Multiple Regression

EDUC 643: Unit 3

David D. Liebowitz

UNIVERSITY OF OREGON

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.

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

Goals for the unit

- Articulate the concepts of multiple regression and "statistical adjustment"
- Distinguish between the substantive implications of the terms "statistical control" and "statistical adjustment"
- Estimate the parameters of a multiple regression model
- Visually display the results of multiple regresion models
- State the main effects assumption and what the implication would be if it were violated
- Conduct statistical inference tests of single predictors (a t-test) and the full model (an F-test) in multiple regression
- Decompose the total variance into its component parts (model and residual) and use the R^2 statistic to describe this decomposition
- Describe problems for regression associated with the phenomenon of "multicollinearity"
- Use visual schema (e.g., Venn diagrams) to assess regression models for the potential of multicollinearity
- Use statistical results (e.g., correlation matrices or heat maps) to assess regression models for the potential of multicollinearity
- Describe and implement some solutions to multi-collinearity

Univariate statistics

We're interested in characterizing the relationship between over-eating frequency (*OE_frequency*) and dietary restraint behaviors (*EDEQ_restraint*), so we can start out by examining each of these variables independently.

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   2.281   4.000  29.000
```

```
sd(do$OE_frequency)
```

```
## [1] 3.733668
```

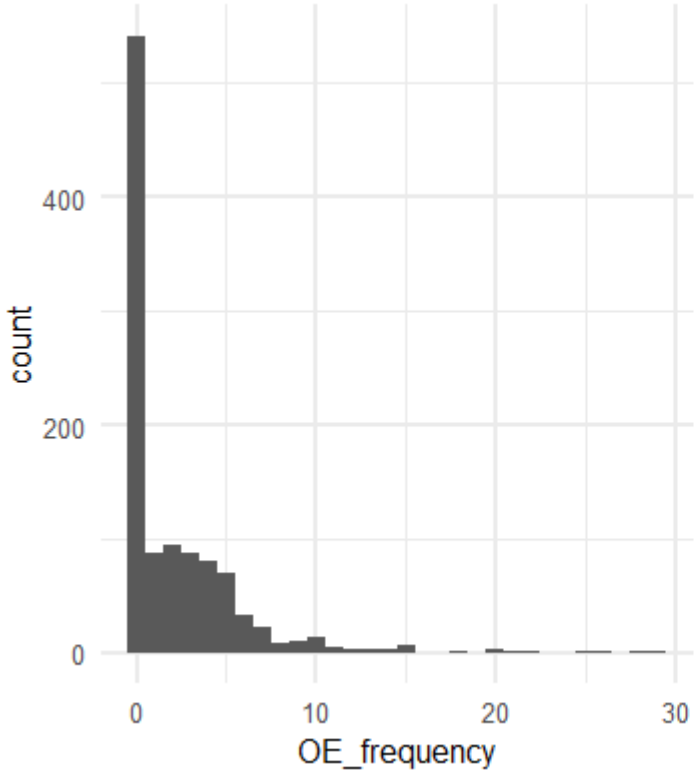
```
summary(do$EDEQ_restraint)
```

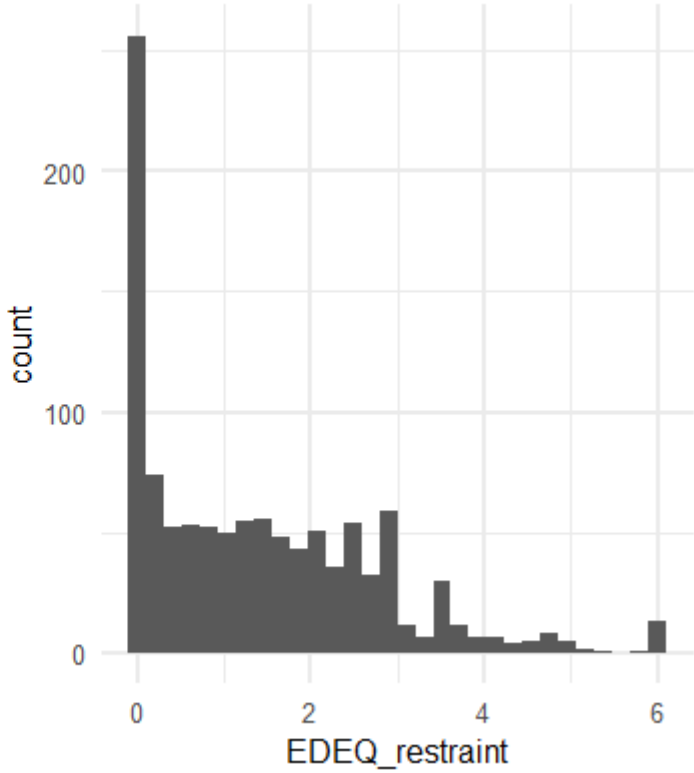
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.200   1.200   1.383   2.200   6.000
```

etc.

Univariate displays

Now some univariate visualizations...





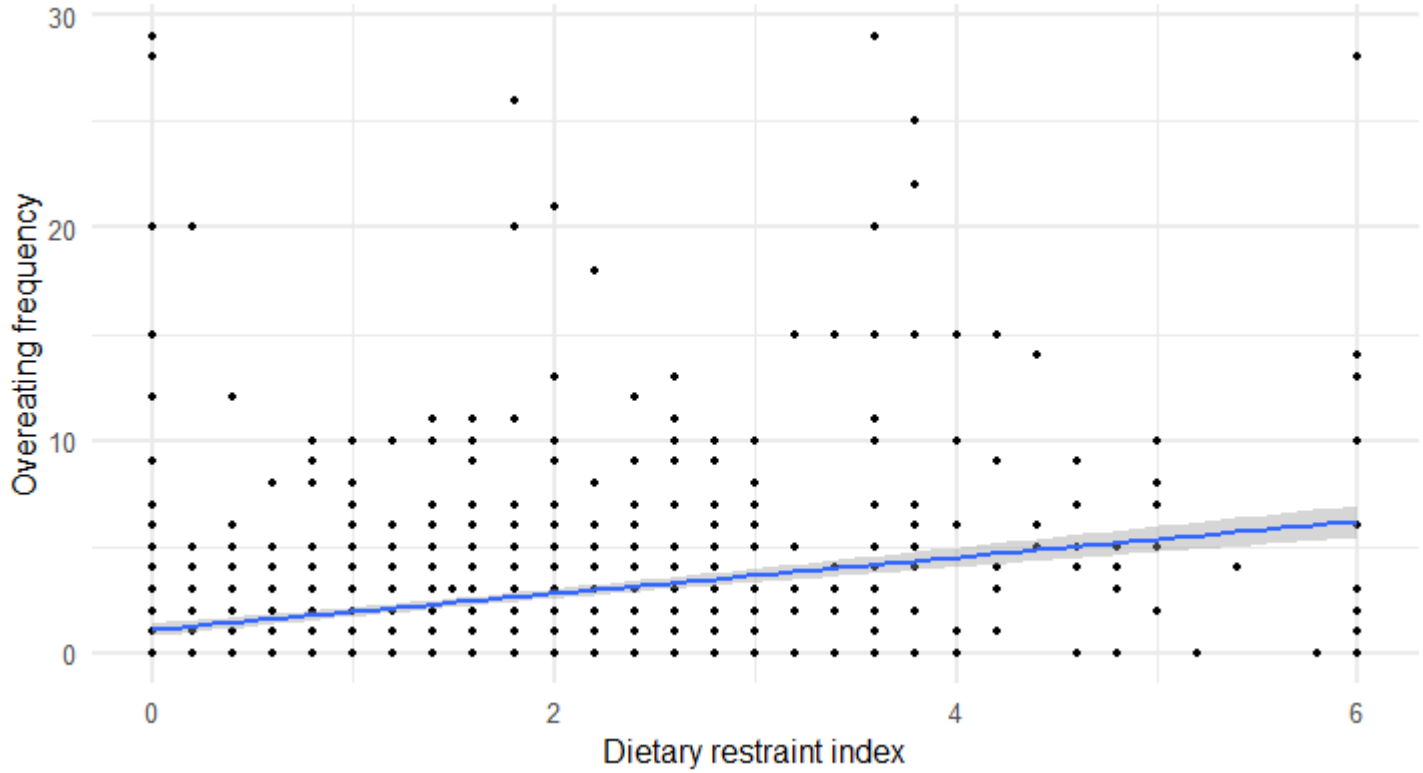
Bivariate relationship

Now some **bivariate** statistics...

```
cor(do$OE_frequency, do$EDEQ_restraint)
```

```
## [1] 0.3079139
```

Bivariate relationship



Based on what you see here and on the previous slides, what can we say about the direction, linearity, existence of outliers, strength and magnitude of this relationship? What evidence from these visuals and statistics supports/threatens our regression assumptions?

Regression results

```
##
## Call:
## lm(formula = OE_frequency ~ EDEQ_restraint, data = do)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2137 -1.7849 -1.1036  0.8964 27.8964
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.10358    0.15448   7.144 1.66e-12 ***
## EDEQ_restraint  0.85168    0.07997  10.651  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.554 on 1083 degrees of freedom
## Multiple R-squared:  0.09481,    Adjusted R-squared:  0.09398
## F-statistic: 113.4 on 1 and 1083 DF,  p-value: < 2.2e-16
```

Can you interpret this relationship substantively? What previous value have we seen related to a number displayed above?

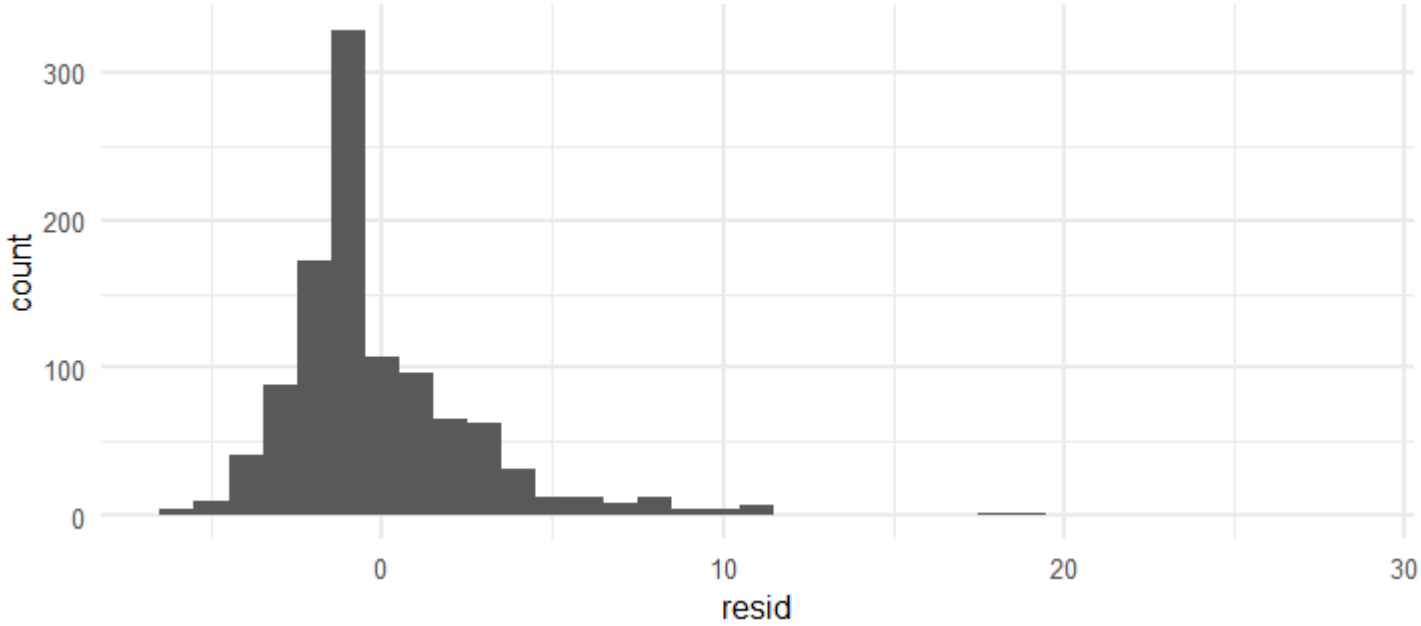
Regression diagnostics

We've looked a little already at the normality and linearity of our relationship. *What else can we look at?*

How can we test for homoscedasticity, linearity and normality in our residuals?

