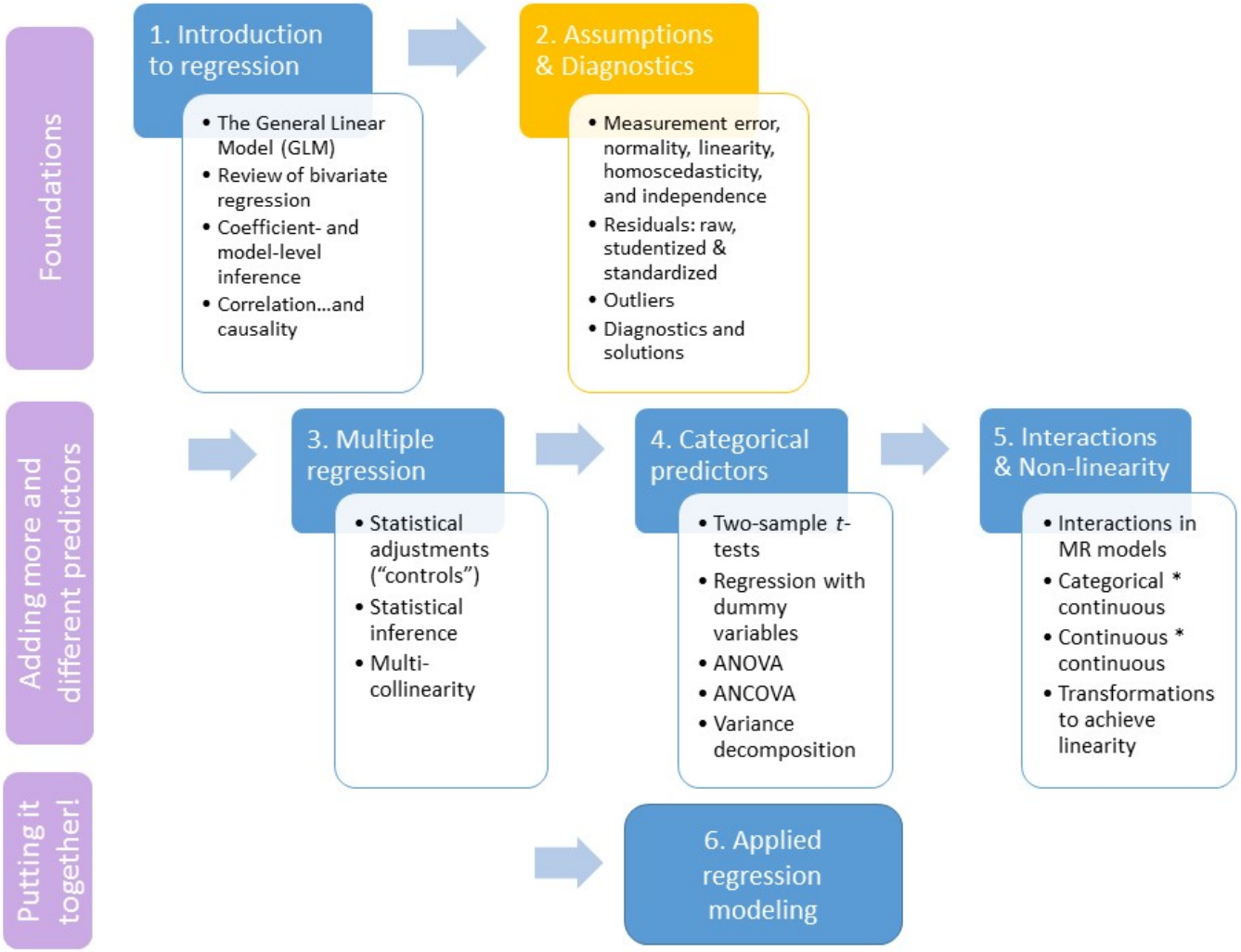
Regression assumptions and diagnostics

EDUC 643: Unit 2

David D. Liebowitz

UNIVERSITY OF OREGON

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

A motivating question

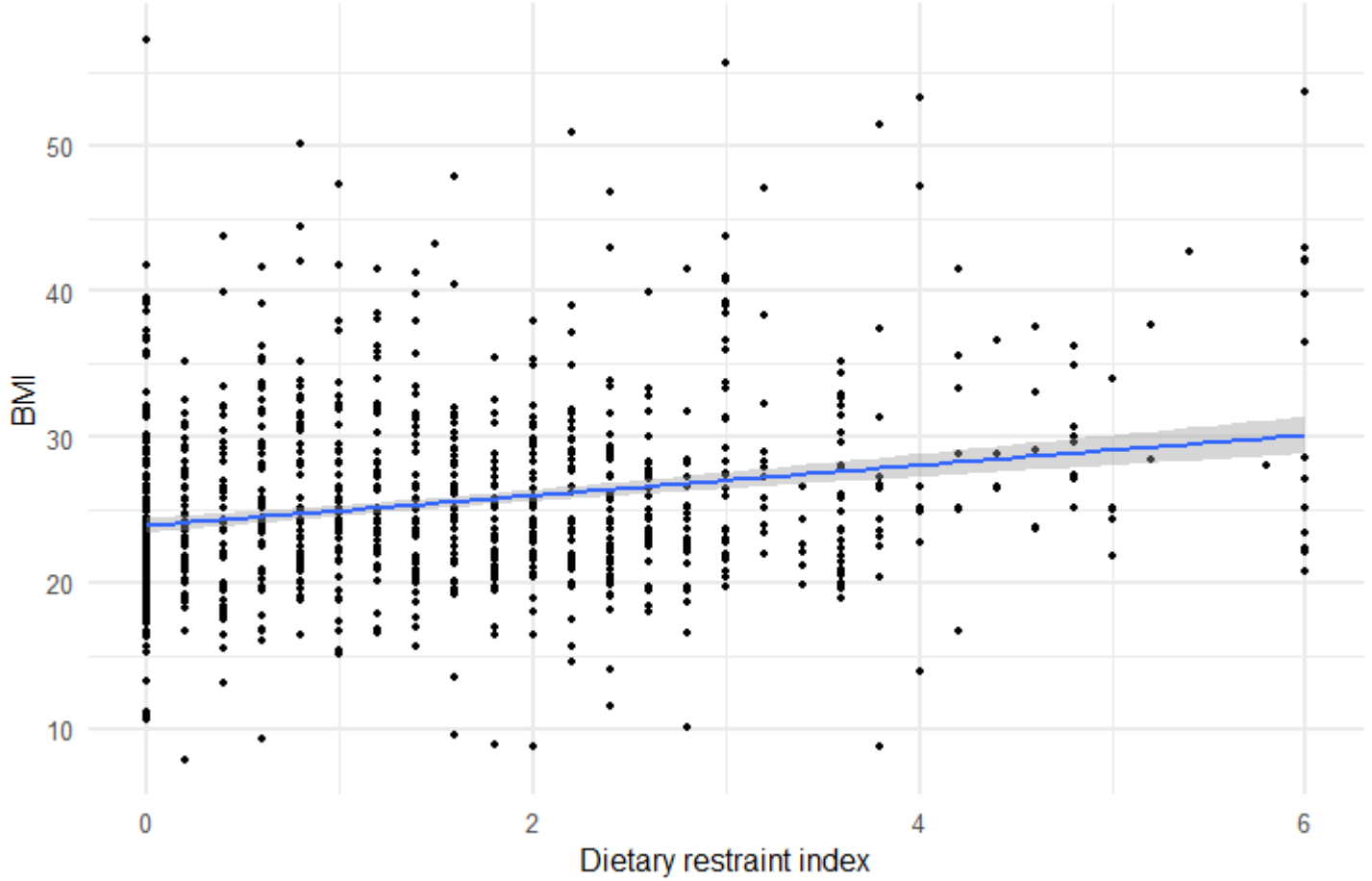
Reminder:

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.

However, before answering this question, we are exploring **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).**

Bivariate relationship



Regression results

(from last unit)

```
fit <- lm(BMI ~ EDEQ_restraint, data=do)
tidy(fit)
```

```
## # A tibble: 2 x 5
##   term           estimate std.error statistic  p.value
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       23.9      0.265     90.4  0
## 2 EDEQ_restraint     1.04     0.137      7.57 8.18e-14
```

- Each 1 unit difference in dietary restraint behaviors (*EDEQ_RESTRAINT*) is associated with a 1.04 unit difference in Body Mass Index (*BMI*).
  - This represents 1/6th of a standard deviation of the outcome (*BMI*), or a relatively large magnitude relationship
- Relationship is statistically significant
  - Unlikely that in the population of 18-30 yo males that there is no relationship between dietary restraint behaviors and BMI.
- There is still a lot of residual variation

Have we accurately characterized this relationship???

Goals of the unit

- Articulate the assumptions of the General Linear Model broadly and least squares estimation and inference particularly
