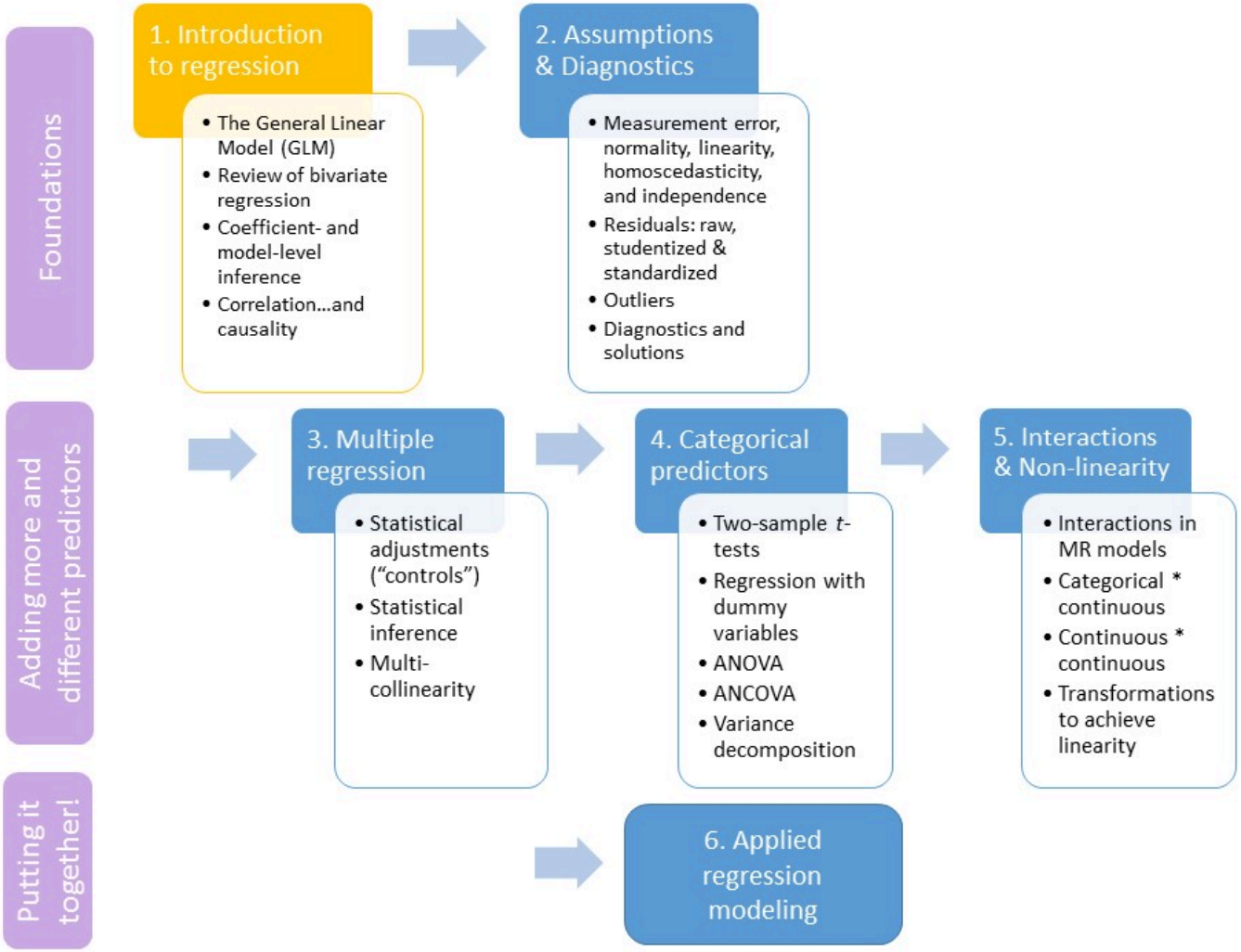
General Linear Model (GLM)

EDUC 643: Unit 1

David D. Liebowitz

Roadmap


Foundations
Adding more and different predictors
Putting it together!
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
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
6. Applied regression modeling

Goals for the unit

- 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

The General Linear Model

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).

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*** are pedadogically 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.

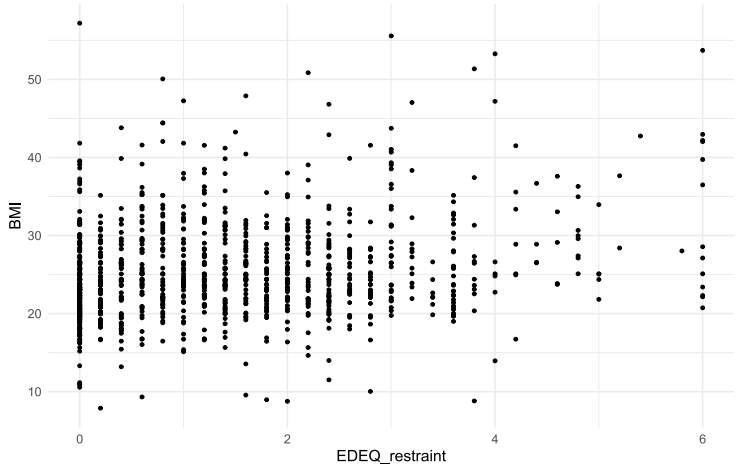
Reading in the data

```
do <- read_spss(here("data/male_do_eating.sav")) %>%
    select(OE_frequency, EDEQ_restraint, EDS_total,
           BMI, age_year, income_group) %>%
    mutate(EDS_total = ifelse(EDS_total==-99, NA, EDS_total)) %>%
    drop_na()
```