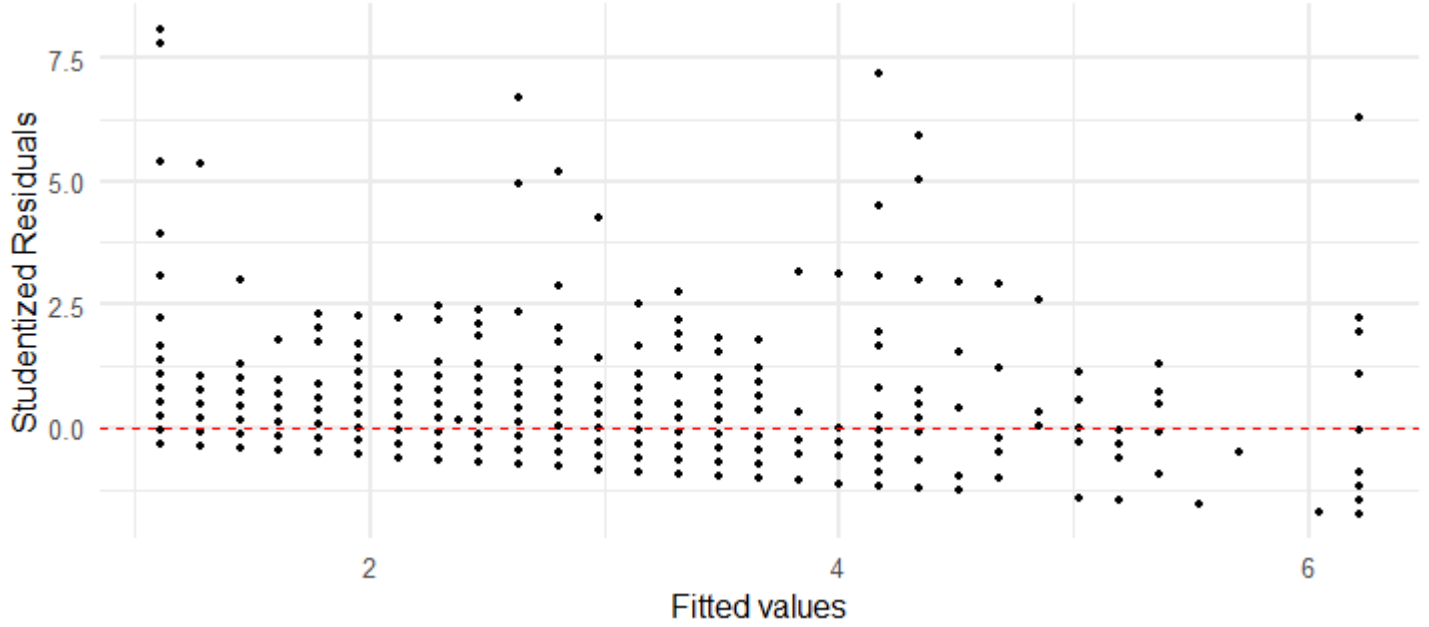
```
do$predict <- predict(fit)
do$resid <- resid(fit)
do$stu_resid <- rstudent(fit)
```

Regression diagnostics: normality



The bulk of our residuals are roughly normally distributed, but we clearly have a long right tail. Perhaps we would want to test the sensitivity of our models for the exclusion of these outlying values? What is driving this long right tail?

Residuals v. fitted



We are clearly under-predicting for some set of individuals. There is also some evidence of heteroscedasticity.

There are some orange flags in our estimates here. In fact, we are conducting a somewhat different analysis than Kelly et al. (2018). They conducted a logistic regression for the presence of *any* medical or psychological challenges. We will learn how to do that analysis in EDUC 645. However, we are going to proceed with our analysis while noting that some of our assumptions may not be fully met.

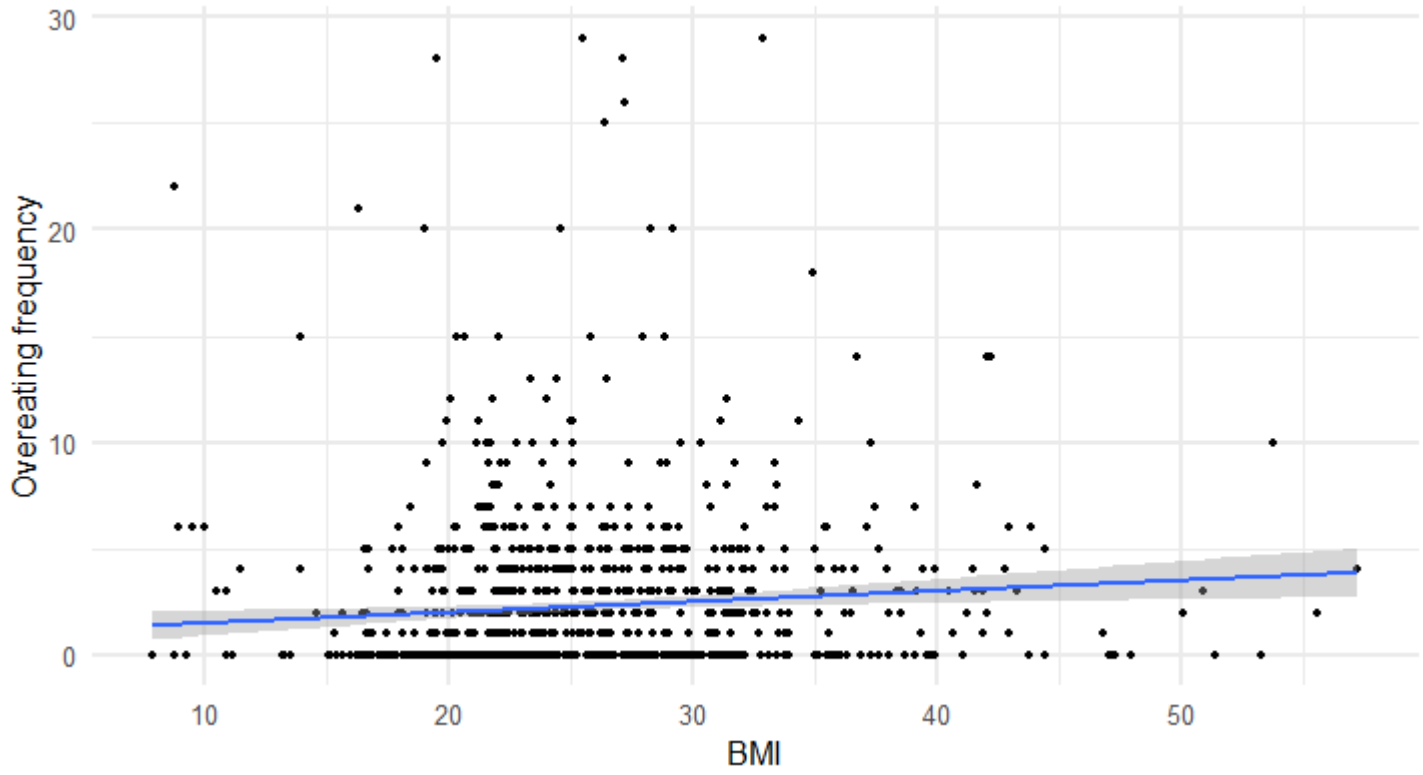
Regression results

```
##
## Call:
## lm(formula = OE_frequency ~ EDEQ_restraint, data = do)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2137 -1.7849 -1.1036  0.8964 27.8964
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.10358    0.15448   7.144 1.66e-12 ***
## EDEQ_restraint  0.85168    0.07997  10.651  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.554 on 1083 degrees of freedom
## Multiple R-squared:  0.09481,    Adjusted R-squared:  0.09398
## F-statistic: 113.4 on 1 and 1083 DF,  p-value: < 2.2e-16
```

Let's assume that we trust the way we've characterized this bivariate relationship. But perhaps there are other features among the participants that also influence their eating behaviors?

Another variable

Perhaps we should consider another variable that might also be related to overeating frequency (*OE_FREQUENCY*). What about one we already know a good deal about...*BMI*?



Another variable

Perhaps we should consider another variable that might also be related to overeating frequency (*OE_FREQUENCY*). What about one we already know a good deal about...*BMI*?

```
summary(lm(OE_frequency ~ BMI, data=do))
```

```
...
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.690 -2.194 -1.693  1.085 26.711
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.00178    0.47270   2.119  0.03430 *
## BMI          0.05046    0.01810   2.787  0.00541 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.722 on 1083 degrees of freedom
## Multiple R-squared:  0.007122,    Adjusted R-squared:  0.006205
## F-statistic: 7.769 on 1 and 1083 DF,  p-value: 0.005409
...
```

Multiple regression

aka, statistical adjustment

Multiple regression

Mathematically, we simply add additional terms to our equation like this:

$$OE\_frequency_i = \beta_0 + \beta_1 DietaryRestraint_i + \beta_2 BMI_i + \varepsilon_i$$

or more generally for k predictors...

$$Y_i = \beta_0 + \beta 1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon_i$$

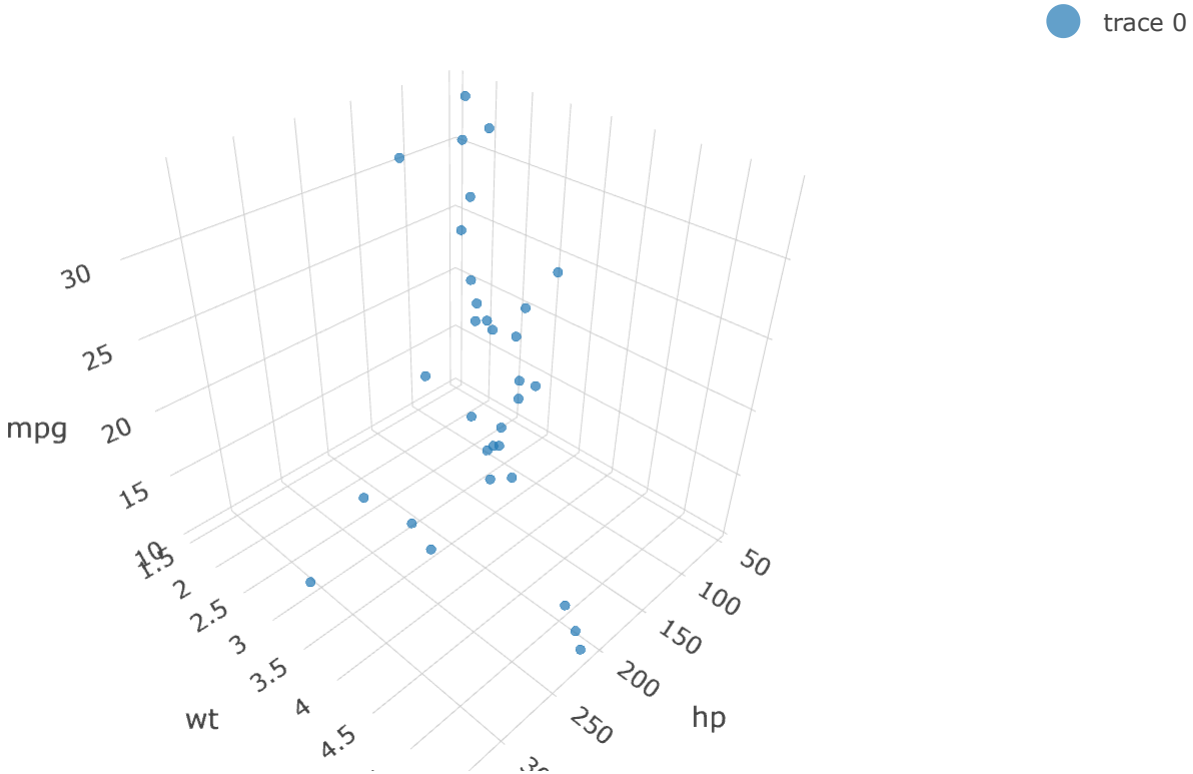
Implement in R

We can estimate our postulated linear model in R as follows:

```
summary(lm(OE_frequency ~ EDEQ_restraint + BMI, data=do))
```

```
...
## lm(formula = OE_frequency ~ EDEQ_restraint + BMI, data = do)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3068 -1.7590 -1.0768  0.9079 27.8809
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.871725   0.451670   1.930   0.0539 .
## EDEQ_restraint 0.841637   0.082079  10.254   <2e-16 ***
## BMI            0.009692   0.017741   0.546   0.5850
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.555 on 1082 degrees of freedom
## Multiple R-squared:  0.09506,    Adjusted R-squared:  0.09339
## F-statistic: 56.83 on 2 and 1082 DF,  p-value: < 2.2e-16
```

