in our measurement and we can include that error in the model



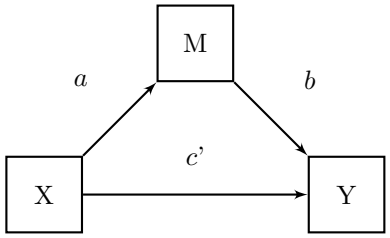
Factor analysis

Discover a single underlying (latent) factor that explains the variance in a set of observed variables



Mediation

Statistical mediation occurs when three variables occur in a causal chain, such that X causes M, which then causes Y



Regression with latent variables