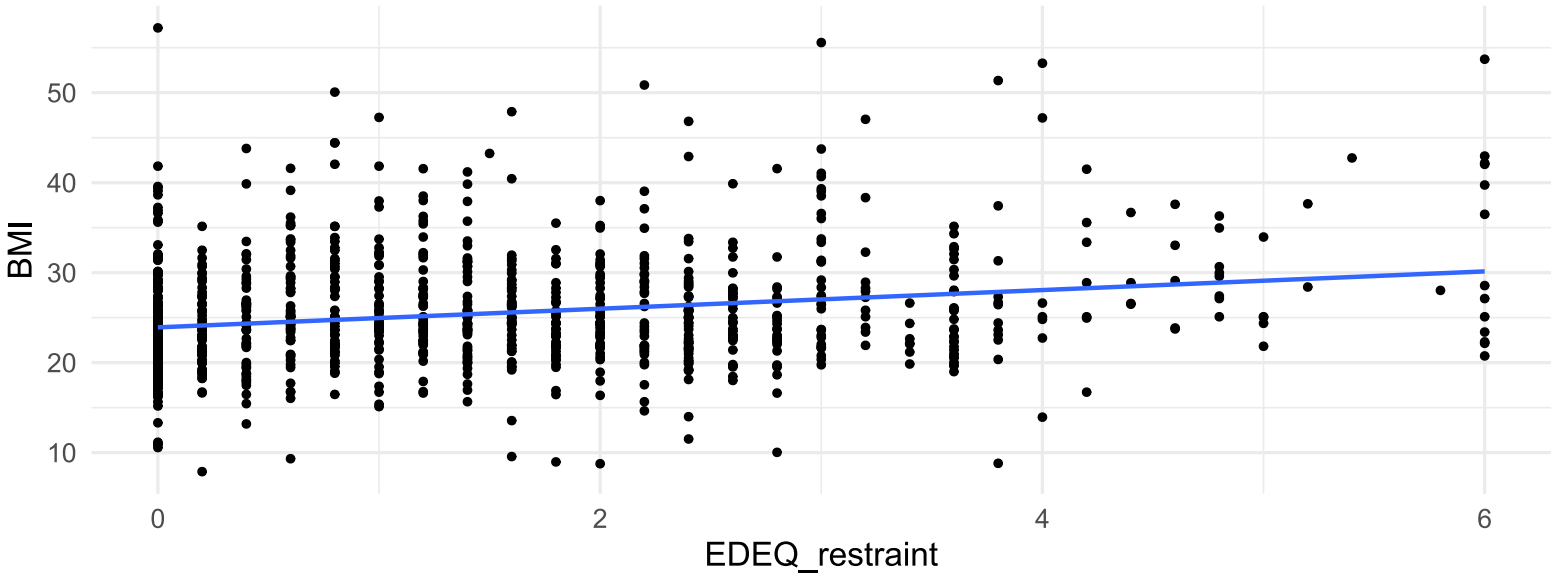
Bivariate relationships

Five ways to characterize them

- Direction
- Linearity
- Outliers
- Strength
- Magnitude



A line through our cloud



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...

Mathematical representations

In addition to visual representations, we can borrow from mathematical models to construct a statistical relationship between variables. However, these are not identical.

Mathematical models

- Are deterministic
- The area of any square is always s^2
- Once we know the rule, we can use it to fit the model to empirical data perfectly