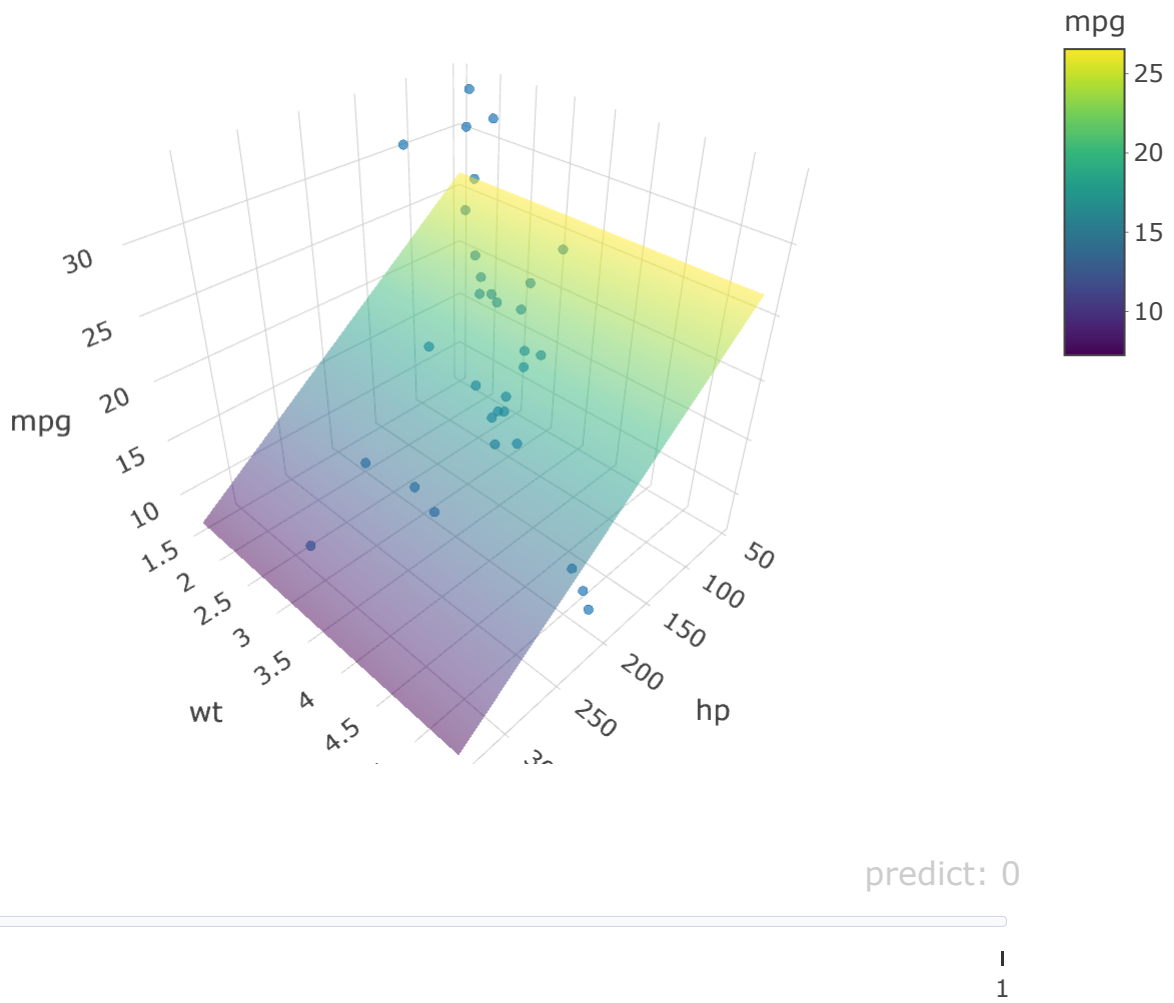
What does MR look like?



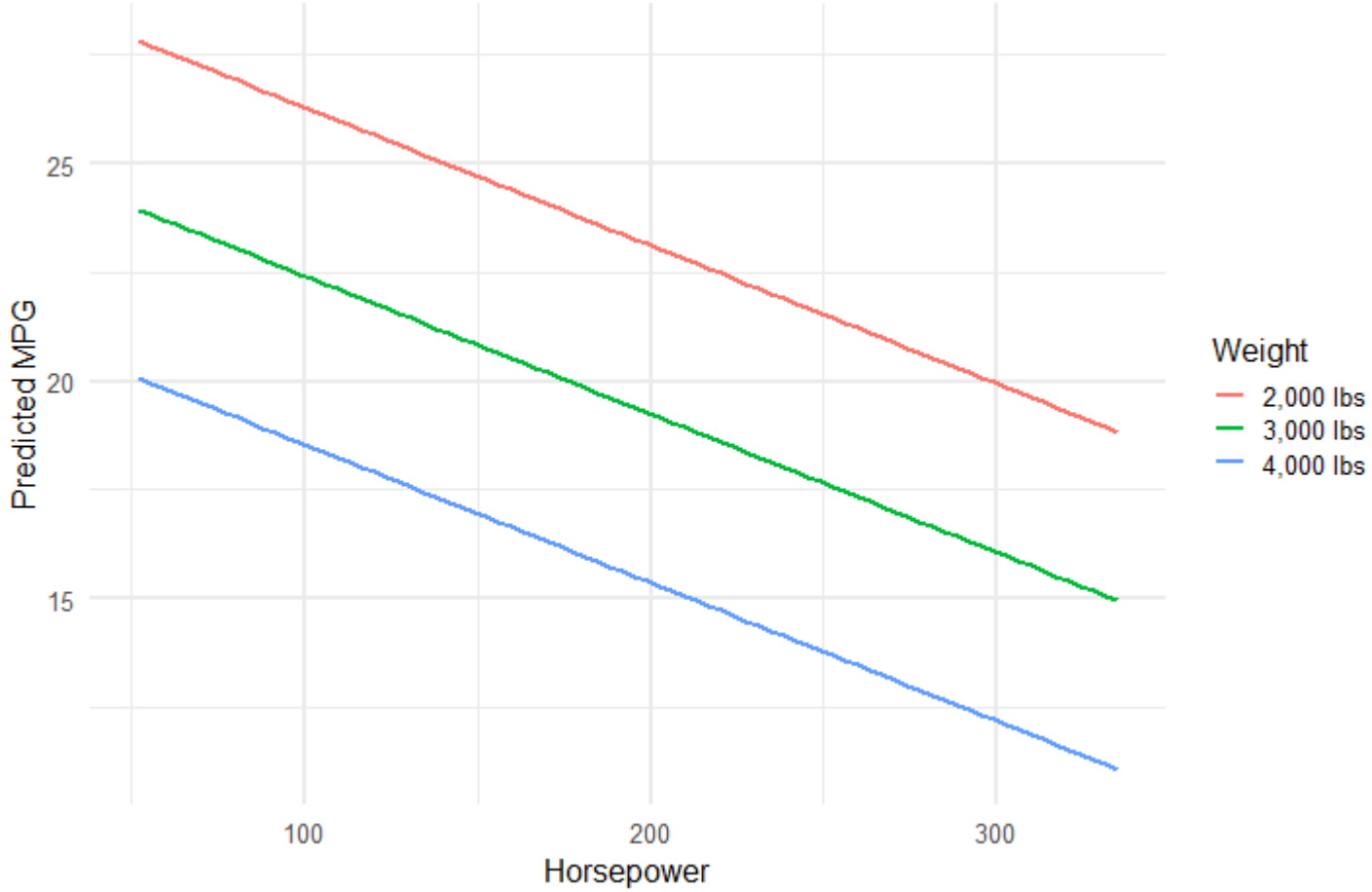
predict: 0

Play

What does MR look like?

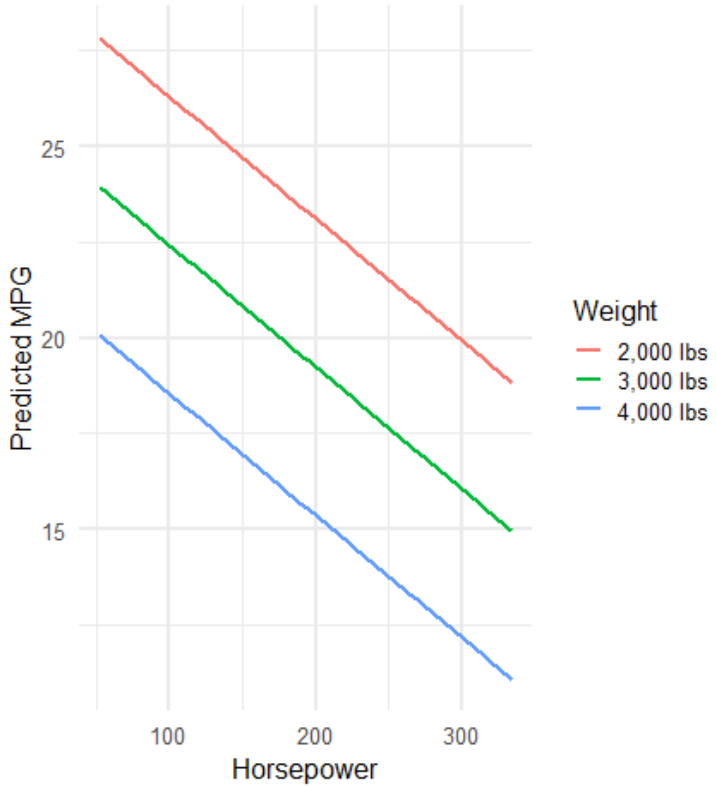


A return to 2D-land



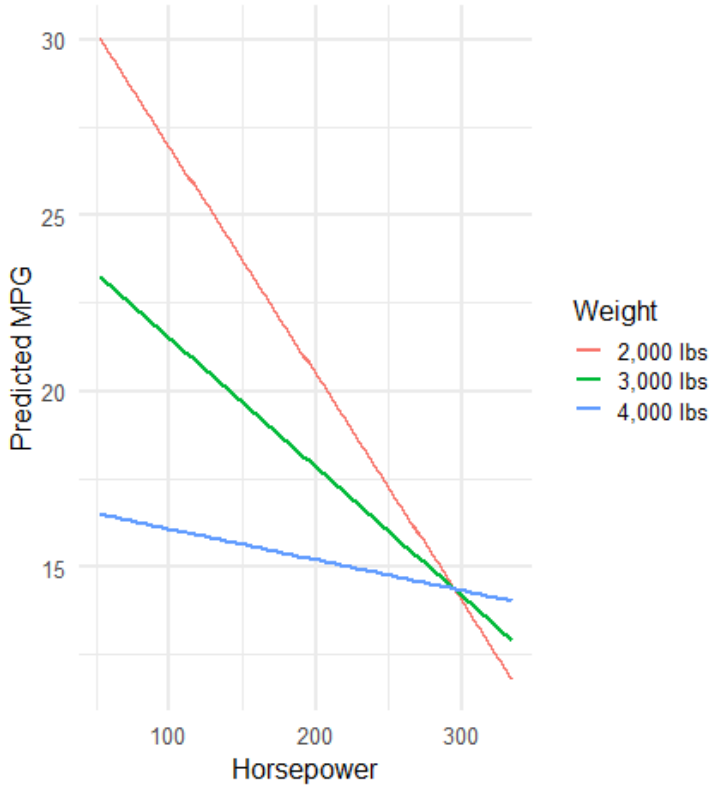
And we could do the same for any other predictor variable (e.g., wt)!

Interpreting multiple regression



- On average, cars with worse horsepower have worse mpg
  - This is true at various car weights
- **AND** on average, heavier cars have worse mpg
  - This is true across the distribution of cars' horsepower
- These lines are parallel (have a common slope)
  - This is only true because we have assumed them to be; aka **the main effects assumption**
  - We assume the relationship between each predictor and the outcome are equivalent, independent of the levels of the other predictors

Interpreting multiple regression



- On average, cars with worse horsepower have worse mpg
  - This is true at various car weights
- **AND** on average, heavier cars have worse mpg
  - This is true across the distribution of cars' horsepower
- Note that these lines are parallel (have a common slope)
  - This is only true because we have assumed them to be; aka **the main effects assumption**
  - We assume the relationship between each predictor and the outcome are equivalent, independent of the levels of the other predictors
- This doesn't need to be true, and we will relax this assumption in Unit 5

Multiple regression assumptions

We make **the same** assumptions as in bivariate regression, but now extended to multiple variables:

1. At each ***combination of the Xs***, there is a distribution of Y with a given mean $(\mu_{Y|X_1 \ldots X_k})$ and variance $(\sigma^2_{Y|X_1 \ldots X_k})$
2. The relationship between the points can be correctly characterized by a ***flat plane*** through the means
3. The variances (σ^2) of the distributions ***at the combination of the Xs*** are homoscedastic
4. Conditional on the ***combination of the Xs***, the values of Y are independent of each other
5. At each ***combination of the Xs***, the values of Y are normally distributed

Interpreting our results

```
...
## lm(formula = OE_frequency ~ EDEQ_restraint + BMI, data = do)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3068 -1.7590 -1.0768  0.9079 27.8809
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.871725   0.451670   1.930   0.0539 .
## EDEQ_restraint 0.841637   0.082079  10.254   <2e-16 ***
## BMI            0.009692   0.017741   0.546   0.5850
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.555 on 1082 degrees of freedom
## Multiple R-squared:  0.09506,    Adjusted R-squared:  0.09339
## F-statistic: 56.83 on 2 and 1082 DF,  p-value: < 2.2e-16
...
```

Interpretation of the intercept: we estimate that individuals with a dietary restraint index=0 and BMI score=0 will have an overeating score of 0.87

Interpreting our results

```
...
## lm(formula = OE_frequency ~ EDEQ_restraint + BMI, data = do)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3068 -1.7590 -1.0768  0.9079 27.8809
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.871725   0.451670   1.930   0.0539 .
## EDEQ_restraint 0.841637   0.082079  10.254   <2e-16 ***
## BMI            0.009692   0.017741   0.546   0.5850
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.555 on 1082 degrees of freedom
## Multiple R-squared:  0.09506,    Adjusted R-squared:  0.09339
## F-statistic: 56.83 on 2 and 1082 DF,  p-value: < 2.2e-16
...
```

Interpretation of *EDEQ_restraint*: adjusting for individuals' BMI, we estimate that young men who score one unit apart on the dietary-restraint index will have 0.84 unit different overeating scores.