- Describe sources of assumption violation in the regression model including: measurement error, non-linearity, heteroscedasticity, non-normally distributed residuals, correlated errors, and outliers.
- Articulate properties of residuals and describe their centrality in understanding the regression model assumptions
- Conduct diagnostic tests on regression model assumption violations
- Implement a consistent screening protocol to identify regression model assumption violations
- Implement solutions to regression model assumption violations, when appropriate

Regression fundamentals:

The remix

Some definitions

When you conduct any kind of statistical analysis, there are four phenomenon you tend to be interested in:

- the parameter: the underlying substantive concept you are seeking to learn about ***in the population***
  - "the effect of the program on the dropout rate", "disordered eating", etc.
- the estimand: the "true" distribution of the conceptual relationship ***in the population***; identified as a result of our research design
  - "a randomized control trial of the program", "adjusting for demographics", etc.
- the estimator: the method used to obtain an approximation of this quantity ***in the sample***
  - "Ordinary Least Squares regression", "*t*-test", "LASSO", etc.
- the estimate(s): the value obtained from conducting a particular statistical test ***in a particular sample***
  - "38.1 scale points", "5.5 percentage points", "a log-odds ratio of 2.3", etc

If we posit a linear model, we are suggesting that the estimand we are attempting to recover via our analysis will be linear in nature.