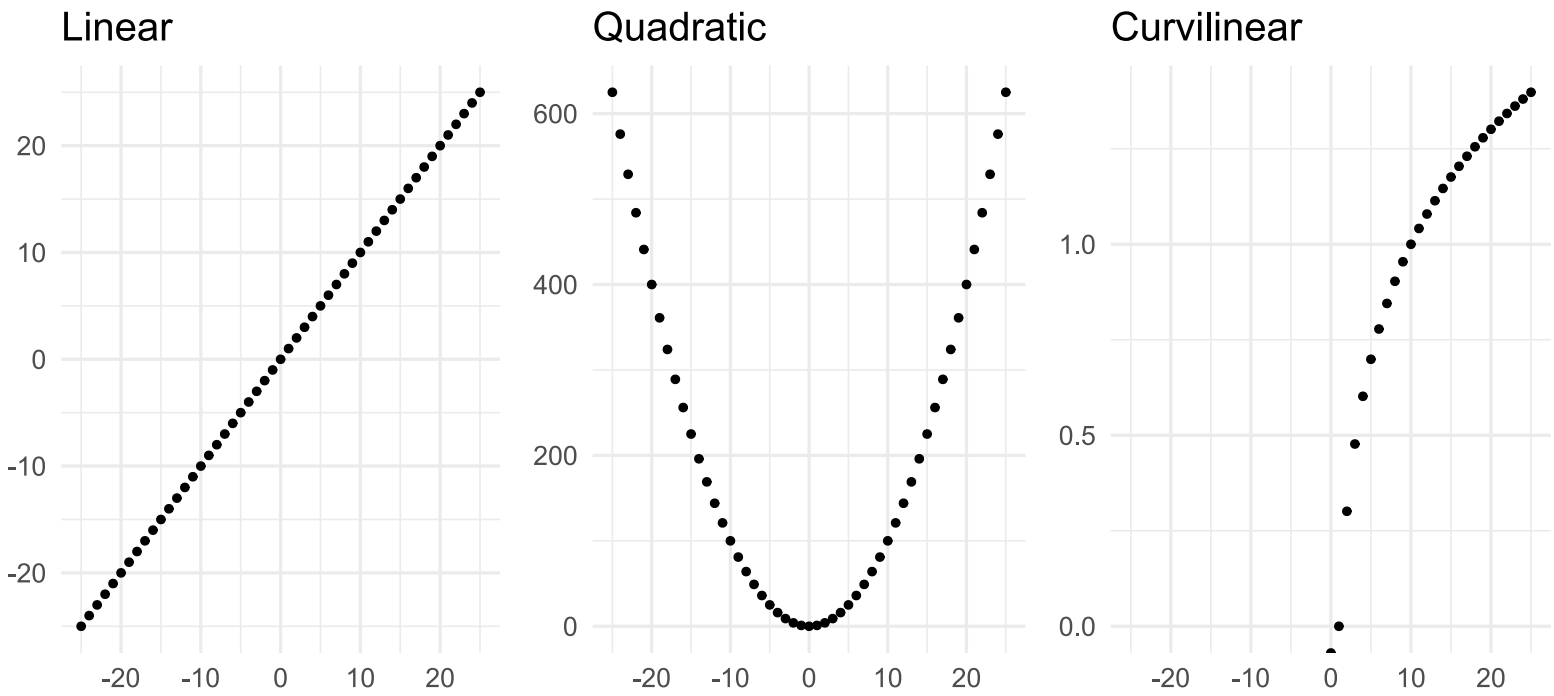
Statistical models

- Include individual variation
- Other systematic components exist that are either not measured or observable
- Outcome = Systematic component + residual

What is wrong about describing the relationship between BMI and eating restraint as we did on the previous slide?

Statistical model

In order to develop our statistical model (inspired by a deterministic mathematical model), we need to first determine our model's functional form.



All of the models we will focus on in this course are linear, but we're going to "cheat" to allow ourselves to model non-linear relationships with linear models.

Linear models

Why are straight lines so popular in statistics?

1. Mathematical simplicity: straight lines are one of the simplest (and consistent) ways of characterizing relationships
2. Actual linearity: many relationships are best characterized linearly; over a small enough interval, every relationship is linear

What if my data isn't related in linear ways

1. Transformations: we will learn how to use these later to fit linear models to curvilinear data
2. Limited ranges of X may yield linearity: most things are "locally linear"

In fact, linear modeling is so tractable that this is what we will spend the entire course on!

We'll learn about lots of different modeling approaches. **They are all part of the same family of models, known as the General Linear Model (GLM)**. We will learn more about this GLM in Unit 4.

Linear reqression equations

Okay, so once we have selected our model's functional form (which for now and the foreseeable future is going to be "linear"), we can move on to a mathematical representation. In this case, we are going to postulate a linear regression model.

A linear regression equation borrows the mathematical framework for a line to summarize this relationship.

$$BMI = \beta_0 + \beta_1(EDEQ\_restraint)$$

Regression equations

The regression line describes the mean value of Y for each possible "input" value of X.

$$BMI = \beta_0 + \beta_1(EDEQ\_restraint)$$

Our data already provides us with observations of over-eating frequency and dietary restraint:

```
do %>%
  select(EDEQ_restraint, BMI) %>%
  tail()
```

```
#> # A tibble: 6 x 2
#>   EDEQ_restraint   BMI
#>            <dbl> <dbl>
#> 1              3  36.0
#> 2              0  29.7
#> 3              6  22.3
#> 4              0.2  26.5
#> 5              1  38.0
#> 6              0  23.0
```

So, we need to find the best intercept (β_0) and slope (β_1) to represent the relationship.

Regression equation components

$$BMI = (\beta_0) + (\beta_1)(EDEQ\_restraint)$$

Intercept (β_0): Predicted outcome when x is equal to 0.

Slope (β_1): Predicted difference in the outcome for every one unit difference in x.

Write out the appropriate regression equation for a line that has an intercept of 20 and a Dietary Restraint slope of 1.5.

Interpret this equation in words, being careful to avoid making causal statements.

Regression equation components

Intercept (β_0): Predicted outcome when x is equal to 0.

Slope (β_1): Predicted difference in the outcome for every one unit difference in x.

$$BMI = 20 + 1.5(EDEQ\_restraint)$$

What is an individual's expected BMI given a Dietary Restraint rating of 2?

Regression equation components

Intercept (β_0): Predicted outcome when x is equal to 0.

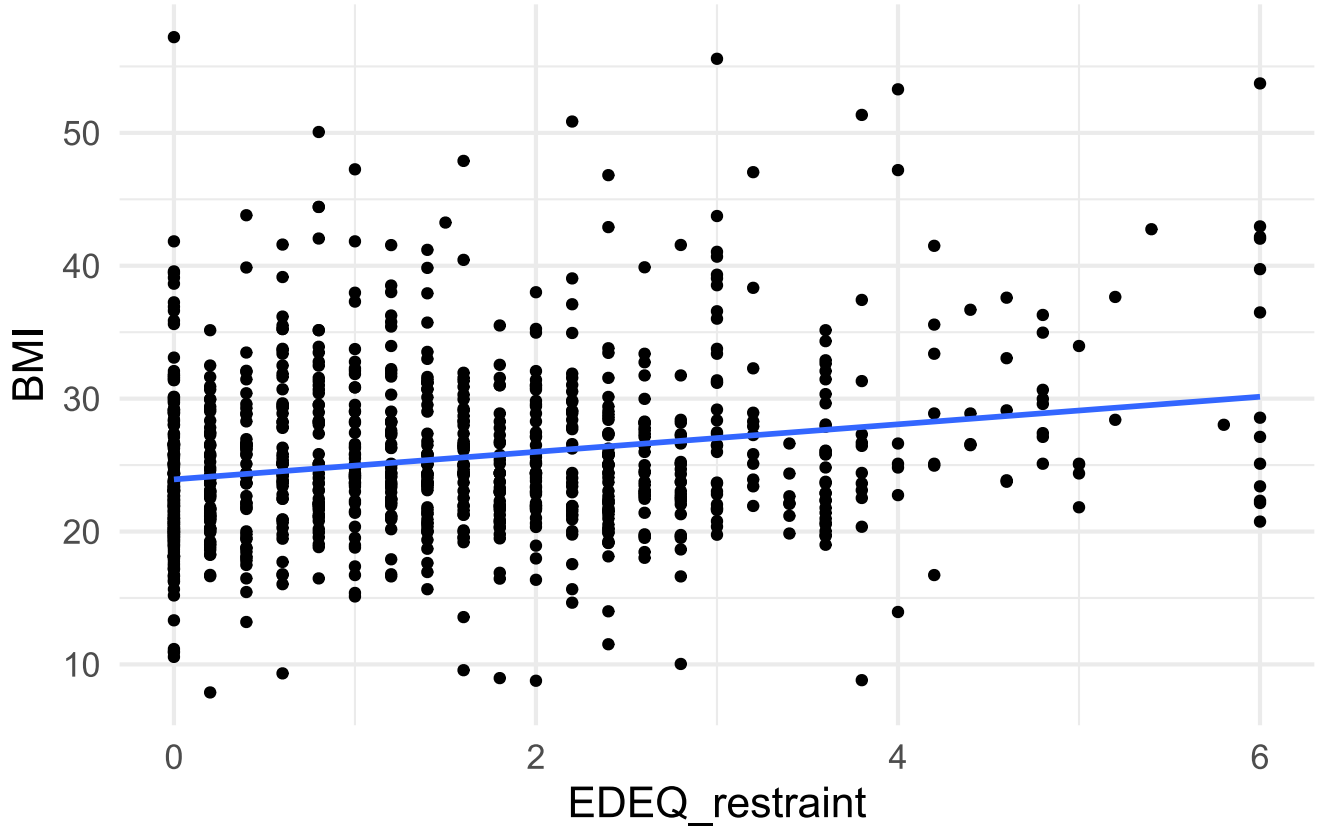
Slope (β_1): Predicted difference in the outcome for every one unit difference in x.

