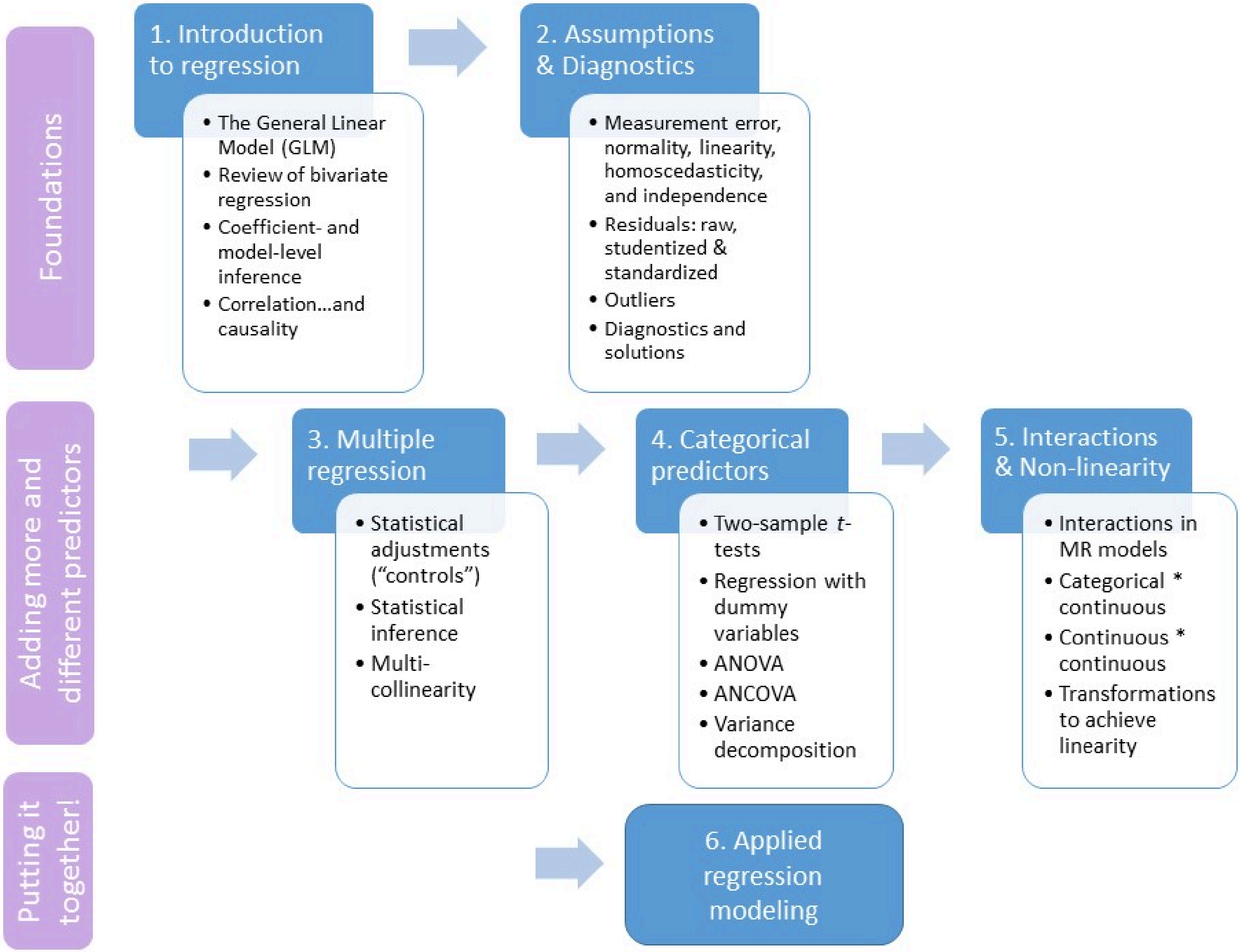
# Introduction and bivariate relationships

EDUC 643: Unit 1

David D. Liebowitz

UNIVERSITY OF OREGON

# Roadmap



Foundations

1. Introduction to regression

- The General Linear Model (GLM)
- Review of bivariate regression
- Coefficient- and model-level inference
- Correlation…and causality

2. Assumptions & Diagnostics

- Measurement error, normality, linearity, homoscedasticity, and independence
- Residuals: raw, studentized & standardized
- Outliers
- Diagnostics and solutions

Adding more and different predictors

3. Multiple regression

- Statistical adjustments ("controls")
- Statistical inference
- Multi-collinearity

4. Categorical predictors

- Two-sample *t*-tests
- Regression with dummy variables
- ANOVA
- ANCOVA
- Variance decomposition

5. Interactions & Non-linearity

- Interactions in MR models
- Categorical * continuous
- Continuous * continuous
- Transformations to achieve linearity

Putting it together!

6. Applied regression modeling

# Goals of EDUC quant sequence

- Develop the basic quantitative skills necessary to conduct applied data analysis
  - The full-year sequence (EDUC 641 → EDUC 643 → EDUC 645) prepares you to make valuable contributions to a research team
  - Not all the skills you will need (and not the only courses you should take), particularly for those interested in analysis-heavy positions, or becoming an applied quantitative doctoral-level researcher
  - Foundations of statistics, methods and data science
- Understand the (in)appropriate application of those skills
  - "Building a toolbox, not a cookbook"
  - Evaluate the credibility of published research
  - Understand the affordances, limitations and dangers of quantitative analysis

This is a re-designed and modernized core quantitative sequence at the COE. We welcome your feedback!

Given the relative new-ness of these courses, there are likely to be hiccups. We are committed to solving them, but we ask for your grace in allowing us to do so.

# EDUC pedagogical orientation

- Analysis follows research design that emerges from substantive questions
- Students learn statistical analysis by doing statistical analysis
- Create an inclusive, supportive environment in which we learn from each other
- Balance support and academic stretch across a variety of levels of prior experience and comfort with quantitative analysis
- *We started with an assumption of no prior background in mathematics, programming, statistics or research*

We assume that you enter this course (EDUC 643) with the ability to:

- Create an RStudio project, read in a dataset in csv format, filter rows, select columns, understand structure of data, recode variables, assess missingness, and calculate summary statistics for continuous and categorical variables in the R programming language
- Create simple visualizations describing categorical and continuous variables, and their relationships with each other in R
- State a quantitative research question and its corresponding null hypothesis
- Articulate the conceptual basis for a null-hypothesis significance test
- Conduct a one-sample $t$-test and fit a bivariate Ordinary Least Squares regression
- See EDUC 641 data management guide for data management pointers

# Goals of EDUC 643

In this intermediate course, we will focus on applying the General Linear Model to Ordinary Least Squares regression analysis. Students will progress from bivariate to multiple regression, developing an understanding of the associated assumptions of these models and tools to solve instances in which those assumptions are unmet. The course seeks to blend a conceptual, mathematical and applied understanding of basic statistical concepts.