OLS is BLUE

When fitting a linear model to data, we are seeking an estimator that will return a BLUE estimate:

- Best
- Linear
- Unbiased
- Estimate

If the assumptions of linear regression are met, OLS estimators are always BLUE.

Key assumptions of regression

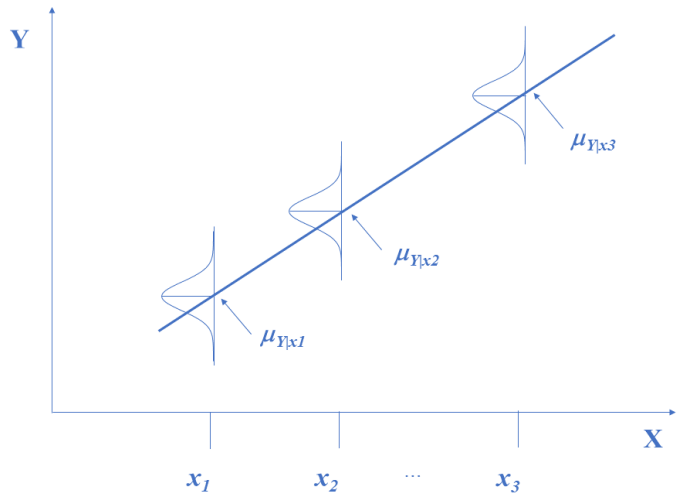
In fitting these models, we assume that the relationships are characterized by:

- No measurement error
- Linearity
- Homoscedasticity (or homoskedasticity?)
- Normally distributed residuals
- Independent errors
- No unduly influential outliers

Key assumptions of regression

At each value of X, there is a distribution of Y. These distributions have a mean $\mu_{Y|X}$ and a variance of $\sigma^2_{Y|X}$

1. **Measurement error**: Values of X are non-overlapping
2. **Linearity**: The mean of each X distribution ($\mu_{Y|X}$) can be joined together by a straight line
3. **Homoscedasticity**: Variance of each distribution ($\sigma^2_{Y|X}$) is identical
4. **Normality**: At each given value of X, the values of Y are normally distributed
5. **Independence**: Conditional on values of X, the values of Y are independent of each other



(6. **Outliers** are baked into several of above)

These assumptions are not about the sample or the original variables, but are about the **individual error terms (residuals)** ***in the population***.

Regression diagnostics: measurement error

Original estimates

Classical measurement error (our predictor variable is measured imprecisely) will bias our results to zero and increase residual standard error.

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

```
##
## Call:
## lm(formula = BMI ~ EDEQ_restraint, data = do)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.047  -3.955  -0.922   2.701  33.282
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     23.9223     0.2647  90.384  < 2e-16 ***
## EDEQ_restraint   1.0367     0.1370   7.566 8.18e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.089 on 1083 degrees of freedom
## Multiple R-squared:  0.05021,    Adjusted R-squared:  0.04933
## F-statistic: 57.25 on 1 and 1083 DF,  p-value: 8.177e-14
```