$$BMI = 20 + 1.5(2) = 20 + 3 = 23$$

We predict that individuals with an observed Dietary Restraint rating of 2 will have a body-mass index of 23.

Residual (error) term

"No model is perfect, but some are useful." - George Box



Is Dietary Restraint a perfect predictor of BMI? **NO!**

Omitted variables

A "perfect" regression equation would probably include a lot more variables:

$$BMI = \beta_0 + \beta_1(EDEQ\_restraint) + \beta_2(\text{Meal Frequency}) + \beta_3(\text{Nutritional Habits}) + \ldots \beta_\infty$$

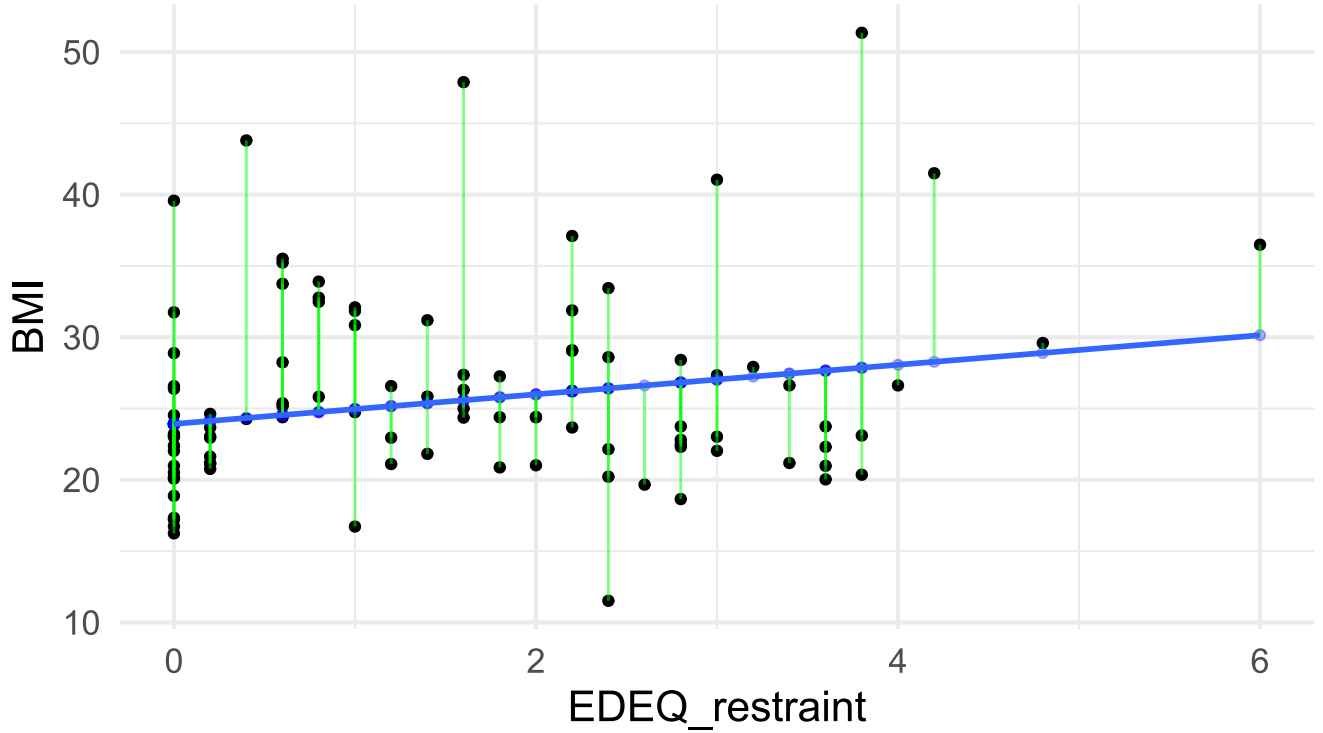
Many of these might not even be measurable!

Omitted variables are not a problem in themselves. Our estimates of the relationship between two (or more) variables might still be unbiased (i.e., accurately describe the nature of the relationship in the population), but they may be less precise because we have not explained all of the variation.