Concrete learning objectives:

1. Articulate the framework of the General Linear Model as a method to describe relationships between quantitative variables
2. Distinguish between research designs and analyses that permit different forms of inferences (e.g., relational or causal, inferential or descriptive)
3. Conduct and interpret (orally and in writing) least-squares regression analyses with continuous outcomes and predictors
4. Describe the assumptions of least-squares regression analysis and test analytic models for the extent to which they satisfy these assumptions
5. Generalize the least-squares regression model (conceptually and in practice) to predictors that are categorical, interacted and non-linear
6. Build taxonomies of sensible regression models in response to independently developed research questions
7. Use an open-source, object-oriented statistical programming language (specifically R) to conduct all such analyses

# My personal goals

- Teach you the basics of applied statistics
- Challenge you:
  - Everyone did very well in EDUC 641; this course will introduce more challenging material
  - If you're struggling, come talk to me
  - If the course is too basic, consider either the recommended texts for more formal mathematical treatment **AND/OR** programming extensions (see the resources on the syllabus **AND/OR** reflect on pedagogical approaches as a future instructor
- Induct you into a graduate-level focus on mastery, rather than an undergraduate-level focus on success
- Contribute to building a community of live learners and life-long professional colleagues
- Assign everyone an A
- Help you see that many common statistical approaches with fancy names (multivariate regression, ANOVA, Spearman rank correlation coefficient, MANCOVA, Welch's two-sample $t$-test, Weighted Least Squares, $\chi^2$ goodness-of-fit test, etc.) are part of the same same "family" of statistical tests known as the General Linear Model (GLM). (aka *demystify the world of statistics*)

# Me

- David Liebowitz (aka, "David" or "Dr. Liebowitz")
- he/him
- English lit undergrad major
- Former ELA teacher and principal
- Policy analyst/advisor
- **Applied** quantitative social scientist focused on improving leadership and policy in schools
- **Not** a methodologist

# Learnings from the fall

- Maintain inclusivity, support, clarity, communuication, relevance, organization and accessibility
- Improve active learning (also more "chew time") and accessibility
- Clarity in use of technical language (and its nuances)
- Review topics on which students struggled in assignments; make more explicit what about an assignment response does not meet expectations
- 81% response rate on SES 😍 [everyone gets a free percentage point!!]

# Roadmap of EDUC 643

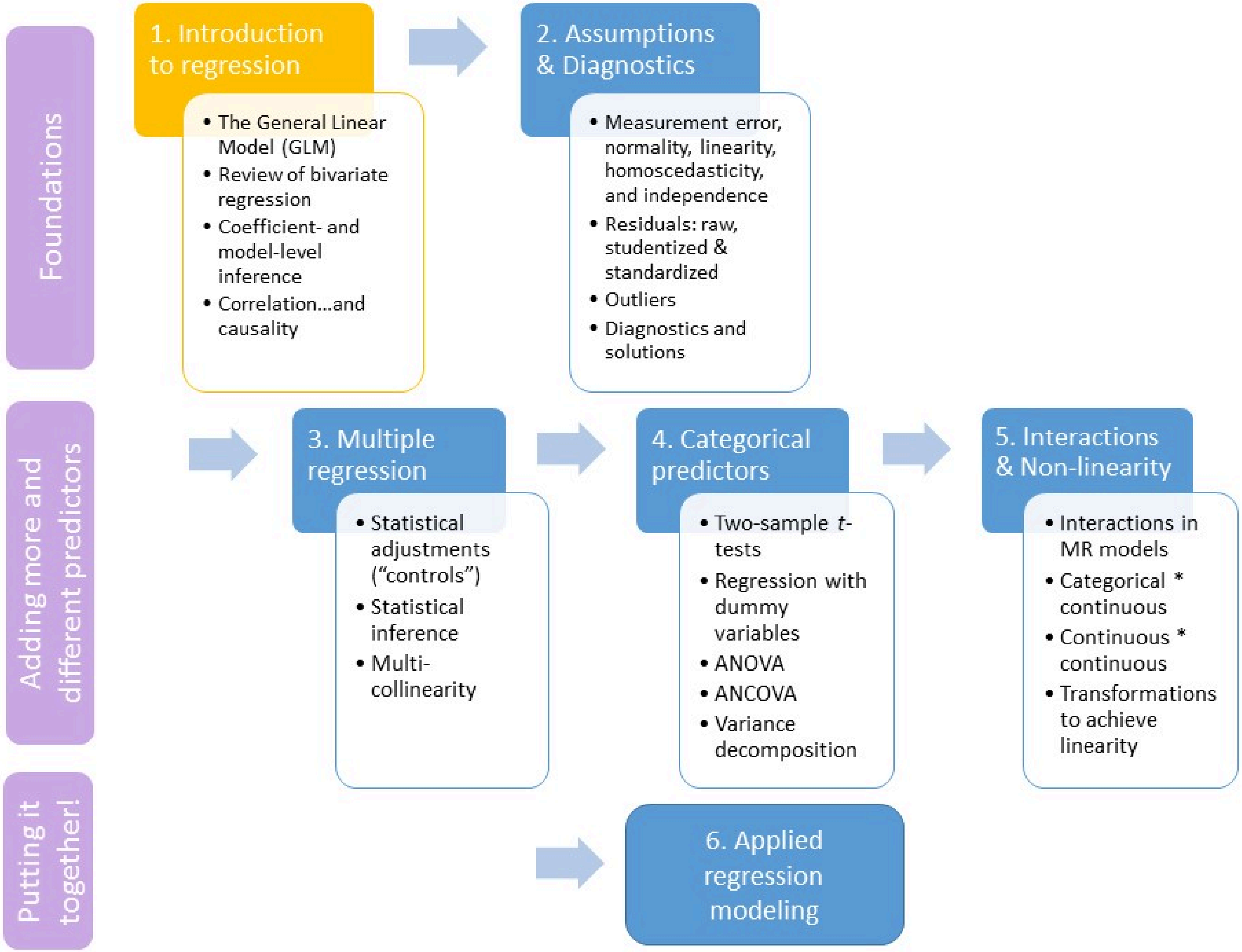
| Unit | Week(s) | Topic |
|---|---|---|
| 1 | 1–2 | Continuous relationships redux |
| 2 | 3 | Regression assumptions |
| 3 | 4 | Multiple regression |
| 4 | 5–6 | Categorical predictors |
| 5 | 7–9 | Non-linearities and interactions |
| 6 | 10 | Model building and applied analysis |

# Weekly schedule of activities

1. Two 1.5-hour lectures will introduce concepts in interactive lectures, discussion and activities
2. Readings are intended to supplement material from lectures (these can be completed after the first class of the unit)
3. Two *OPTIONAL, UNGRADED* weekly lab meetings intended to provide support for R programming tasks. Each lab will cover identical material.
4. Five (5) take-home quizzes worth trivial amount of points each (2% of grade each)
5. Four (4) data analytic assignments (1st: 20 points; 2nd-4th: 13 points each) + final project (30% of grade)

Course website is (we hope!) a valuable resource. Let's check it out!

# Roadmap



Foundations

1. Introduction to regression

- The General Linear Model (GLM)
- Review of bivariate regression
- Coefficient- and model-level inference
- Correlation…and causality

2. Assumptions & Diagnostics

- Measurement error, normality, linearity, homoscedasticity, and independence
- Residuals: raw, studentized & standardized
- Outliers
- Diagnostics and solutions

Adding more and different predictors

3. Multiple regression

- Statistical adjustments (“controls”)
- Statistical inference
- Multi-collinearity

4. Categorical predictors

- Two-sample *t*-tests
- Regression with dummy variables
- ANOVA
- ANCOVA
- Variance decomposition

5. Interactions & Non-linearity

- Interactions in MR models
- Categorical * continuous
- Continuous * continuous
- Transformations to achieve linearity

Putting it together!