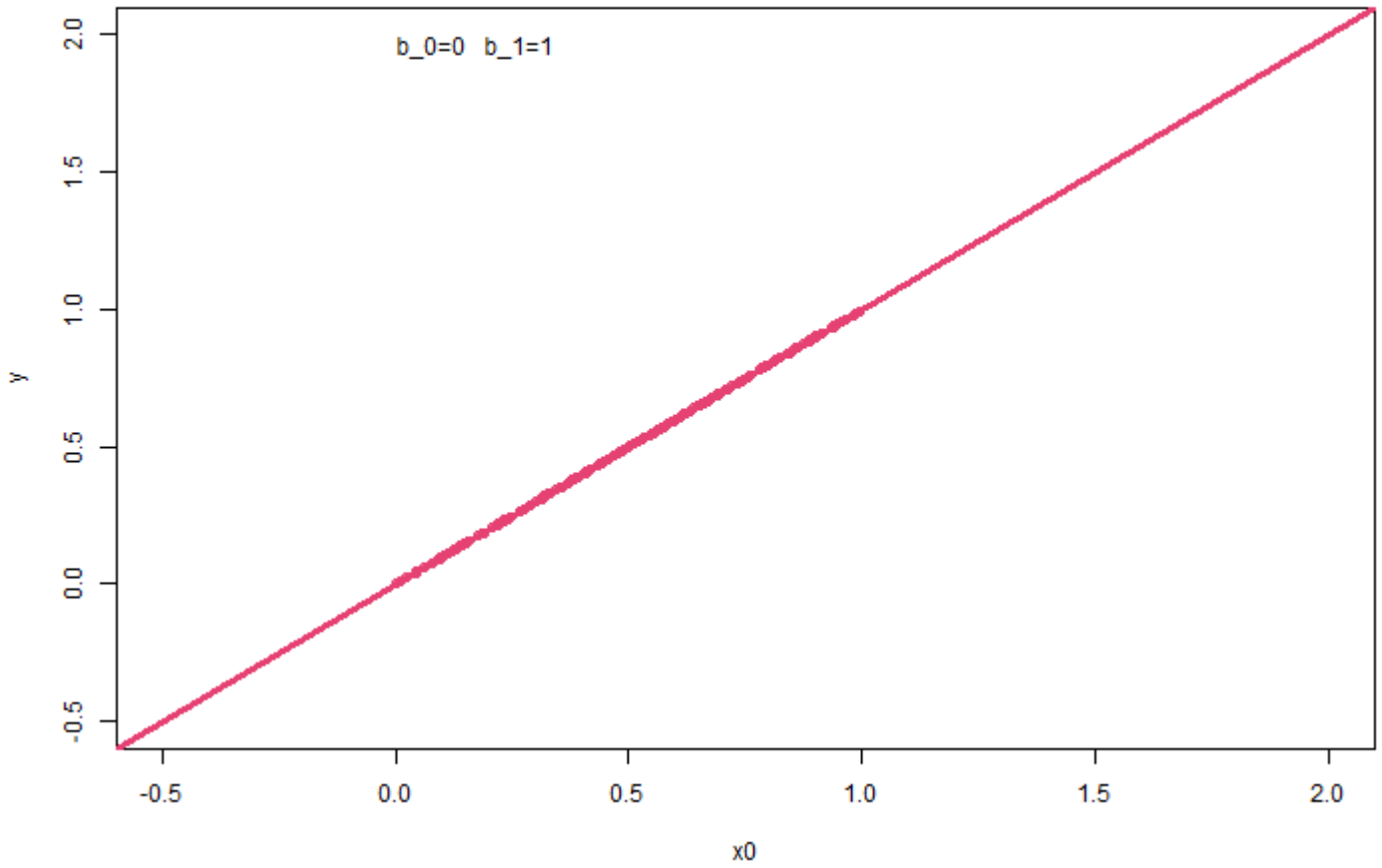
Addding measurement error

Classical measurement error (our predictor variable is measured imprecisely) will bias our results to zero and increase residual standard error.

```
do$EDEQ_noise <- jitter(do$EDEQ_restraint, factor = 5, amount = 1)
summary(lm(BMI ~ EDEQ_noise, data=do))
```

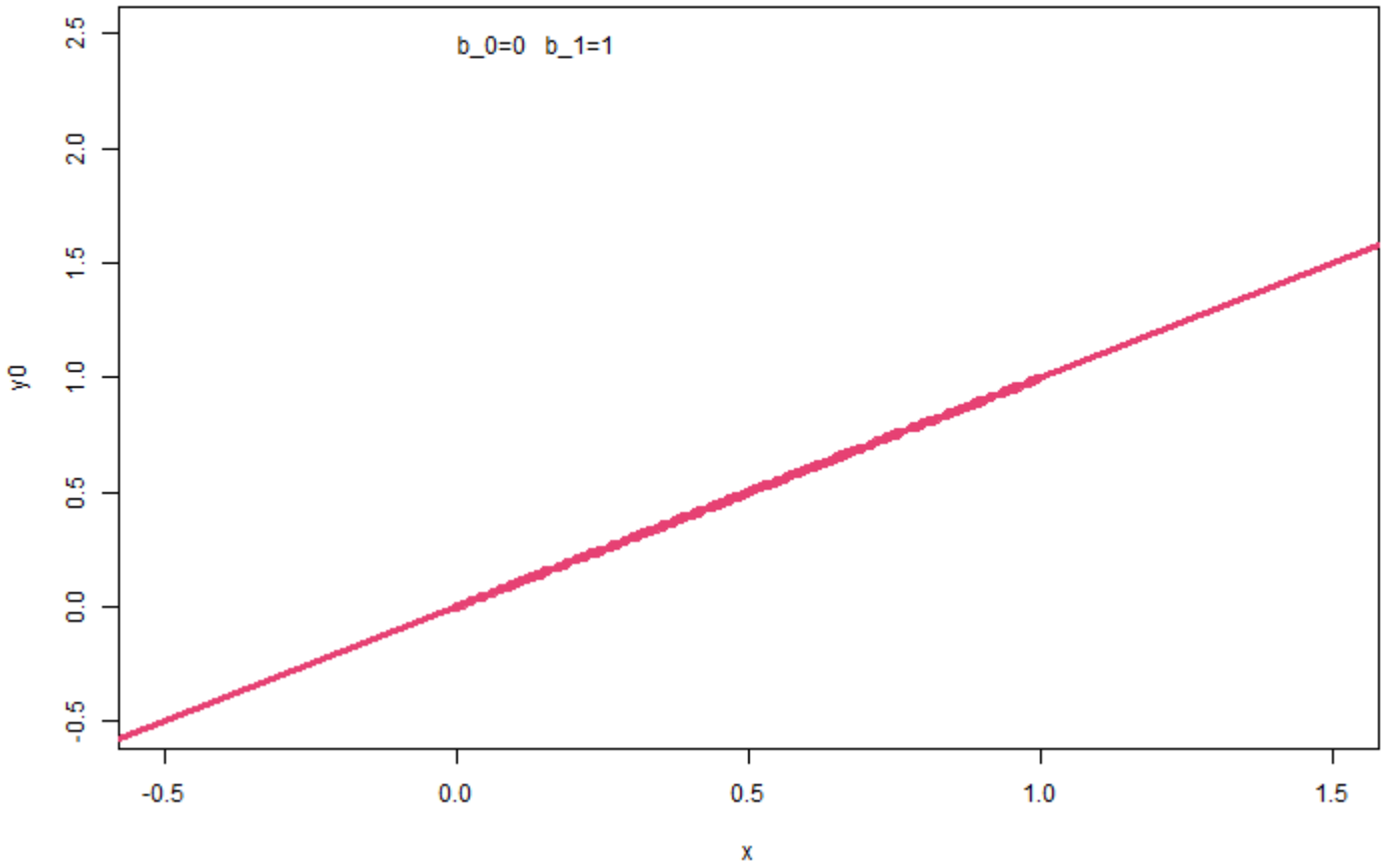
```
##
## Call:
## lm(formula = BMI ~ EDEQ_noise, data = do)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.713  -3.904  -0.916   2.831  33.246
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.1216     0.2541  94.938  < 2e-16 ***
## EDEQ_noise    0.8966     0.1262   7.103 2.21e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.107 on 1083 degrees of freedom
## Multiple R-squared:  0.04451,    Adjusted R-squared:  0.04363
```