Interpreting our results

```
...
## lm(formula = OE_frequency ~ EDEQ_restraint + BMI, data = do)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3068 -1.7590 -1.0768  0.9079 27.8809
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.871725   0.451670   1.930   0.0539 .
## EDEQ_restraint 0.841637   0.082079  10.254   <2e-16 ***
## BMI            0.009692   0.017741   0.546   0.5850
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.555 on 1082 degrees of freedom
## Multiple R-squared:  0.09506,    Adjusted R-squared:  0.09339
## F-statistic: 56.83 on 2 and 1082 DF,  p-value: < 2.2e-16
...
```

Interpretation of *BMI*: adjusting for individuals' dietary restraint behavior, we estimate that young men who have one unit different body-mass indices will have 0.01 unit different overeating scores. In fact...

Multiple regression

We can now use these regression estimates to construct our fitted equation:

$$\hat{OE\_frequency_i} = 0.872 + 0.842 * DietaryRestraint_i + 0.010 * BMI_i + \varepsilon_i$$

Looking at the results from the previous slides, has including the covariate *BMI* clarified the relationship between *OE_frequency* and *EDEQ_restraint*? Why or why not?

What if I have more than two predictors???

We can no longer display this graphically (after all we don't live in a 11-dimensional world), but the same theoretical ***and mathematical*** principles apply. Our regression estimates the coefficient for each predictor, while adjusting for all other relationships between our other covariates and our outcome; each time collapsing the multi-dimensional relationship to a two-dimensional one.

Multiple regression:

Affordances and limitations

Power of multiple regression

Multiple regression helps us as follows:

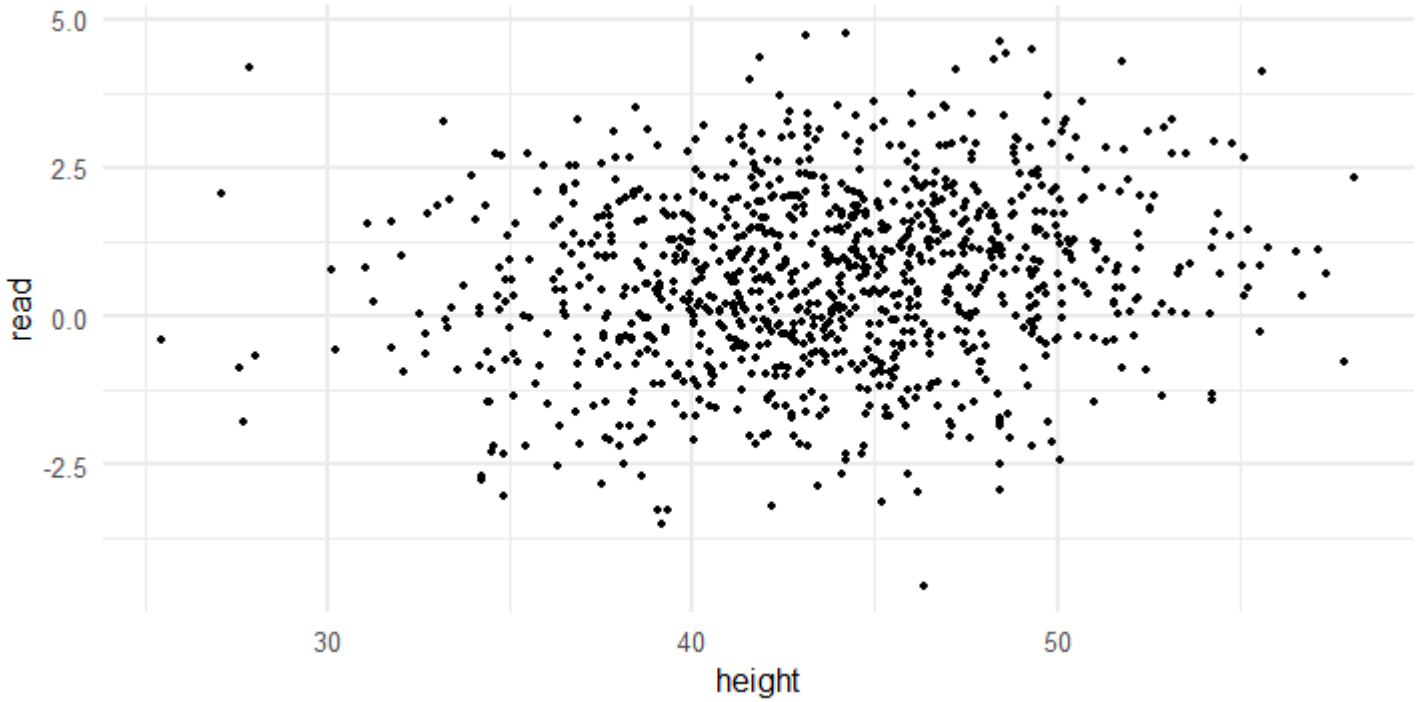
1. Allows us to simultaneously consider many contributing factors in the relationship
2. We explain more of the variation in Y, and
3. We make more accurate predictions of Y (#2 and 3 both make our residuals smaller)
4. Provides a separate understanding of the relationship between each predictor and our outcome, adjusting for the effects of the other predictors (that is, holding the other predictors constant at their means)

If we are ready to believe some **very** strong assumptions, multiple regression might even be able to characterize a credibly causal relationship between two variables of interest. However, there are some important limits to what multiple regression can do. More in a bit!

Power of multiple regression

Multiple regression can be a powerful tool to adjust for sample differences that depend on a variable other than the one in which we are interested and focus on the key question we have.

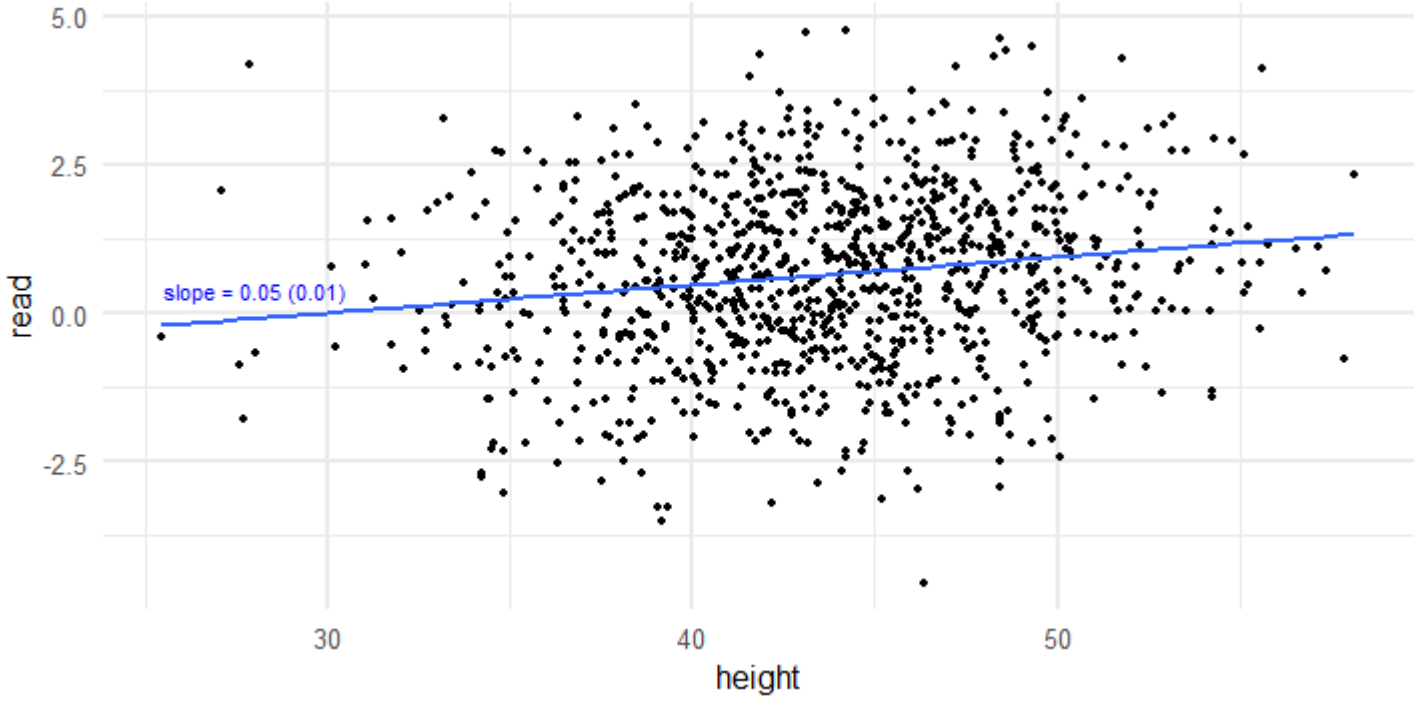
Take this example of a theoretical relationship between height and reading ability:



Power of multiple regression

Multiple regression can be a powerful tool to adjust for sample differences that depend on a variable other than the one in which we are interested and focus on the key question we have.

Take this example of a theoretical relationship between height and reading ability:

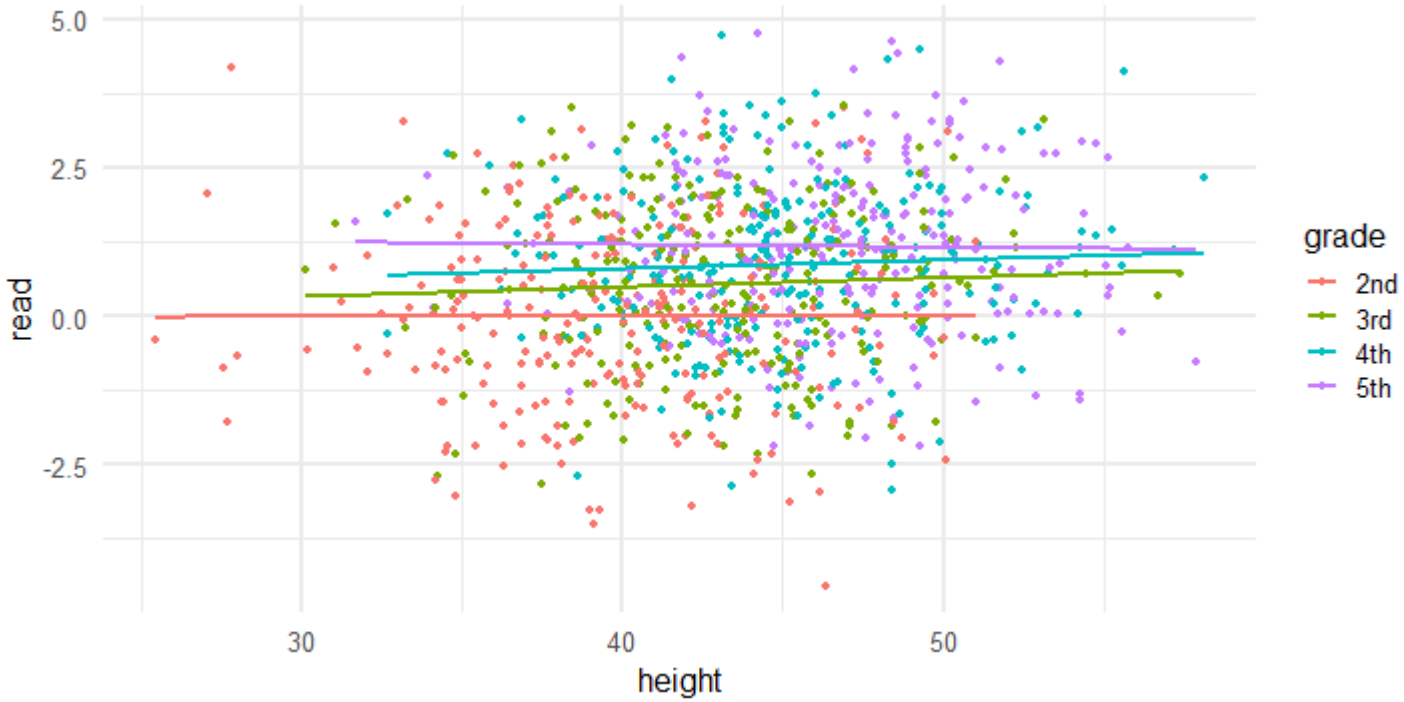


Do we really believe this or are there statistical adjustments we can make to reveal the true nature of the relationship?