6. Applied regression modeling

# Bivariate regression

# Goals for the unit

- Describe goals and structure and of the course (and the quantitative EDUC sequence more broadly)
- Characterize a bivariate relationship along five dimensions (direction, linearity, outliers, strength and magnitude)

- Describe how statistical models differ from deterministic models
- Mathematically represent the population model and interpret its deterministic and stochastic components
- Formulate a linear regression model to hypothesize a population relationship
- Describe residuals and how they can describe the degree of our OLS model fit
- Explain $R^2$, both in terms of what it tells us and what it does not
- Estimated a fitted regression line using Ordinary-Least Squares regression
- Conduct an inference test for a regression coefficient and our regression model
- Calculate a correlation coefficient $(r)$ and describe its relationship to $R^2$
- Distinguish between research designs that permit correlational associations and those that permit causal inferences

# A motivating question

Researchers (including two from the **University of Oregon**), Nichole Kelly, Elizabeth Cotter and Claire Guidinger (2018), set out to understand the extent to which young men who exhibit overeating behaviors have weight-related medical and psychological challenges.

Using real-world data (generously provided by Nichole Kelly) about the dietary habits, health, and self-appraisals of males 18-30, we are going to attempt to answer a similar question.

In particular, we are going to explore the **relationship** between **dietary restraint behaviors** (self-reports on the extent to which participants consciously restricted/controlled their food intake) and **over-eating frequency** (participants' self-reported frequency of over-eating episodes).

# Quantitative research design components

1. **Research questions**
   - Descriptive
   - Relational
   - Causal
2. **Question predictors**, *sometimes called independent variables (IV)*
   - Fixed attributes (e.g., race, age)
   - Potentially changeable characteristics (e.g., trauma, class size)
   - Interventions (e.g., new curriculum, counseling, programs/policies)
3. **Outcomes**, *sometimes called dependent variables (DV)*
4. **Analytic strategy**

**What do you anticipate each of these will be here?**

# A glance at the data

```
# Not all data comes to us in .csv format
# Here we'll use the `haven` package to read in SPSS data
# There are many other such packages, including: `foreign` and `rio`

# install.packages("haven")
do <- read_spss(here("data/male_do_eating.sav")) %>%
          select(OE_frequency, EDEQ_restraint, EDS_total,
          BMI, age_year, income_group)

head(do)
```

```
#> # A tibble: 6 x 6
#>   OE_frequency EDEQ_restraint EDS_total   BMI age_year income_group
#>          <dbl>          <dbl>     <dbl> <dbl>    <dbl> <dbl+lbl>
#> 1            0            0          21  23.3       27 1 [less than 19,999]
#> 2            0            1.2        21  16.6       30 1 [less than 19,999]
#> 3            0            0.4        22  18.7       21 1 [less than 19,999]
#> 4            0            0.4        22  15.4       21 6 [60-69,999]
#> 5            0            0          40  20.2       29 1 [less than 19,999]
#> 6            0            1.2        34  26.3       27 2 [20-29,999]
```

# More glance-ing

```r
str(do)
```

```
#> tibble [1,114 x 6] (S3: tbl_df/tbl/data.frame)
#>  $ OE_frequency  : num [1:1114] 0 0 0 0 0 0 0 0 0 0 ...
#>   ..- attr(*, "label")= chr "Frequency of overeating episodes"
#>   ..- attr(*, "format.spss")= chr "F8.2"
#>  $ EDEQ_restraint: num [1:1114] 0 1.2 0.4 0.4 0 1.2 0.2 1.6 2.2 0 ...
#>   ..- attr(*, "label")= chr "Mean dietary restraint score (0-6, higher score
#>   ..- attr(*, "format.spss")= chr "F8.2"
#>   ..- attr(*, "display_width")= int 16
#>  $ EDS_total     : num [1:1114] 21 21 22 22 40 34 21 49 46 24 ...
#>   ..- attr(*, "label")= chr "range is 21-126; all scores added up; higher sc
#>   ..- attr(*, "format.spss")= chr "F8.2"
#>   ..- attr(*, "display_width")= int 11
#>  $ BMI           : num [1:1114] 23.3 16.6 18.7 15.4 20.2 ...
#>   ..- attr(*, "label")= chr "Body mass index in kg/m-squared"
#>   ..- attr(*, "format.spss")= chr "F8.2"
#>   ..- attr(*, "display_width")= int 10
#>  $ age_year      : num [1:1114] 27 30 21 21 29 27 22 28 24 19 ...
#>   ..- attr(*, "label")= chr "Total years of age"
#>   ..- attr(*, "format.spss")= chr "F8.2"
#>   ..- attr(*, "display_width")= int 10
#>  $ income_group  : dbl+lbl [1:1114] 1, 1, 1, 6, 1, 2, 6, 6, 8, 8
```

# Check for missingness

```
# Check for missingness
sapply(do, function(x) sum(is.na(x)))
```

```
#>   OE_frequency EDEQ_restraint      EDS_total            BMI       age_year
#>              0              5              0             15              0
#>   income_group
#>              3
```

```
# Things look good, but we'll focus on those cases that had complete
# records for all variables
do <- do %>% drop_na()
```

What is this approach to missingness called? Listwise deletion

# Even more glance-ing

Let's start by focusing on a few variables of interest: dietary restraint (***EDEQ_restraint***), over-eating frequency (***OE_frequency***), and exercise-dependency (***EDS_total***).

```
summary(do$EDEQ_restraint)
```

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   0.000   0.200   1.200   1.386   2.200   6.000
```

```
summary(do$OE_frequency)
```

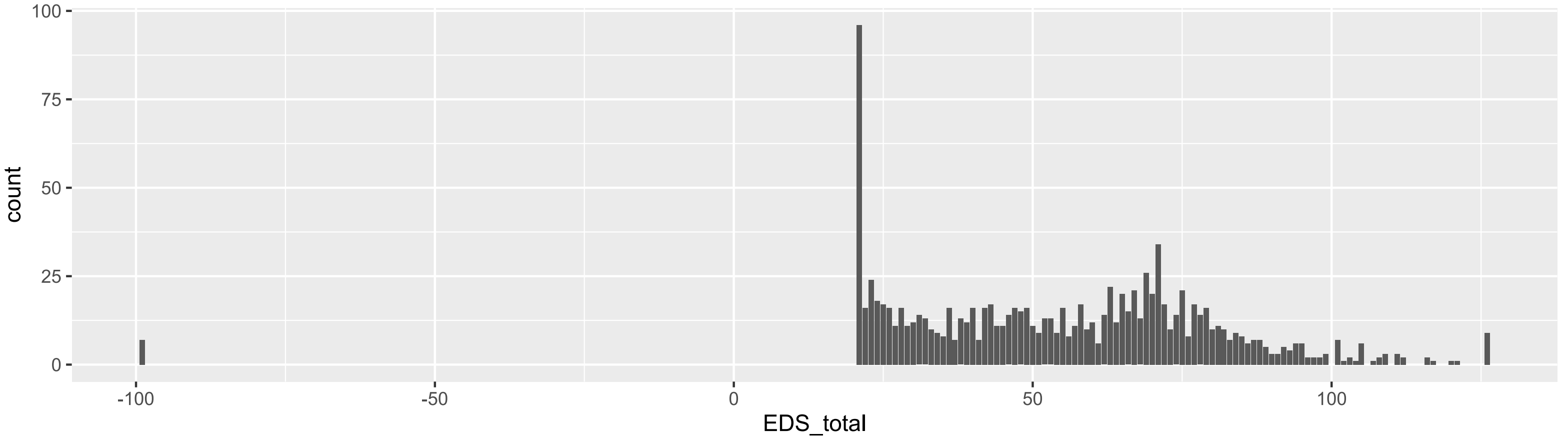
```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   0.000   0.000   1.000   2.285   4.000  29.000
```