Measurement error in x



h/t Simon Hess (@simonhhess)

Measurement error in y



h/t Simon Hess (@simonhhess)

Addressing measurement error

- Test for presence of measurement error with reliability statistics
- Solve by:
  - Creating factors of underlying constructs
  - Adjusting standard errors post-hoc
  - SEM
  - Get more (better) data
  - Improve research design

We'll learn about reliability in EDUC 645 as well as in more advanced courses such as EDLD 667

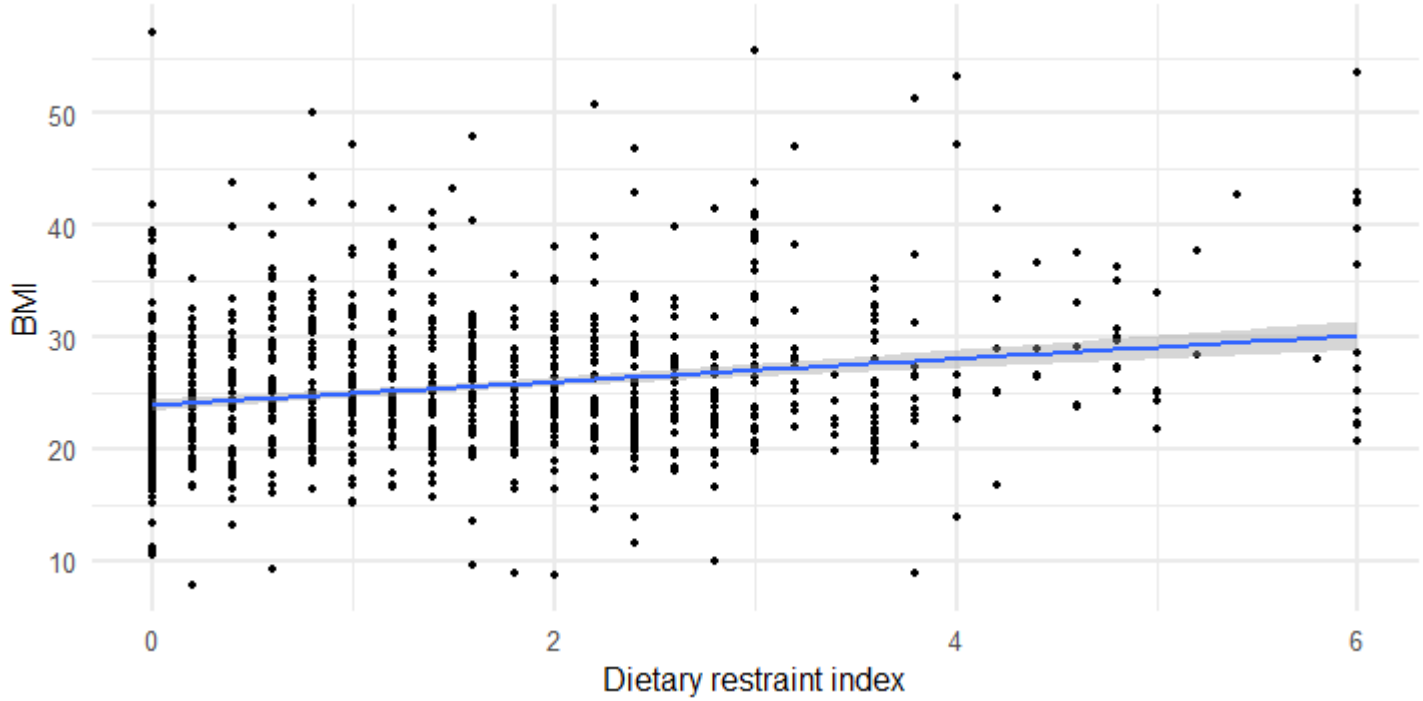
We will also learn about some solutions in 633/634 (SEM) and EDLD 650 and 679 (causal inference)