Power of multiple regression

Multiple regression can be a powerful tool to adjust for sample differences that depend on a variable other than the one in which we are interested and focus on the key question we have.

Take this example of a theoretical relationship between height and reading ability:

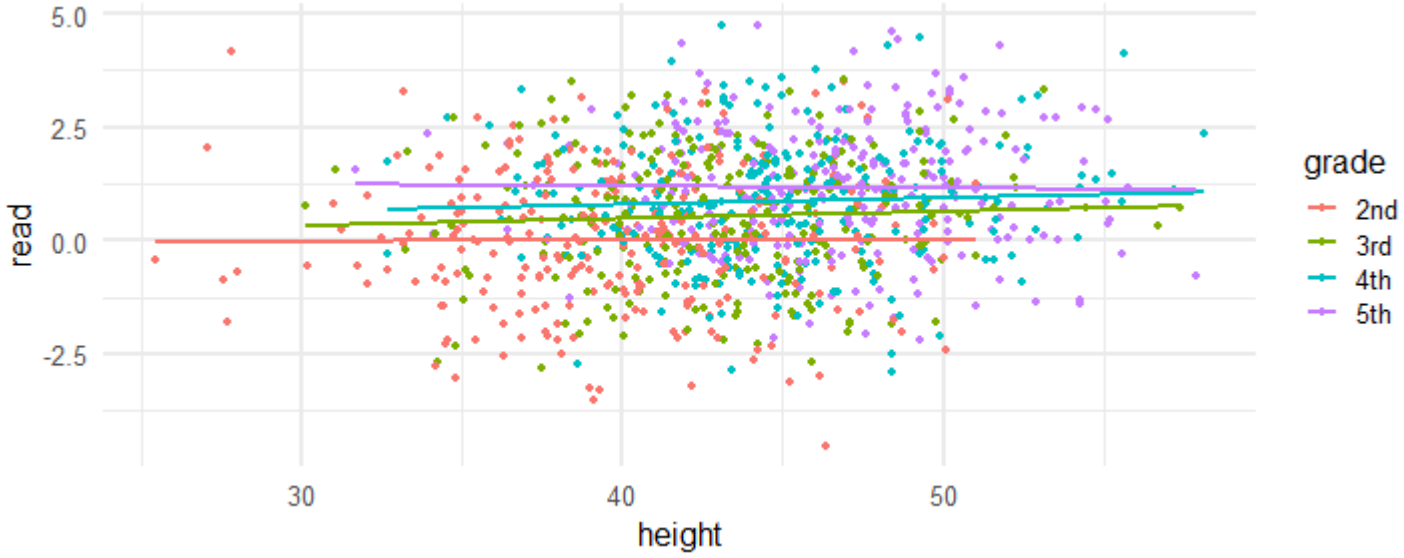


The weighted average of these slopes is the relationship between height and reading score, *after accounting for students' grade*.

Implementing in R

How to add a third variable in `ggplot`:

```
ggplot(data=reading, aes(x=height, y=read, color=grade)) +
  geom_point() +
  geom_smooth(method='lm', se=F) +
  theme_minimal(base_size = 16)
```



The weighted average of these slopes is the relationship between height and reading score, *after accounting for students' grade*.

Power of multiple regression

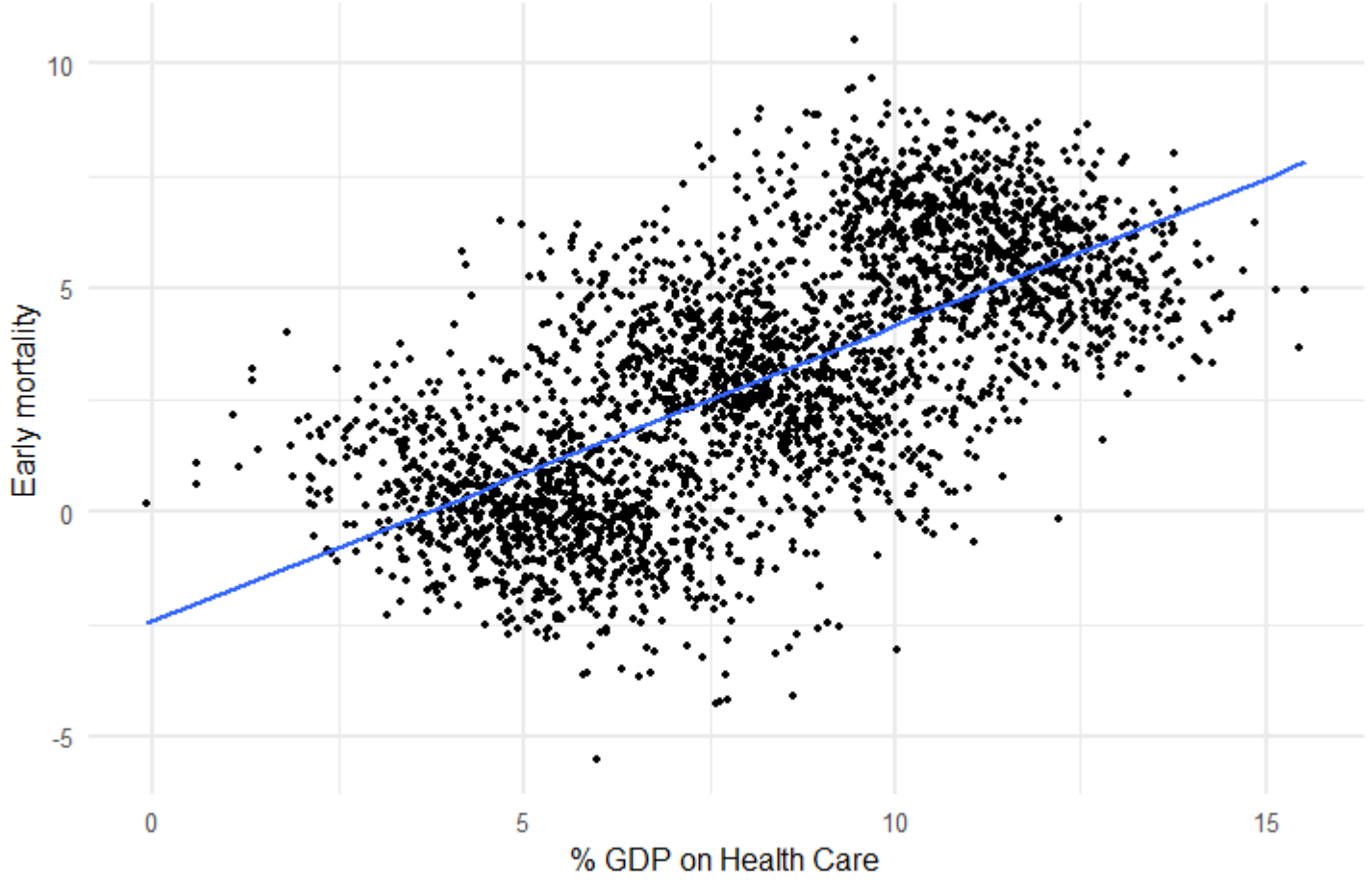
Formally testing this:

```
summary(lm(read ~ height + grade, data=reading))
```

```
##
## Call:
## lm(formula = read ~ height + grade, data = reading)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5747 -0.9787  0.0229  0.9910  4.2605
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.266133   0.407565  -0.653 0.513918
## height       0.006718   0.009997   0.672 0.501727
## grade3rd     0.510637   0.131011   3.898 0.000104 ***
## grade4th     0.831877   0.136910   6.076 1.75e-09 ***
## grade5th     1.123386   0.144658   7.766 2.01e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.417 on 995 degrees of freedom
```

Power of multiple regression II

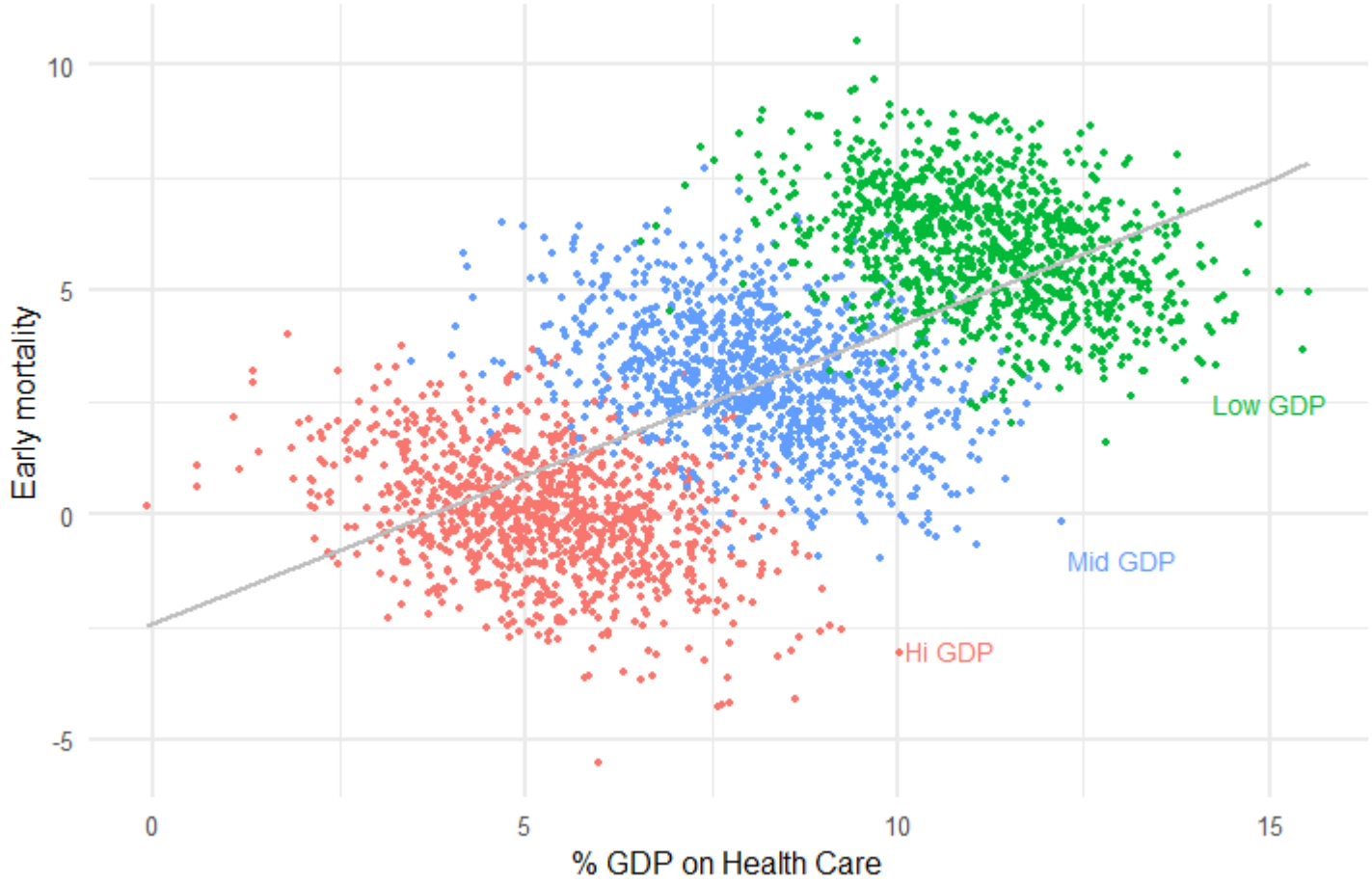
Multiple regression can solve a dilemma we introduced earlier known as Simpson's Paradox.