```
summary(do$EDS_total)
```

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>  -99.00   33.00   55.00   54.03   72.00  126.00
```

# Oh the tangled web SPSS weaves!

```
ggplot(data=do, aes(EDS_total)) + geom_bar()
```



```
do <- do %>% mutate(EDS_total = ifelse(EDS_total==-99, NA, EDS_total))

sum(is.na(do$EDS_total))
```

```
#> [1] 7
```

```
do <- filter(do, !is.na(EDS_total))
```

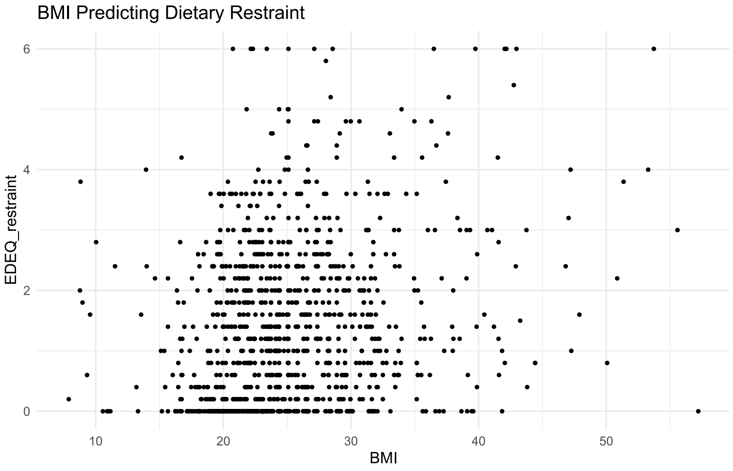
# A preliminary analysis

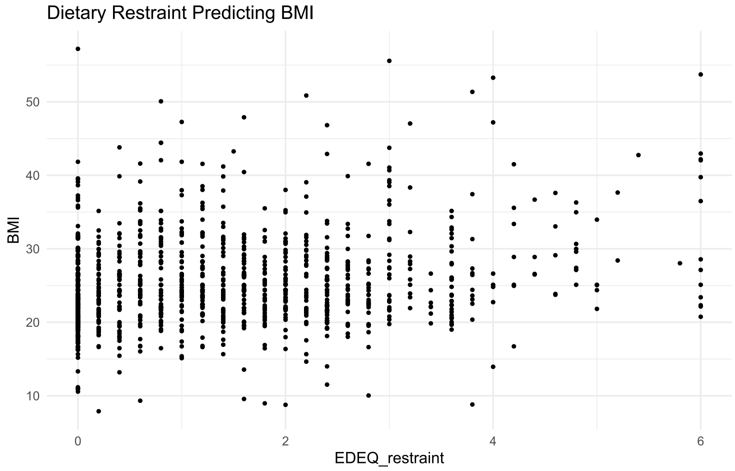
Before we get to the core question of the Kelly et al. study--how are dietary restraint behaviors related to over-eating frequency?--we are going to explore another important relationship in the data that may also be related to our main research question: the relationship between dietary restraint behaviors (self-reports on the extent to which participants consciously restricted/controlled their food intake) and body-mass index (BMI). In particular, we are going to operationalize this by examining the relationship in our sample of young men between our predictor variable (***EDEQ_restraint***) and their body-mass index (***BMI***).

> We are examining this relationship so that we can better understand how all three of these variables (***OE_frequency***, ***EDEQ_restraint*** and ***BMI***) are related in Unit 3. Additionally, the properties of the variable ***BMI*** in this particular dataset are pedagogically helpful in demonstrating the assumptions of OLS.
>
> However, we recognize that BMI has been shown to be relatively uninformative about individuals' overall health and categorizes individuals based on distributions initially derived exclusively from white Western European (French and Scottish) study participants. We use the measure for pedagogical purposes because the variable is one of the few continuous measures in one of the few datasets that our UO colleagues shared with us, while noting its problematic historical use, particularly at the individual level.

# Bivariate relationships

We are interested in the *relationship* between dietary restraint (*EDEQ_restraint*) and body-mass index (*BMI*). Statistically, the relationship is the same regardless of which variable is the "outcome."



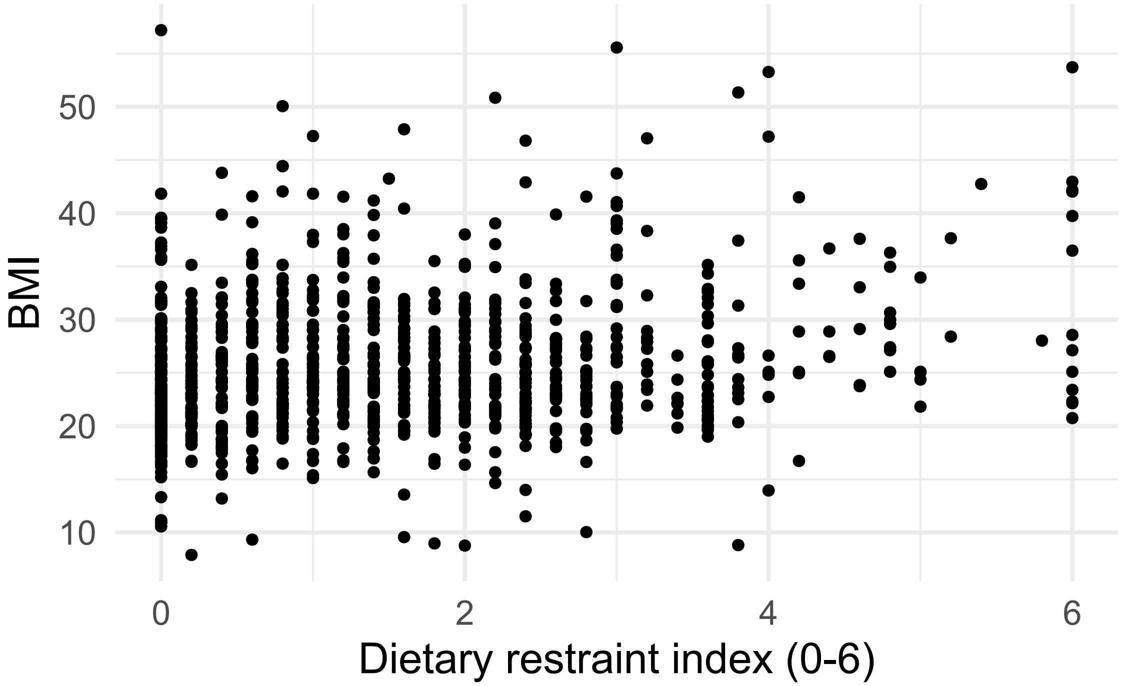


However, our convention is to consider the variable on the Y-axis to be the one that we interpret as the "outcome" or the "dependent" variable.