However, omitted variables often introduce bias into our estimates of the relationship. More on the problem of omitted variable bias later.

Residual (error) term

In a regression model, all the variability that we can't explain with our predictor(s) is condensed into a residual term (often represented by the Greek letter epsilon, ε).



$$BMI = \beta_0 + \beta_1(EDEQ\_restraint) + \varepsilon$$

The regression model

So there it is! Our full population regression model:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Outcome = Systematic component + residual[1]

> Y: our outcome
>
> β_0 and β_1: our population parameters and regression coefficients to be estimated
>
> ε: our error/residual (ε is a fancy way of writing the Greek letter "epsilon")

Also written as:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where we use the subscript i to emphasize that the model estimates the outcome for each of the i units (students, schools, patients, etc.).

[1] Sometimes also called deterministic and stochastic components.

A fitted model

A **fitted model** takes our population model and uses our observed data to derive estimates for the population.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

We denote that we are generating estimates with "hats" above the estimated coefficients. Note that there is ***no error term*** in our fitted model.

Residuals and model fit

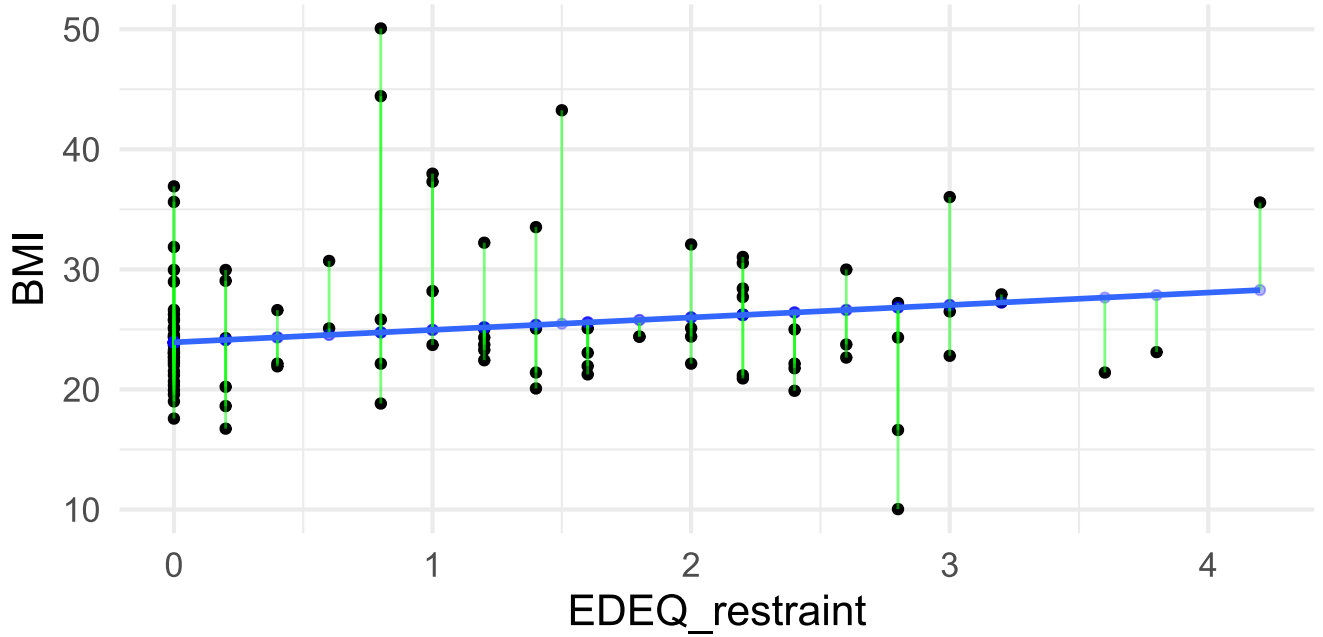
Residuals

For any observation, the residual is the difference between the observed and predicted value.

$$\varepsilon_i = Y_i - \hat{Y}_i$$

Ordinary Least Squares (OLS)

An OLS-fitted regression would go through the "center" of the data, finding the best intercept and slope to minimize the total distance between all of the residual and itself.



For some observations we under-predict, for others we over-predict. But, across the full sample, no matter where we put our line, the residuals will always sum to zero. So how do we calculate the "best" line?

Sum of the squared residuals

For any observation, the residual is the difference between the observed and predicted value.

$$\varepsilon_i = Y_i - \hat{Y}_i$$

Squaring the residual terms (ε_i^2) allows us to treat negative and positive deviations equally, and puts a greater penalty on larger deviations.

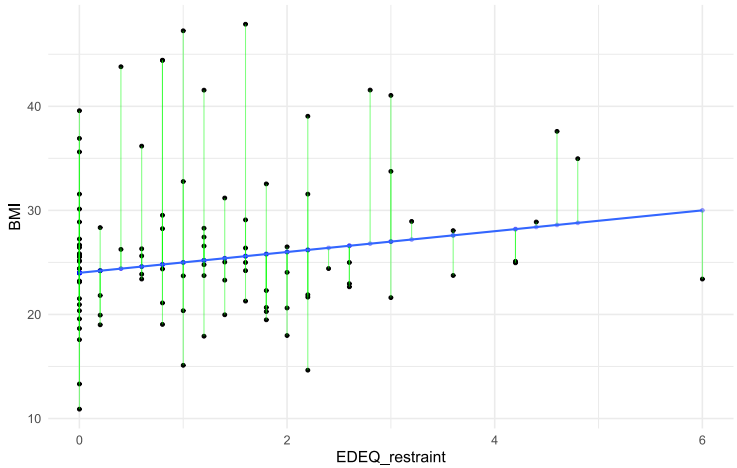
```
# A tibble: 6 x 4
    BMI predicted_BMI residual residual_sq
  <dbl>         <dbl>    <dbl>       <dbl>
1  36.0          27.0     8.99       80.8
2  29.7          23.9     5.76       33.2
3  22.3          30.1    -7.83       61.3
4  26.5          24.1     2.37        5.6
5  38.0          25.0    13.0       169.
6  23.0          23.9    -0.97        0.94
```