These are generally problems that are solved by design rather than stats

Regression diagnostics:

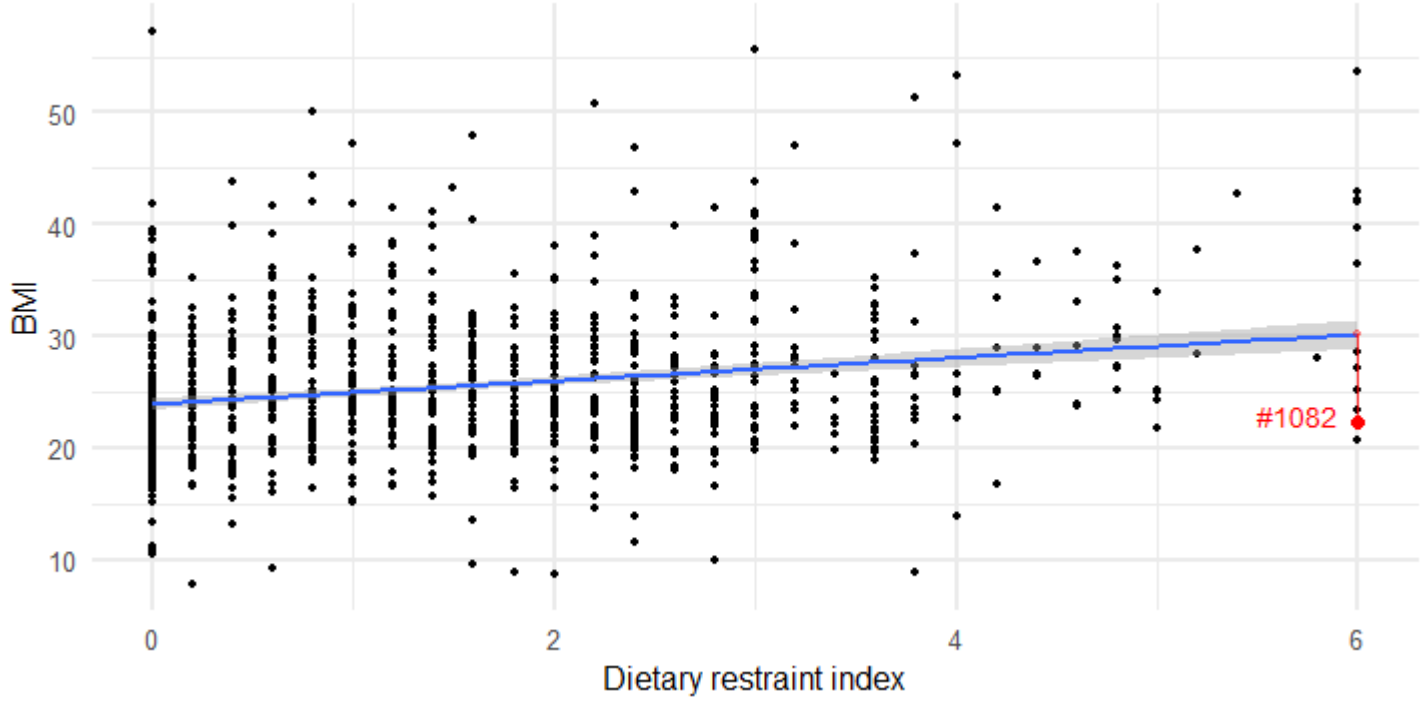
Residuals as tools to assess linearity, homoscedasticity, normality and independence

Residuals



Our fitted regression line contains the "predicted" values of *BMI* for each value of *DIETARY_RESTRAINT*. But almost all of the "actual" values of *BMI* lie off the actual line regression line.

Example: Participant #1082



What can we say about this study participant's BMI, relative to our prediction?

What is a "residual"?

The difference ("vertical distance") between the observed value of the outcome and its predicted value is called the residual.

Residuals can be substantively and statistically useful:

- Represent individual deviations from average trend
- Tell us about values of the outcome after taking into account ("adjusting for") the predictor
    - In this case, tell us whether study participants have higher or lower BMI indices than predicted, given their dietary restraint behaviors

Raw and studentized residuals

Raw residuals give us information on how far each observed value is from its predicted value *in the original units of the outcome*. It may be valuable to quantify residuals in more standard units.