# Bivariate relationships

Given the focus of the Kelly et al. paper and for pedagogical reasons, for the moment, we are choosing to plot BMI against dietary restraint.

```
biv <- ggplot(data=do, aes(x=EDEQ_restraint, y=BMI)) +
          geom_point()
```

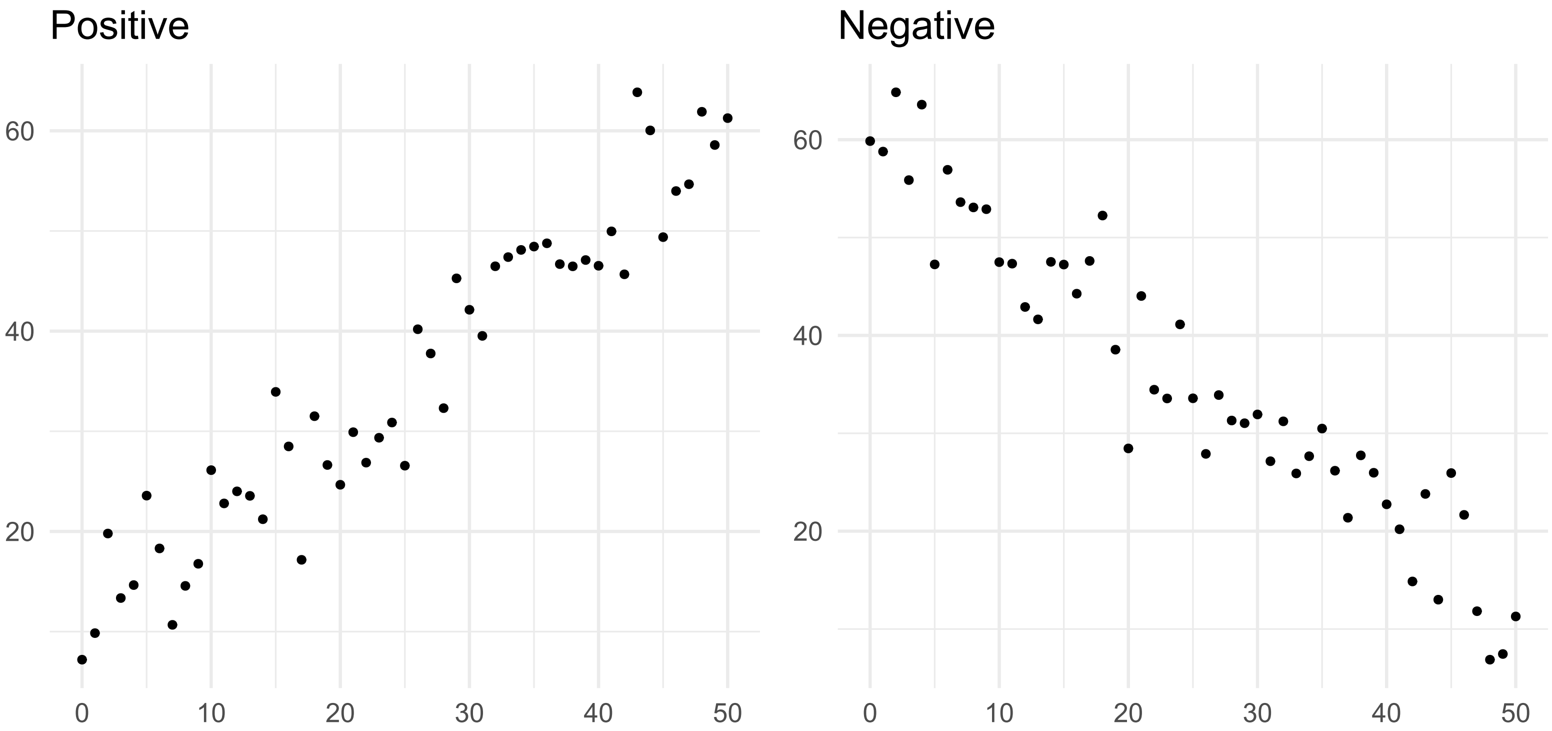


# Bivariate relationships

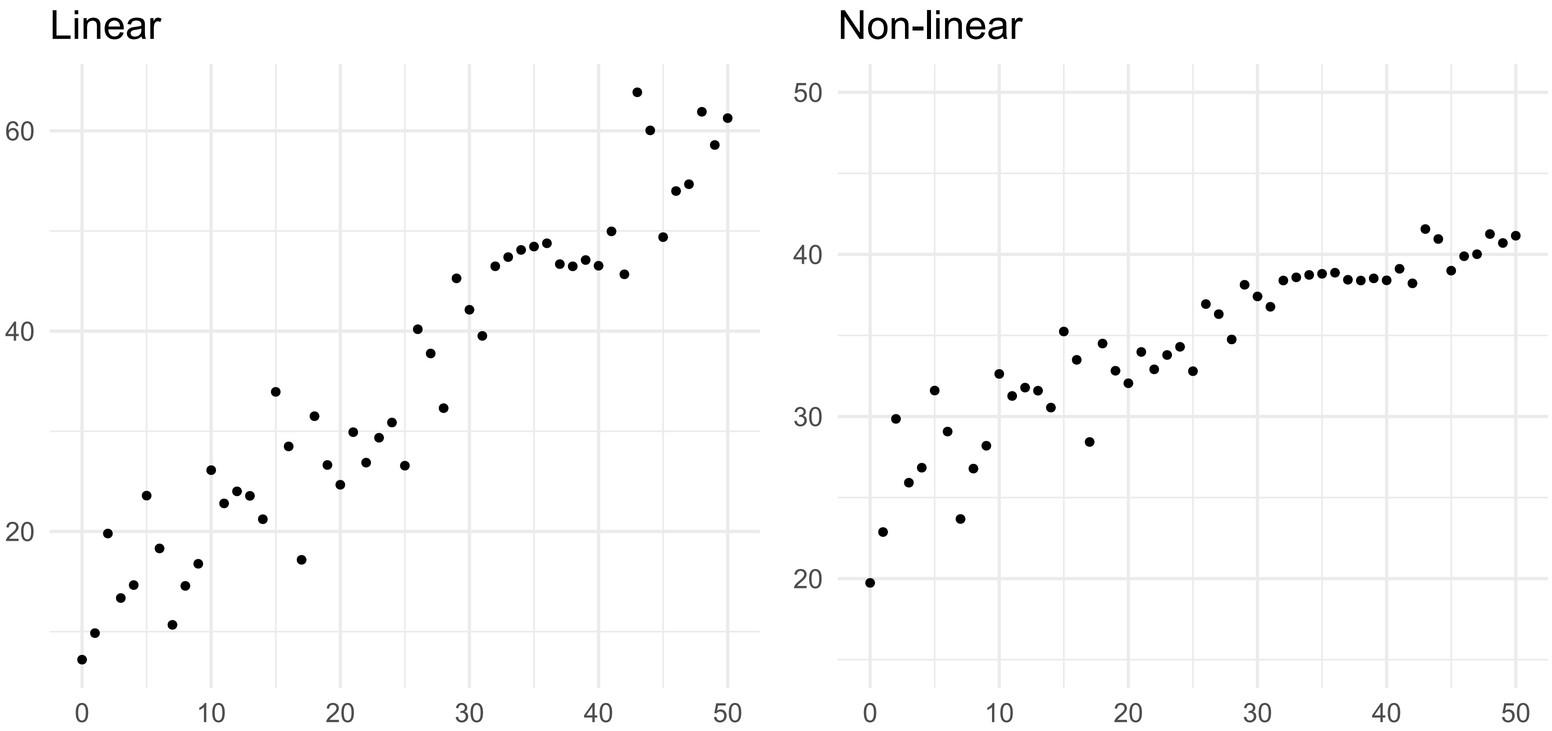
**Consider these five features of bivariate relationships:**