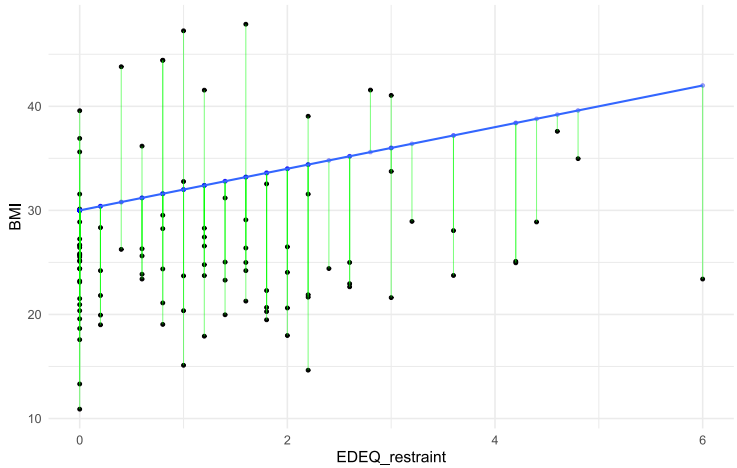
The **sum of squares** is the sum of all our squared residuals:

$$\Sigma(\varepsilon_i)^2$$

Ordinary Least Squares (OLS)

An OLS-fitted regression finds the best intercept and slope values to **minimize the sum of squared residuals**.

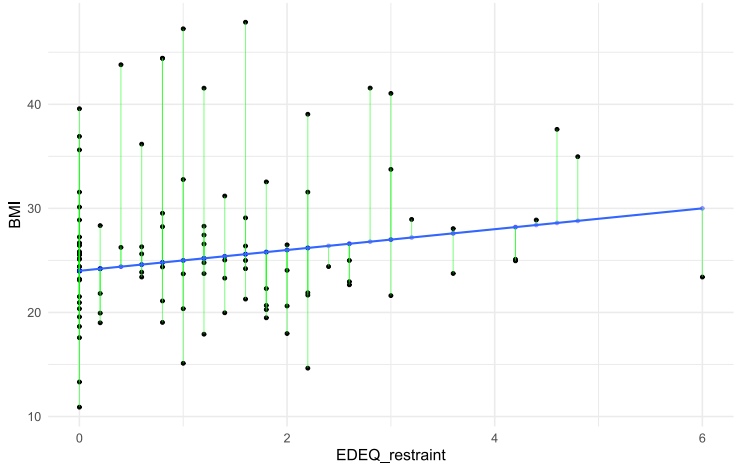




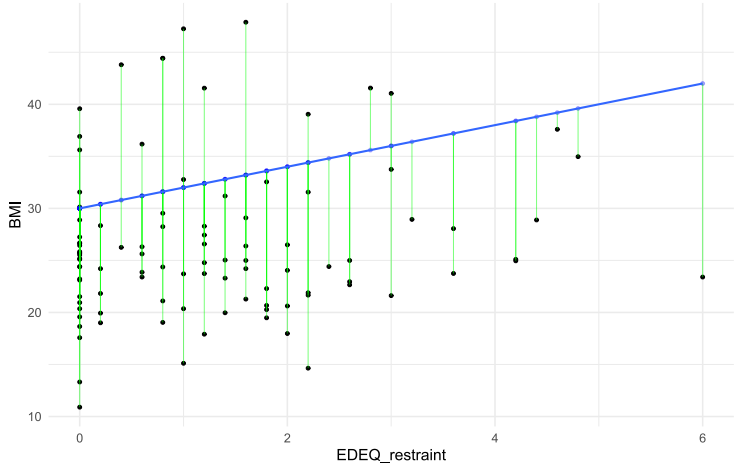
Which regression line appears to be a better fit?

Ordinary Least Squares (OLS)

An OLS-fitted regression finds the best intercept and slope values to minimize the sum of squared residuals.



$\mathbf{\Sigma(e_i)^2 = 40,153}$



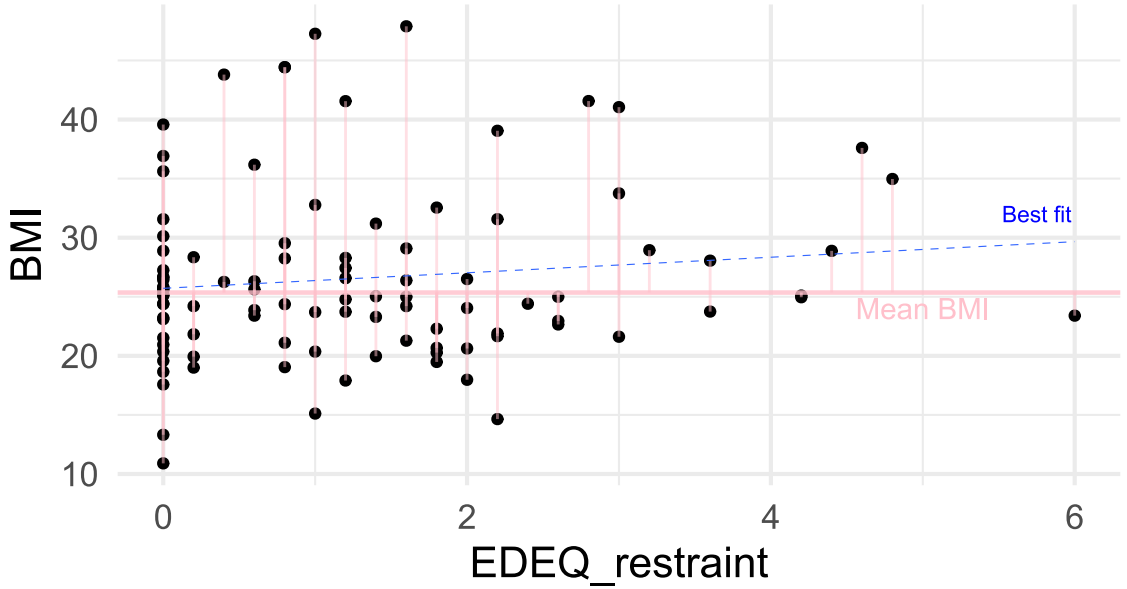
$\Sigma(e_i)^2 = 101,566$

The regression line on the left is a better fit because the sum of its squared errors is smaller (40,153 < 101,566).