- Raw residuals: Observed minus fitted values:

$$r_i = y_i - \hat{y}_i$$

- Standardized residuals: Can transform raw residuals into standardized units by dividing by the Mean Square Error (estimate of standard deviation of the residuals).[1]

$$stdr_i = \frac{r_i}{\sqrt{MSE}}$$

- Studentized residuals: Raw residuals come from models that may be influenced by individual data points (esp. in small samples). To avoid this problem, it is helpful to estimate residuals that come from a model based on all data *except the case at hand (i)*. Here, MSE_i is based on a regression fit without observation i.

$$stur_i = \frac{r_i}{\sqrt{MSE_{-(i)}}}$$

[1] Technically, this is an oversimplification as we usually calculate standardized and studentized residuals by also accounting for the "leverage" of observation *i* (how much does it influence the regression statistics).

Residuals as tools

Residuals can be valuable diagnostic tools to assessing the assumptions of the regression model. In addition to serving as mechanisms to test the overall model, they can point to specific observations that may warrant further inspection. This might include:

- Noticing patterns to change your theory about the correct model or functional form to use
- Inspecting particular cases to determine whether they are mis-measured or belong to a different population than the main one of your study
- Improve your research design overall

Our regression assumptions of homscedasticity, normality and independence are all about the residuals. We often say as shorthand for this that the residuals are independently and identically distributed:

$$\varepsilon_i = \text{i.i.d. } N(0, \sigma^2)$$

Recovering residuals in R

```
fit <- lm(BMI ~ EDEQ_restraint, data=do)

# predict asks for the predicted values
do$predict <- predict(fit)

# residuals asks for the raw residual
do$resid <- residuals(fit)

# rstandard asks for the standardized residual
do$std_resid <- rstandard(fit)

# rstudent asks for the studentized residual
do$stu_resid <- rstudent(fit)
```

We can now treat these residual and predicted values as new variables in our dataset and examine them using all the other univariate and multivariate analysis tools we have.

Examining the residuals

```
summary(do$resid)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -19.047  -3.955  -0.922   0.000   2.701  33.282
```

```
sd(do$resid)
```

```
## [1] 6.086189
```

- Sample mean of the residuals is *always* exactly zero
- On a scale ranging from about 5 to 60, most of our residuals are within 6 BMI units. However, we've done a poor job of predicting BMI for some participants.

Examining the residuals

- Compare to the raw residuals on the previous slide to the standardized residuals. *How similar are they?*[1]

```
summary(do$std_resid)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -3.134177 -0.650214 -0.151496  0.000021  0.443938  5.471042
```

```
sd(do$std_resid)
```