- Direction
- Linearity
- Outliers
- Strength
- Magnitude

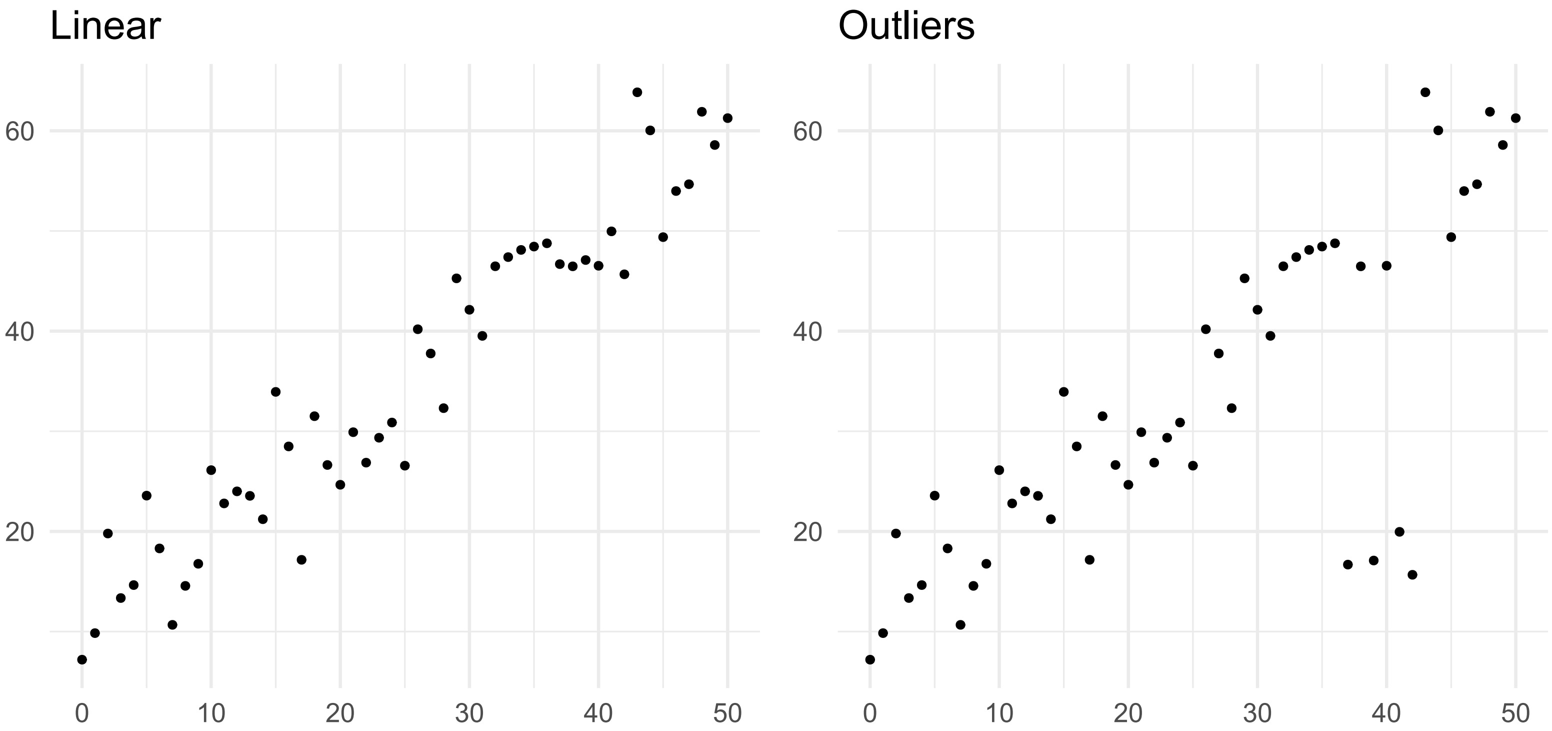
# Bivariate relationships: Direction



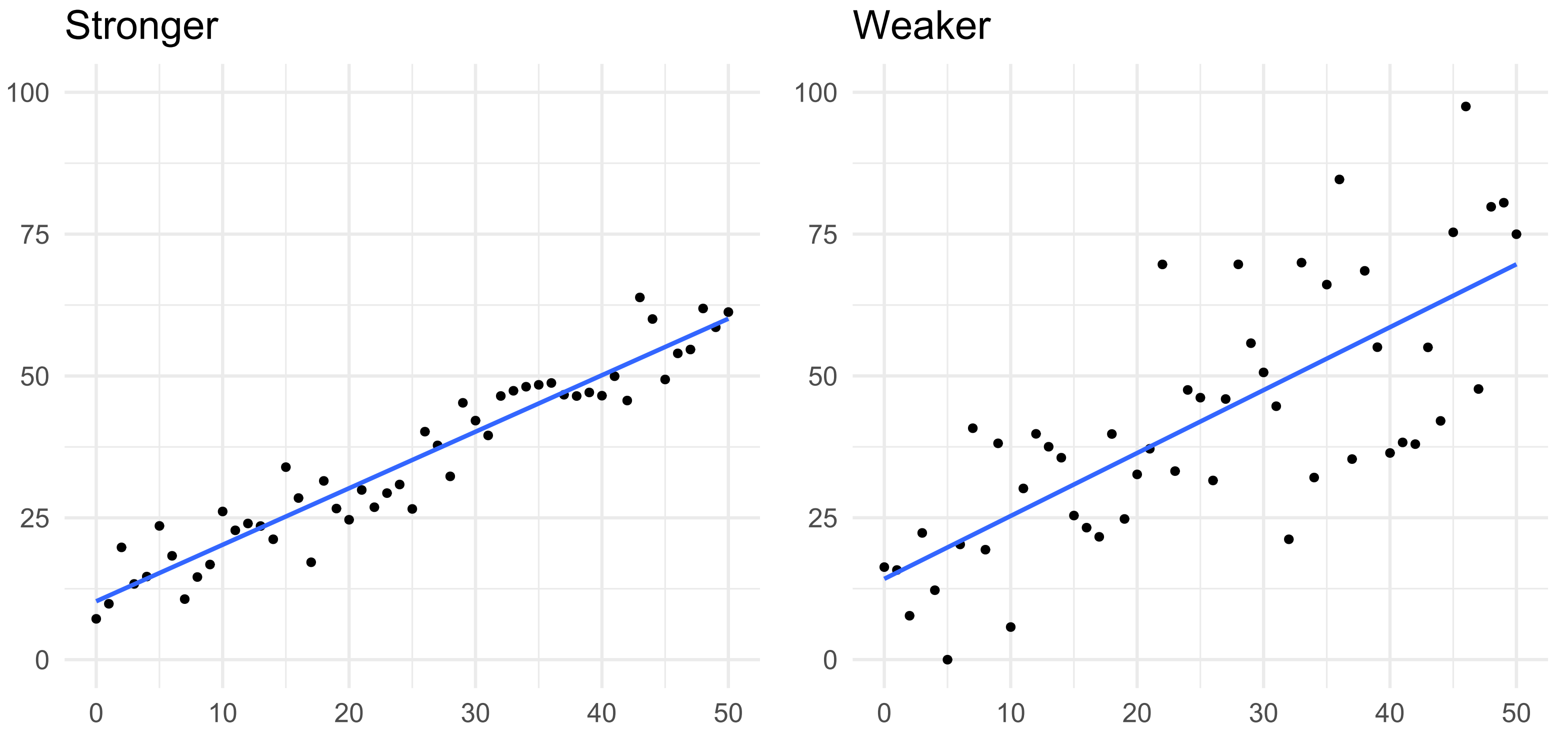
# Bivariate relationships: Linearity



# Bivariate relationships: Outliers

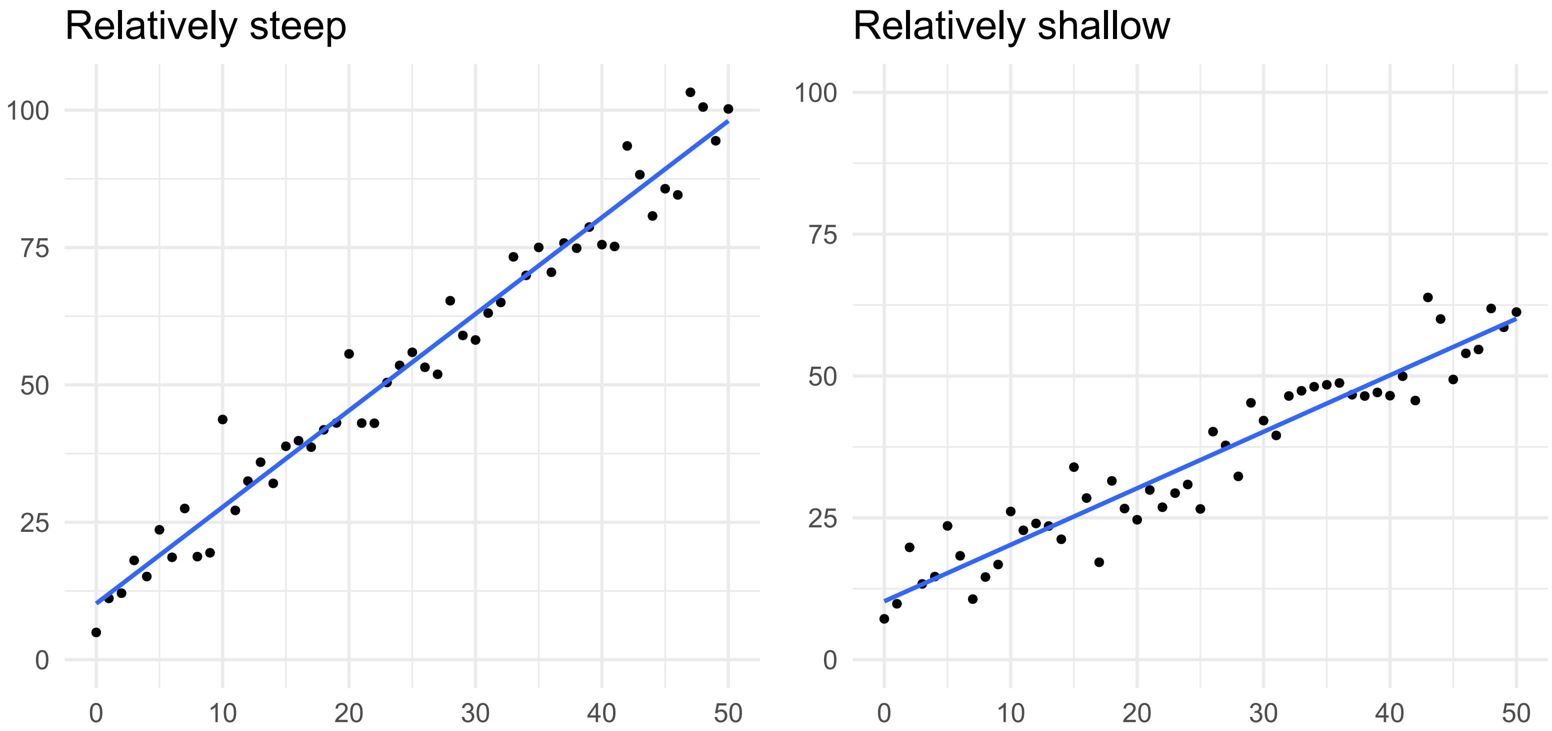