Seems surprising, right?

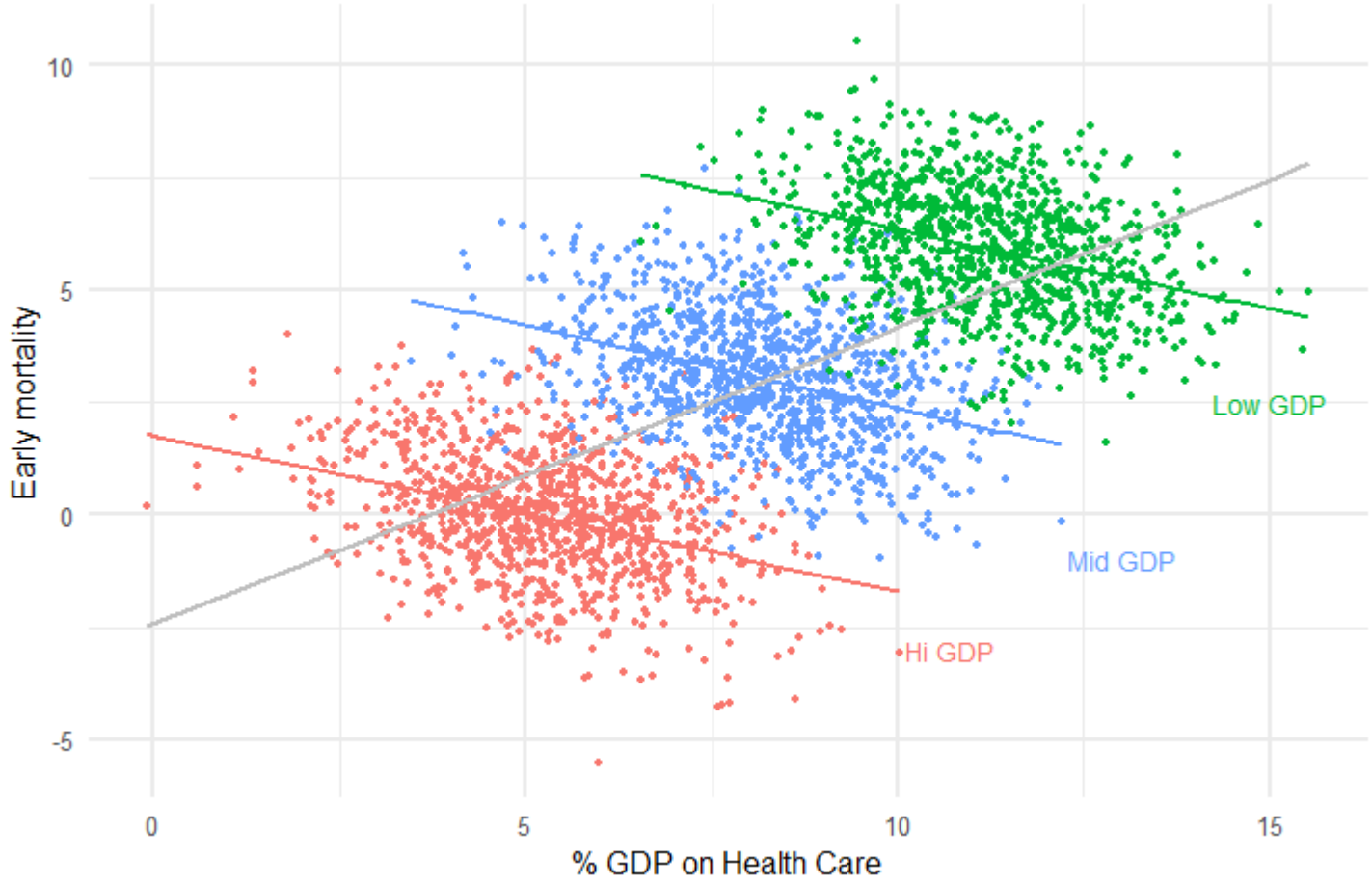
Power of multiple regression II

Multiple regression can solve a dilemma we introduced earlier known as Simpson's Paradox.



Power of multiple regression II

Multiple regression can solve a dilemma we introduced earlier known as Simpson's Paradox.



In pursuit of causality

In fact, if we can be confident that our general linear model includes all observable (prior score, demographics, SES) determinants for why some people receive or have higher values of a "treatment" such that there are no unobservable characteristics (motivation, attitudes, preferences) (and our temporal precedence condition is also met), then our regression estimates for the relationship between that predictor and our outcome can be credibly interpreted as causal in nature. This is sometimes referred to as "selection on observables."

What would have to be true of our postulated model for us to be able to interpret the relationship between *dietary restraint behaviors* and *over-eating patterns* as *causal* in our context?

Again, we have whole classes dedicated to just these topics of experimental and quasi-experimental techniques (EDLD 650 & EDLD 679).

The limits of multiple regression

...or why can't I say "control" like every other stats student everywhere?

We call the most common form of experimental research a randomized controlled trial because the investigator/researcher can **"control"** whether individuals do or do not receive a given treatment. As a result of this design, we can be confident that all observable and unobservable characteristics are equivalent across the two groups. This is not what we are doing in multiple regression!

What are some reasons why a multiple regression with a rich set of covariates might still not "control" for reasons why some people exhibit different patterns of dietary restraint than others?

> Synonyms you may encounter for statistical adjustment: **"controlling for,"** **"partialling out,"** or **"holding constant"**

We encounter a different kind of limit to the power of multiple regression in the presence of multi-collinearity...more on that next class...

Multiple regression in action

Estimate in R

```
summary(lm(OE_frequency ~ EDEQ_restraint + BMI, data=do))
```

```
...
## lm(formula = OE_frequency ~ EDEQ_restraint + BMI, data = do)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3068 -1.7590 -1.0768  0.9079 27.8809
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.871725   0.451670   1.930   0.0539 .
## EDEQ_restraint 0.841637   0.082079  10.254   <2e-16 ***
## BMI            0.009692   0.017741   0.546   0.5850
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.555 on 1082 degrees of freedom
## Multiple R-squared:  0.09506,    Adjusted R-squared:  0.09339
## F-statistic: 56.83 on 2 and 1082 DF,  p-value: < 2.2e-16
...
```

It's all the GLM...

Oh, by the way have you ever heard of partial correlations?

```
ppcor::pcor.test(x=do$OE_frequency, y=do$EDEQ_restraint, z=do$BMI)
```

```
##   estimate      p.value statistic    n gp  Method
## 1 0.297606 1.310824e-23    10.254 1085  1 pearson
```

```
partial1 <- lm(OE_frequency ~ BMI, data=do)
do$OE_partial <- resid(partial1)

partial2 <- lm(EDEQ_restraint ~ BMI, data=do)
do$EDEQ_partial <- resid(partial2)

cor(do$OE_partial, do$EDEQ_partial)
```

```
## [1] 0.297606
```

Goodness of fit statistics

```
summary(lm(OE_frequency ~ EDEQ_restraint + BMI, data=do))
```

```
...
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3068 -1.7590 -1.0768  0.9079 27.8809
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.871725   0.451670   1.930   0.0539 .
## EDEQ_restraint 0.841637   0.082079  10.254   <2e-16 ***
## BMI            0.009692   0.017741   0.546   0.5850
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.555 on 1082 degrees of freedom
## Multiple R-squared:  0.09506,    Adjusted R-squared:  0.09339
## F-statistic: 56.83 on 2 and 1082 DF,  p-value: < 2.2e-16
...
```

This is a better fit (with more variance explained) than with *BMI* alone **...why? What does it mean that this is no better fit than with *EDEQ_restraint* alone?**

Variance decomposition in MR

Just as before:

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

The sum of the squares of the model are now calculated based on the distance each observation has from the fitted regression ***plane*** (or hyper-plane):

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

The total sum of the squares is as before:

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

and R^2 is just $\frac{SS_{\text{Model}}}{SS_{\text{BMI}}}$ or more generally $\frac{SS_{\text{Model}}}{SS_{\text{Total}}}$

Inference in multiple regression

Two diferent hypotheses

Omnibus F-test

- Across all predictors, is any of them related to my outcome?
- $H_0 = \beta_1 \dots \beta_k = 0$
- $H_A = \text{some } \beta_j \neq 0$
- Name comes from distribution generated by this statistic (the F-distribution)

Individual t-tests

- Is this specific predictor related to my outcome, *controlling for all other predictors*?
- $H_0 = (\beta_1 | X_2 \cdots X_k) = 0$
- $H_A = (\beta_1 | X_2 \cdots X_k) \neq 0$

Communicating results I

```
modelsummary(list(fit, fit2),
    stars=T,
    gof_omit = "Adj.|AIC|BIC|Log|RMSE|RSE",
    coef_rename = c("EDEQ_restraint" = "Dietary Restraint Index (0-6)"))
```

	Model 1	Model 2
(Intercept)	1.104***	0.872+
	(0.154)	(0.452)
Dietary Restraint Index (0-6)	0.852***	0.842***
	(0.080)	(0.082)
BMI		0.010
		(0.018)
Num.Obs.	1085	1085
R2	0.095	0.095
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001		

Communicating results II

It can be hard to interpret for a lay reader the substantive meaning of describing a relationship, adjusting for other factors. It can be quite helpful in these instances to plot "prototypical values," allowing your reader to safely return to the land of two-dimensions.

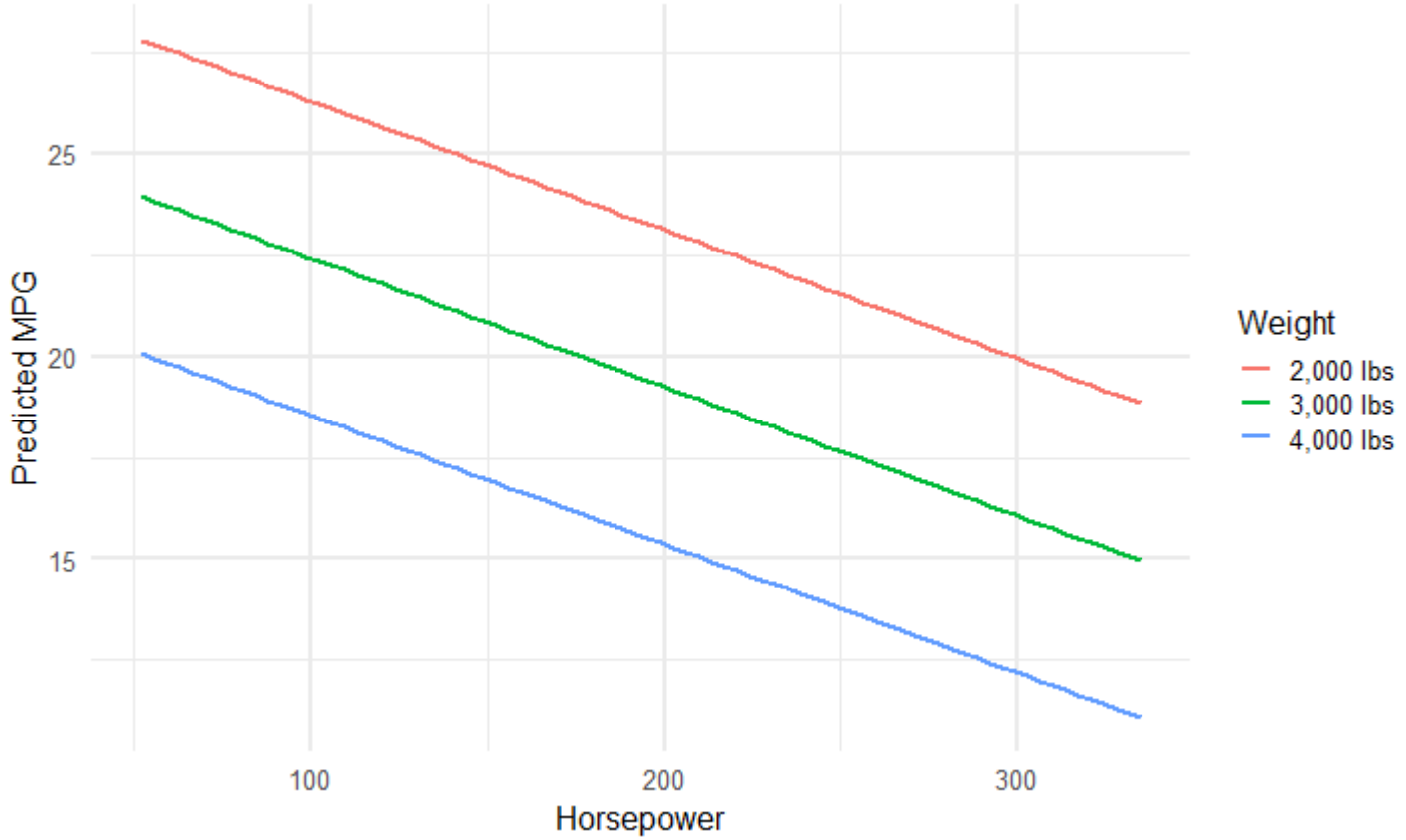
Unfortunately, in our worked example, there is no real variation in over-eating frequency by BMI, so we'll use the toy example we looked at earlier.