Sum of Squared Residuals (SSR)

SSR is a bigger concept than just model predictions. It is a way of thinking about variability. Our outcome's variability is simply the sum of squared deviations from its mean, aka the sum of the squares (SS).[1]

$$SS_{\text{BMI}} = \sum (Y_i - \bar{Y})^2$$



[1] Note that this is a univariate statistic that depends only on the variability of BMI.

Partitioning variance

The goal of regression is to account for some of this variability with our model's predictors.

We can partition the outcome's total variance (SS_{BMI}) into:

- Model-accounted variance (SS_{Model})
- Residual variance (SS_{Residual})

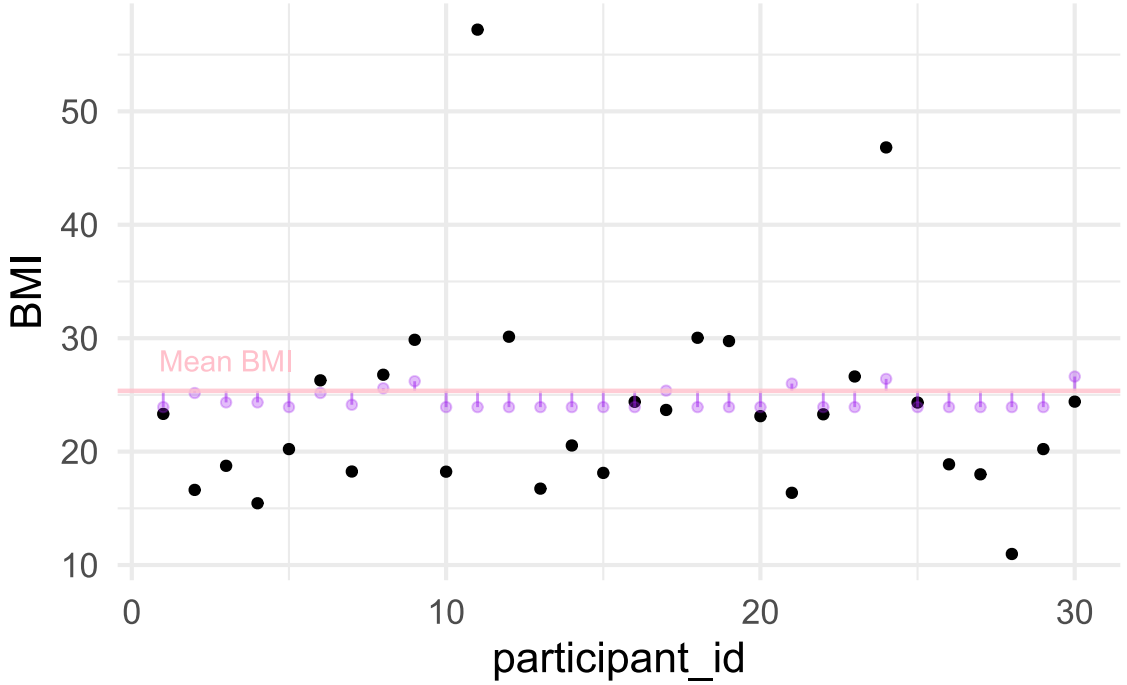
$$SS_{\text{BMI}} = SS_{\text{Model}} + SS_{\text{Residual}}$$

Remember our goal with OLS regression is to find the model coefficients that minimize SS_{Residual}.

Partitioning variance

$$\text{If } SS_{\text{BMI}} = SS_{\text{Model}} + SS_{\text{Residual}}$$

SS_{Model} is the sum of squared deviations between the model predicted values $(\hat{Y})$ and the mean $(\bar{Y})$. Let's visualize this for a set of 30 participants:

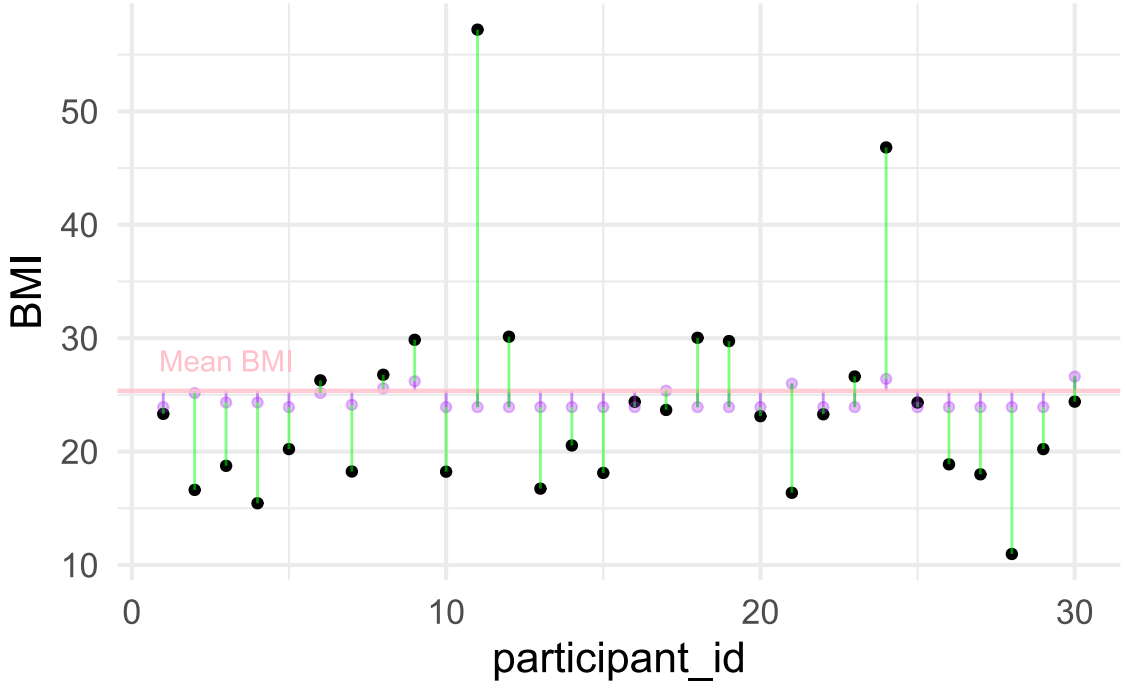


$$SS_{\text{Model}} = \sum (\hat{Y} - \bar{Y})^2$$

Partitioning variance

$$\text{If } SS_{\text{BMI}} = SS_{\text{Model}} + SS_{\text{Residual}}$$

SS_{Residual} is the sum of squared deviations between the model predicted values $(\hat{Y})$ and the observed values (Y).



$$SS_{\text{Residual}} = \sum (Y - \hat{Y})^2$$

Get a feel for trying to minimize the *sum of the square of the residuals*

Sums of Squares Visualization

Intercept:

0

Slope:

1

View Sums of Squares

- (●) Normal View
- ( ) View Residuals
- ( ) View Sums of Squares

Data Simulation

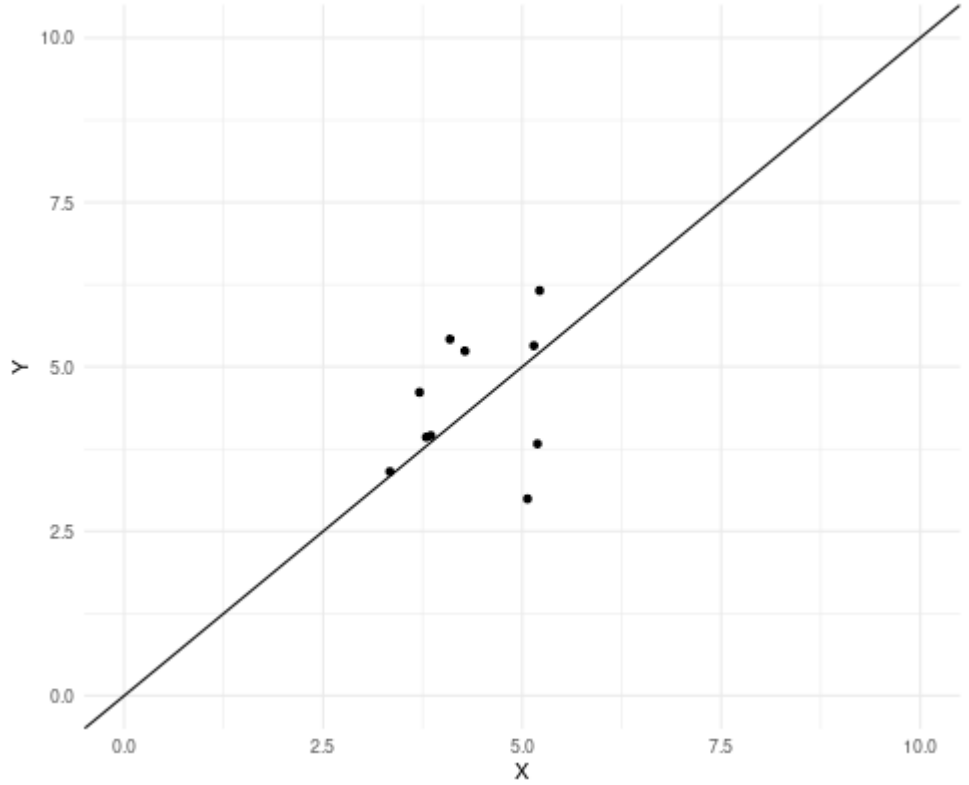
Mean of X:

5

Mean of Y:

5

Correlation



Sums of Squares = 10.59

R^2

Partitioning variance can be useful for model evaluation.

$$SS_{\text{BMI}} = SS_{\text{Model}} + SS_{\text{Residual}}$$

R^2 ("R-squared") is the proportion of variance in the outcome our model accounts for.

$$R^2 = \frac{SS_{\text{Model}}}{SS_{\text{BMI}}}$$

For example, if $R^2 = 0.30$, then our model accounts for 30% of our sample's variance in body-mass index. More on this soon.

Taking stock

We've developed some understanding of the formal structures of linear models (GLMs) and used some graphical representations to understand some of the components of a least-squares regression fit

Bivariate relationship characteristics:

- Direction
- Linearity
- Outliers
- Strength
- Magnitude

Model fit characteristics:

- Residuals
- Sum of Squared Residuals (SSR)
- Sum of Squared Model (SSM)
- R^2

Next, we're going to turn to actually fitting our regression so we can say something substantive about the relationship between dietary restraint behaviors and BMI.

Synthesis and wrap-up

Class goals

- 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

To-Dos

Reading:

- **By January 18 class**: LSWR Chapter 15.1 - 15.2 and 15.4 - 15.7 and Hu (2021)

Review:

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