# Bivariate relationships: Strength



Note that these have the same slope (magnitude)

# Bivariate relationships: Magnitude



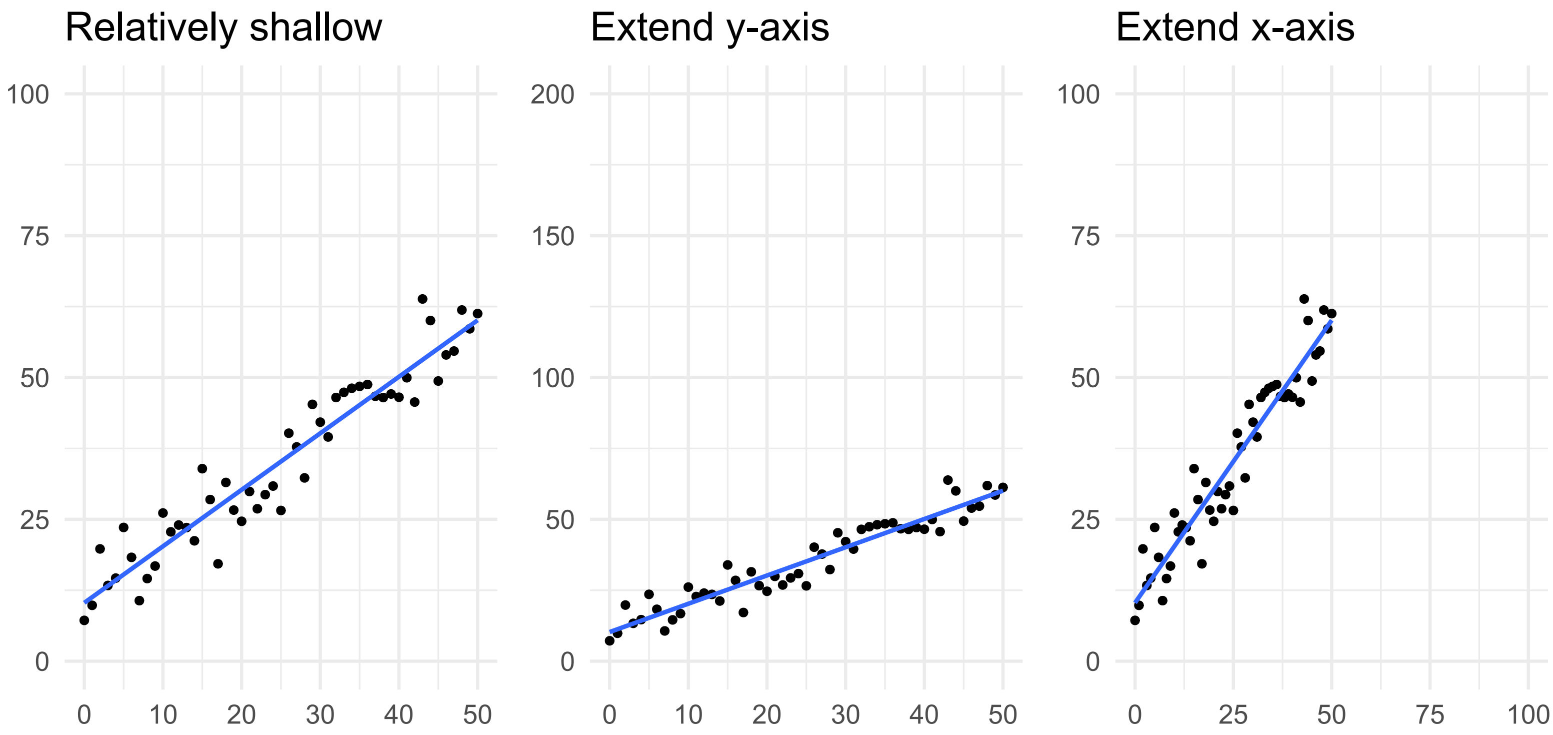
Relatively steep

Relatively shallow

Note that these have the same strength (tightness of fit)