Communicating results II

```
# Fit your regression
car <- lm(mpg ~ hp + wt, data=mtcars)
# Use the margins package and define prototypical values
df2 <- margins::margins(car, at = list(wt = c(2,3,4)))

# Use prototypical values in resulting dataset to show results
proto <-  ggplot(data=df2, aes(x=hp, y=fitted, color=as.factor(wt))) +
   geom_smooth(method='lm', se=F) +
   xlab("Horsepower") + ylab("Predicted MPG") +
   scale_color_discrete(name = "Weight",
                        breaks=c(2,3,4),
                        labels=c("2,000 lbs","3,000 lbs","4,000 lbs")) +
   theme_minimal(base_size=16)
```

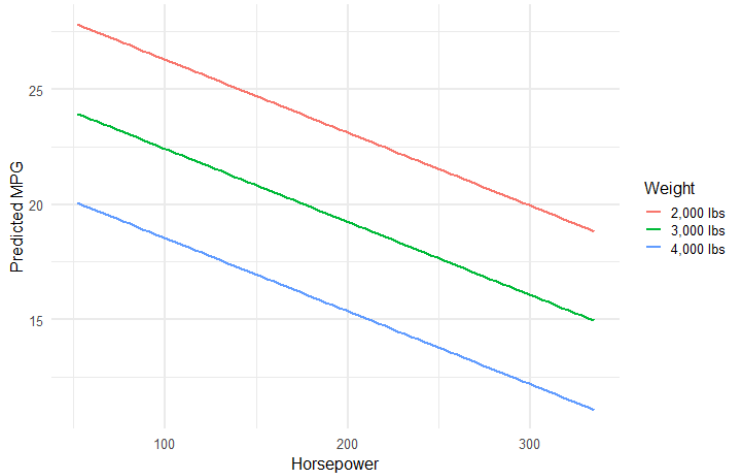
Communicating results II



Communicating results III

In communicating these results, we need to decide:

- Which predictor to display on the x-axis (generally the core question predictor)
- Which prototypical values to use that will be displayed
- Which (if any) prototypical values to use that won't be displayed



Choosing prototypical values

Examine the distribution of the other predictors and consider:

1. **Substantively interesting values**: whole numbers help
2. **A range of percentiles**: e.g., quartiles or 10th, 50th and 90th percentiles
3. **Sample mean $\pm$ 1 SD**: particularly when a symmetric distribution exists
4. **Sample mean**: typically when you are **not** displaying that particular predictor
5. **Categorical predictors**: more in Unit 4

Putting it all together

	Model 1	Model 2
(Intercept)	1.104***	0.872+
	(0.151)	(0.466)
Dietary Restraint Index (0-6)	0.852***	0.842***
	(0.112)	(0.116)
BMI		0.010
		(0.019)
Num.Obs.	1085	1085
R2	0.095	0.095
Cells report coefficients and heteroscedastic-robust standard errors in parentheses.		
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001		

Putting it all together

> We postulated a linear model which we estimated via Ordinary-Least Squares regression to assess whether there is a relationship between dietary-restraint and over-eating behaviors, on average, in the population of young adult males. At an alpha threshold of 0.05, we found that dietary-restraint behaviors were a significant predictor of over-eating patterns and accounted for approximately 10 percent of the overall variance in over-eating. We estimated that young men who are one unit apart on a dietary restraint index will have an over-eating score 0.85 (p<0.001, 95% CI: 0.74, 1.04) points different from each other (Table 1). Adjusting for individuals' body-mass index does not meaningfully alter the nature of our estimated relationship. In fact, when we hold individuals' dietary restraint behaviors constant, we fail to reject the null and conclude that there is no adjusted-relationship between BMI and over-eating, on average in the population. For parsimony, we adopt Model 1 as our preferred specification.

Bivariate v. multiple regression

	Bivariate regression	Multiple regression
Model specification	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{X}_1$	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{X}_1 + \hat{\beta}_2 \mathbf{X}_2 + \cdots \hat{\beta}_k \mathbf{X}_k$
Interpretation of $\hat{\beta}_0$	Predicted value of Y when X=0	Predicted value of Y when *all* Xs = 0
Interpretation of $\hat{\beta}_1$	Difference in Y per 1 unit of X	Difference in Y per 1 unit difference in X_1, adjusting for $X_2 \cdots X_k$
Graphical representation	Fitted line	Fitted plane in 3D (with two Xs) Plot with prototypical lines in 2D
Residuals	Distance between observation and fitted line	Distance between observation and fitted plane
Inference: t-tests	$H_0 = \beta_1 = 0$ Is there a relationship between X and Y in pop?	$H_0 = \beta_1 = 0$ Adjusting for $X_2 \cdots X_k$ is there a relationship between X_1 and Y in the population? Repeat for each X

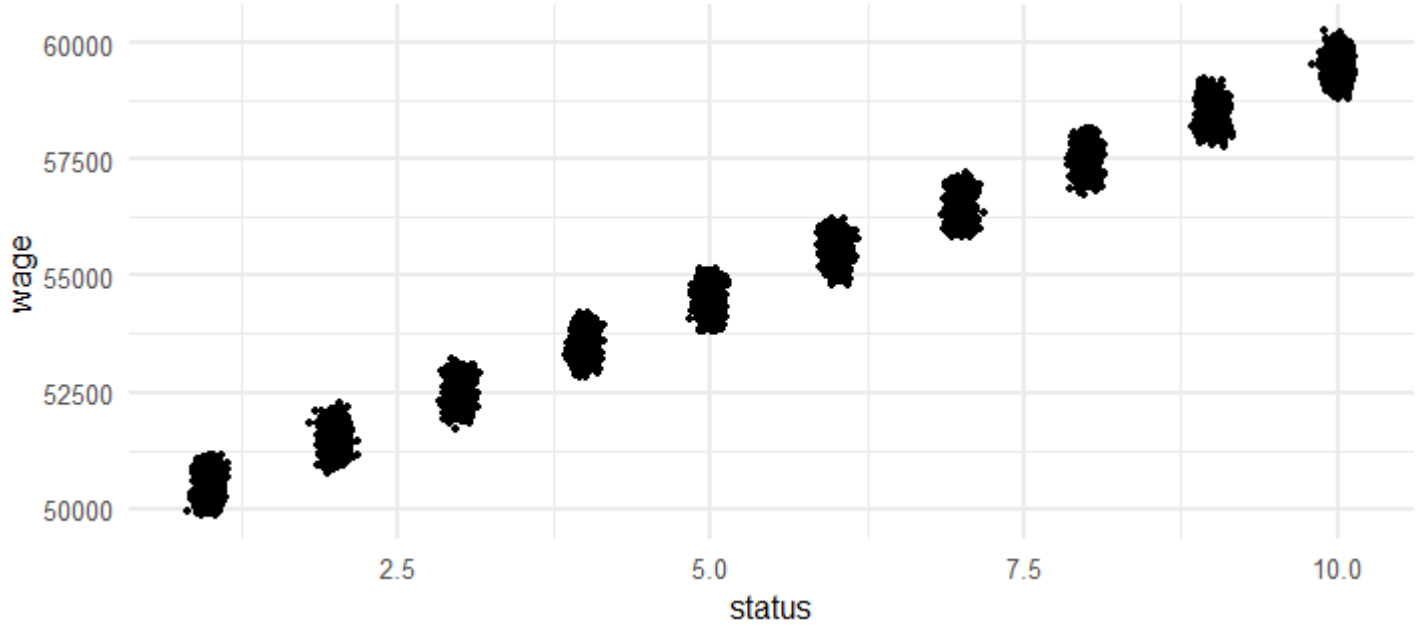
Bivariate v. multiple regression

	Bivariate regression	Multiple regression
Inference: F-test	$H_0 = \beta_1 = 0$ same result as t-test	$H_0 = \beta_1 \dots \beta_k = 0$ Does *any* predictor (or all of them jointly) have a relationship with Y in the population?
R^2	$\frac{ModelSS}{TotalSS}$ % of variation in Y explained by X	$\frac{ModelSS}{TotalSS}$ % of variation in Y explained by $X_1 \cdots X_k$
Regression assumptions	See prior unit	Same as bivariate, but *at each combination of the Xs*. Main effects assumption

Multi-collinearity

Limits of multiple regression

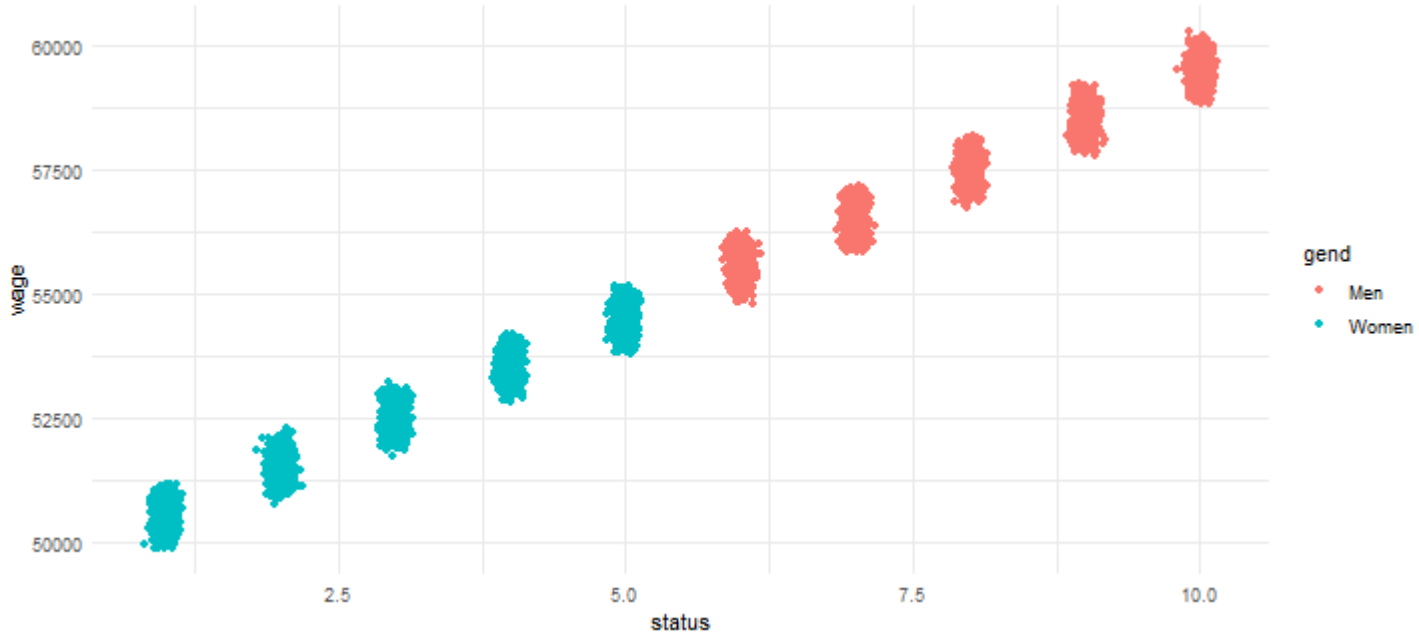
While we've noted several data puzzles multiple regression ***can*** solve, multiple regression cannot uncover the accurate nature of a relationship if predictors are "too highly" correlated. For example, if women and men have unequal access to jobs of different status, adjusting for job status will not recover the relationship between gender and wages.[1]



[1] This was a problem many researchers identified in Google's efforts to document pay disparities in 2019: (https://www.npr.org/2019/03/05/700288695/google-pay-study-finds-its-underpaying-men-for-some-jobs).