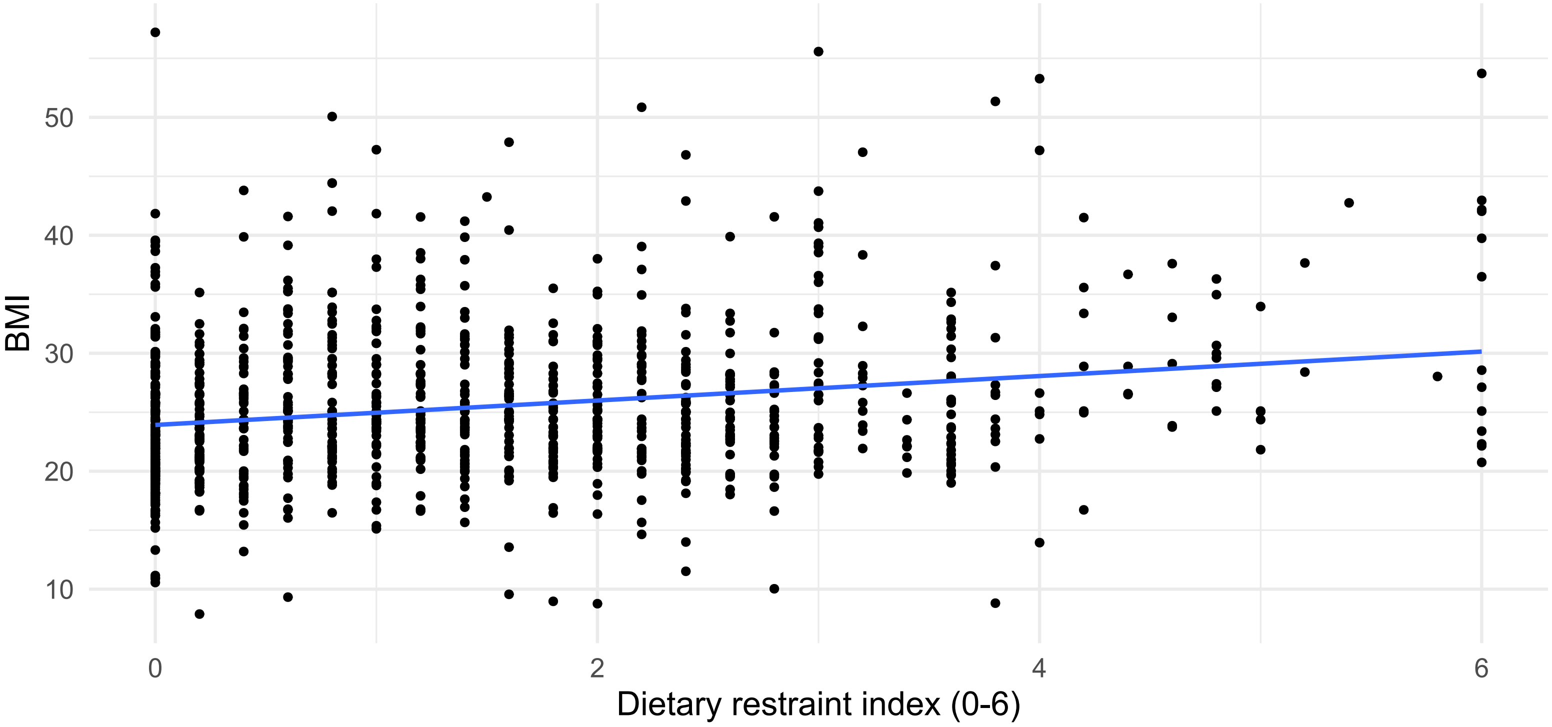
# Bivariate relationships: Magnitude

Note that if you change the scale of either axis, it will change the perceived--**but not the *actual***--size of the magnitude



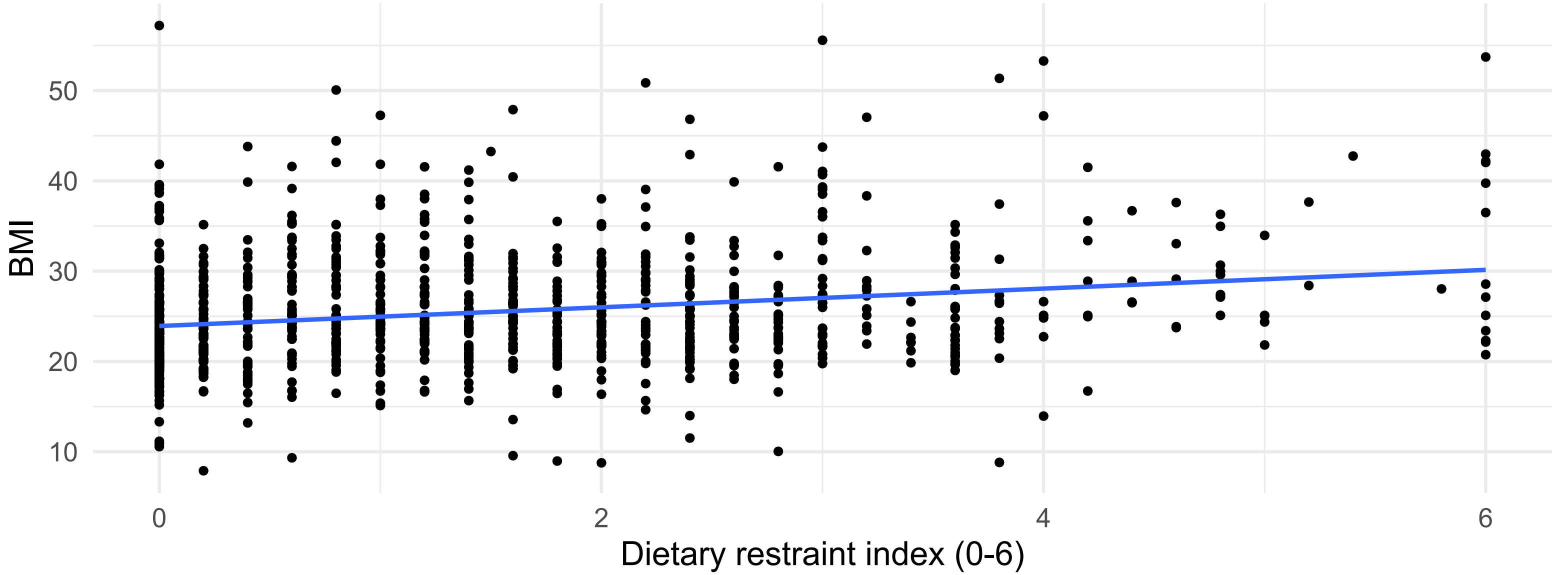
# BMI and over-eating

So what can we say with respect to these five characteristics in our data?



# A line through our cloud

Notice that in the previous slide, we added a line running through our data



That line is defined by the intercept (value $Y$ takes when $X = 0$) and the slope (the difference in $Y$ per 1 unit difference in $X$)

> $Y = intercept + slope * X$ (you may have seen this in HS as Y = mX + b)
>
> We could think of this relationship, therefore, as $BMI = slope * EDEQ\_restraint + intercept$ ... In fact, that's how we described this in EDUC 641, but that's not quite right...

# Synthesis and wrap-up

# Class goals

- Describe goals and structure and of the course (and the quantitative EDUC sequence more broadly)
- Characterize a bivariate relationship along five dimensions (direction, linearity, outliers, strength and magnitude)

# To-Dos

## Reading:

- **By January 16 class**: LSWR Chapter 15.1 - 15.2 and 15.4 - 15.7 and Hu (2021)

## Review:

- Review EDUC 641, Unit 4 (Lectures 13 - 16)