Limits of multiple regression

While we've noted several data puzzles multiple regression ***can*** solve, multiple regression cannot uncover the accurate nature of a relationship if predictors are "too highly" correlated. For example, if women and men have unequal access to jobs of different status, adjusting for job status will not recover the relationship between gender and wages.[1]



Limits of multiple regression

On average, women have lower wages than men:

```
tidy(lm(wage ~ gend, data=discr))
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   57500.      20.4     2813.       0
## 2 gendWomen     -4998.      28.9     -173.       0
```

On average, women are in lower status jobs than men:

```
tidy(lm(status ~ gend, data=discr))
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)     8.00    0.0200      400.       0
## 2 gendWomen      -5.00    0.0283     -177.       0
```

Limits of multiple regression

However, once we adjust for job status, there is no wage differential, on average in the population, between men and women:

```
tidy(lm(wage ~ gend + status, data=discr))
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept) 49517.       18.0   2747.      0
## 2 gendWomen      -8.32     12.6     -0.662   0.508
## 3 status        998.        2.19   457.      0
```

If two predictors are "too highly" correlated, we can't adjust for one to evaluate the effects of the other. This is known as multicollinearity. It can also be described as a problem of collider bias.

Multicollinearity

Multicollinearity occurs when predictor variables are highly related to each other.

This can be a simple relationship, such as when two of our predictors are strongly related to one another. This is usually straightforward to recognize, interpret, and correct for:

- Simple and adjusted slopes differ dramatically for two or more predictors
- Estimated adjusted slopes seem substantively wrong
- Standard errors *increase* with added predictors
- Reject omnibus F-test, but fail to reject individual t-tests

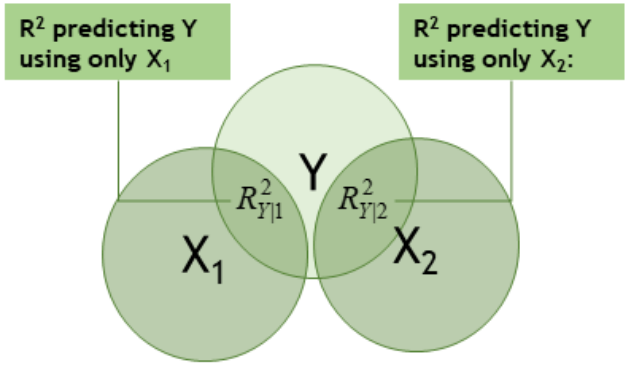
Sometimes multicollinearity is difficult to detect, such as when our variable of interest (e.g., X_1) is not strongly correlated with any one X_k, but the combination of the Xs is a strong predictor of X_1.

Multicollinearity biases our regression estimates and increases the standard errors of our regression coefficients.

Venn diagrams of collinearity

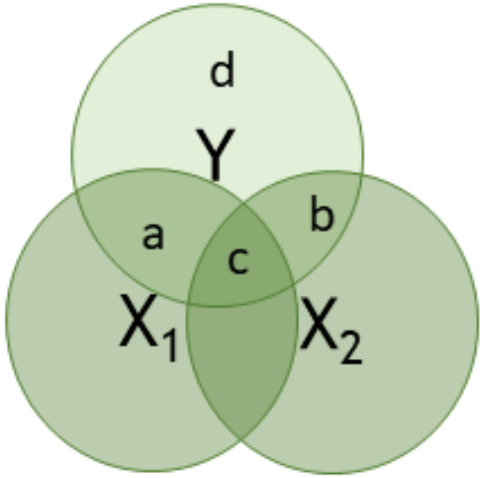
- Totally **uncorrelated predictors** are rare (almost exclusively in experiments)
- Can compute R2 just by summing separate R2s

- $R^2_{Y|1,2} = R^2_{Y|1} + R^2_{Y|2}$



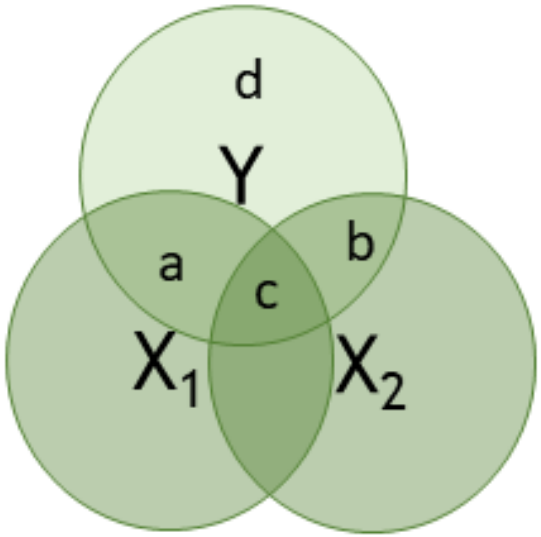
- **Correlated predictors** are very common
- Can't just sum the separate R2 because the predictors overlap

- $R^2_{Y|1,2} = \frac{a+b+c}{a+b+c+d}$



How much do extra predictors help?

- **Highly correlated predictors**: joint explained proportion of "c" is large; independent proportions of "a" and "b" are quite small; adding the other predictor will not help much
- **Fairly uncorrelated predictors**: jointly explained proportion of "c" is relatively small; independent proportions of "a" and "b" are independently or both fairly meaningful; addign predictors will help



Simple correlation: = $\sqrt{\frac{b+c}{a+b+c+d}}$

Partial correlation: = $\sqrt{\frac{b}{b+d}}$

So, the relationship between simple and partial correlations depends on the size of "a" and "c", relative to "b" and "d." As before, it's all the GLM...

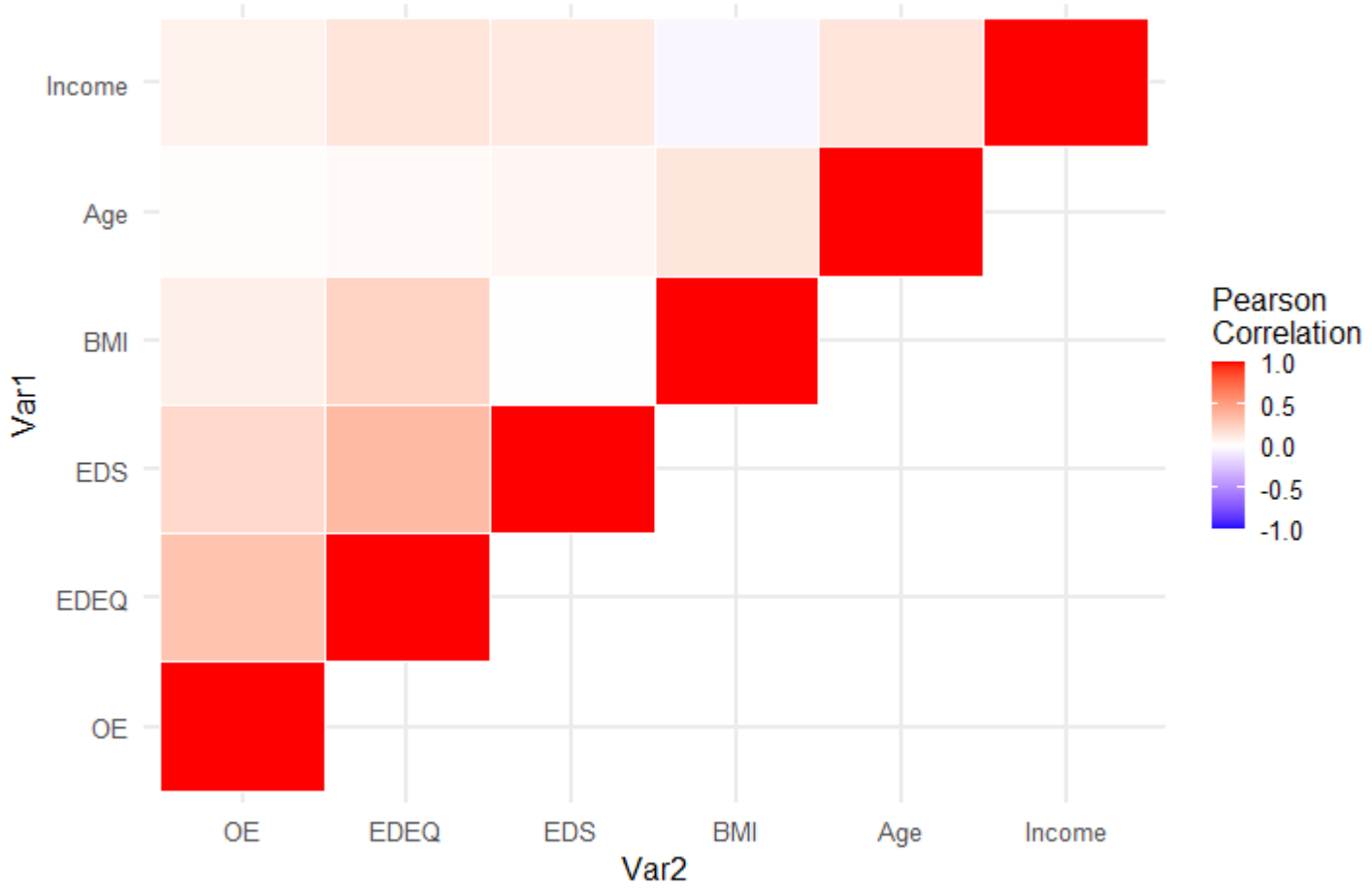
Estimates of collinearity in R

```
datasummary_correlation(cordat,
                        fmt = 3,
notes = "Notes: cells report Pearson correlation coefficients.")
```

	OE	EDEQ	EDS	BMI	Age	Income
OE	1	.	.	.	.	.
EDEQ	.308	1	.	.	.	.
EDS	.198	.358	1	.	.	.
BMI	.084	.224	-.006	1	.	.
Age	.012	.030	.047	.131	1	.
Income	.071	.144	.117	-.041	.139	1
Notes: cells report Pearson correlation coefficients.						

We can examine these correlations for the presence of multi-collinearity in our data using a correlation matrix. Do you see evidence of this here?

Visual heatmap



Correlation and collinearity

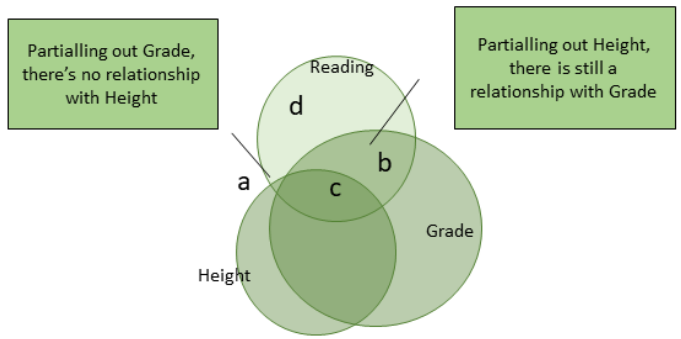
Perfect collinearity never happens (except in the instance of a duplicated variable). There are degrees of multicollinearity.

More multicollinearity = more problematic model.

In practice, when we detect problems with collinearity, what we are really detecting is strongly correlated predictors.

However, not all strongly correlated predictors are "collinear." In the example of height and grade, once we partial out grade, there is no relationship between height and reading. However, after partialling out height, there is still a relationship between grade and reading. This is because there is still variation in grade at each value of height. Don't abuse the term collinear!

	read	height	grade
read	1	.	.
height	.164	1	.
grade	.291	.492	1



Putting multicollinearity together

1. **Statistical adjustments can help recover the "true" relationship in your data that is obscured by confounding variables.**
2. **When two variables are highly correlated, it may be impossible to adjust for one**
    - This is known as the problem of **multicollinearity** (*though the term is not quite right!*)
3. **Graphical representations (such as Venn diagrams) can help you conceptualize the potential for multicollinearity**
4. **Use correlation matrices to detect for the phenomenon of highly correlated variables**
    - Consider visual representations to detect patterns more easily
5. **Solutions to multicollinearity**:
    - Increase sample size, remove a variable, create a composite or factor score (more to come in EDUC 645 and beyond!)

Synthesis and wrap-up

Goals for the unit

- Articulate the concepts of multiple regression and "statistical adjustment"
- Distinguish between the substantive implications of the terms "statistical control" and "statistical adjustment"
- Estimate the parameters of a multiple regression model
- Visually display the results of multiple regresion models
- State the main effects assumption and what the implication would be if it were violated
- Conduct statistical inference tests of single predictors (a t-test) and the full model (an F-test) in multiple regression
- Decompose the total variance into its component parts (model and residual) and use the R^2 statistic to describe this decomposition
- Describe problems for regression associated with the phenomenon of "multicollinearity"
- Use visual schema (e.g., Venn diagrams) to assess regression models for the potential of multicollinearity
- Use statistical results (e.g., correlation matrices or heat maps) to assess regression models for the potential of multicollinearity
- Describe and implement some solutions to multi-collinearity

To-Dos

Reading:

- **Finish by Feb. 2**: LSWR Chapter 15.3

Quiz 2:

- Opens 3:45 Tuesday, Jan. 31 (closes 5pm on 2/1)

Assignment 2:

- Due Feb 3., 11:59pm (**note extension!**)

Assignment 3:

- Due Feb. 10, 11:59pm

Midterm SES!!!