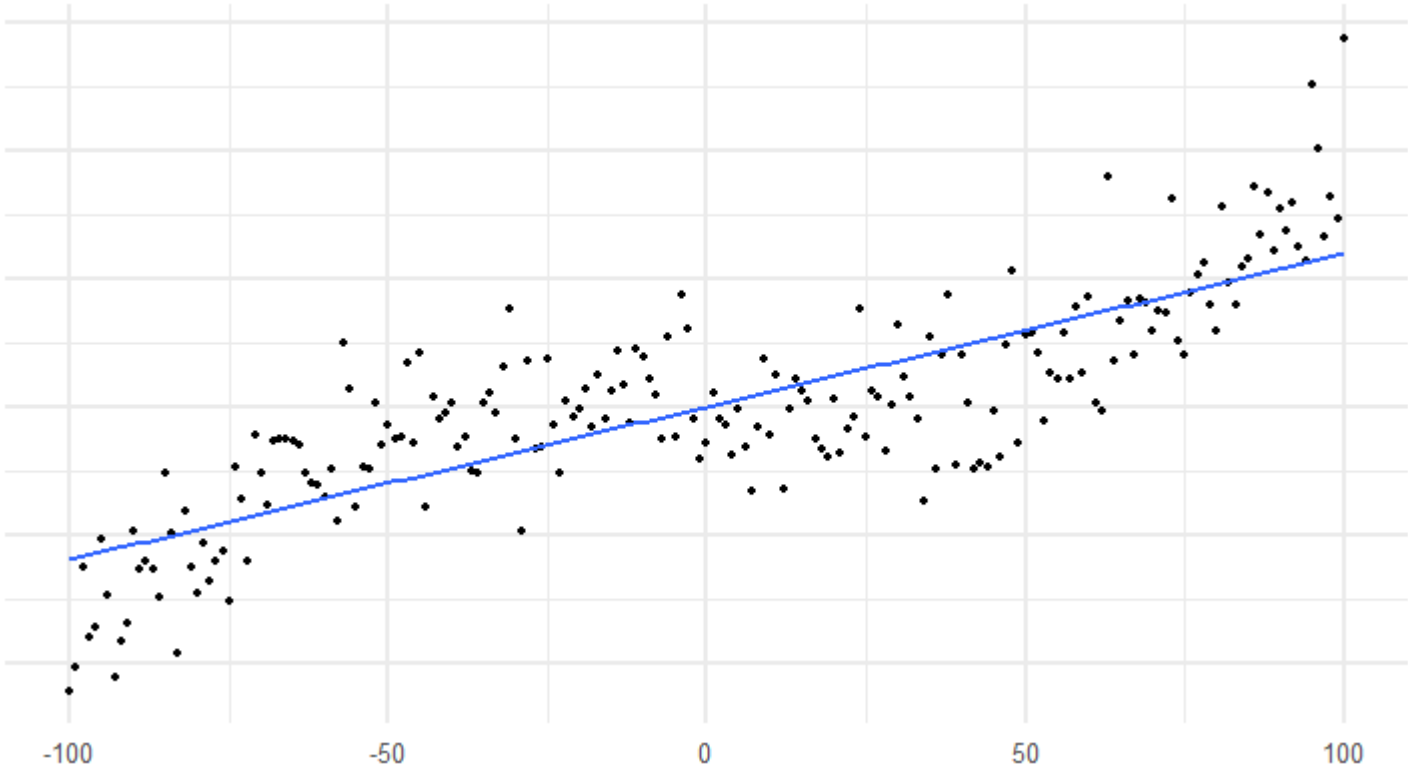
```
## [1] 1.000632
```

- Note the minima and maxima standardized residual values. *Given a sample of ~1100 individuals, what does this make you think?*

[1] Our standardized and studentized residuals have means *close to* zero, but not quite because of "leverage" corrections (see slide 22)

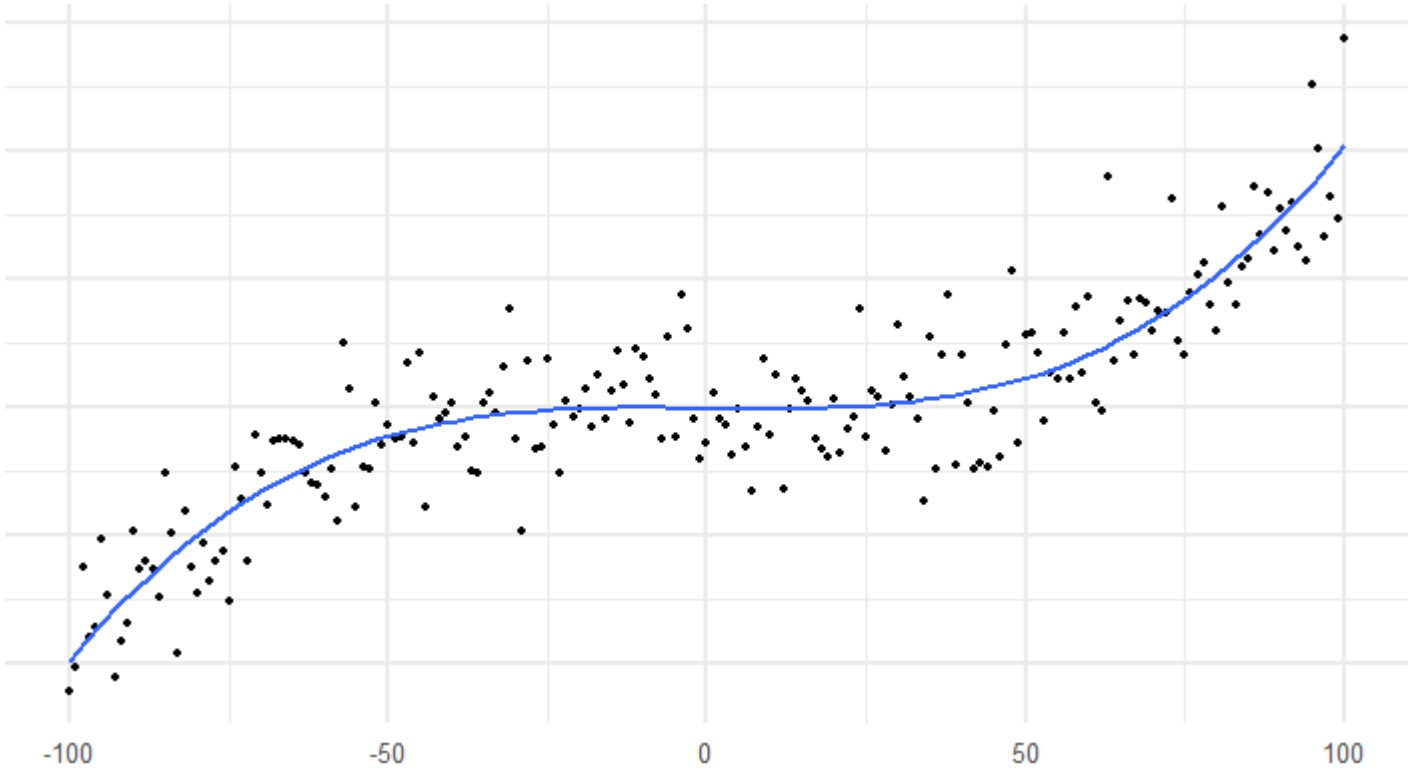
Non-linearity

In addition to examining the original scatterplot to detect the presence of non-linearity...



Non-linearity

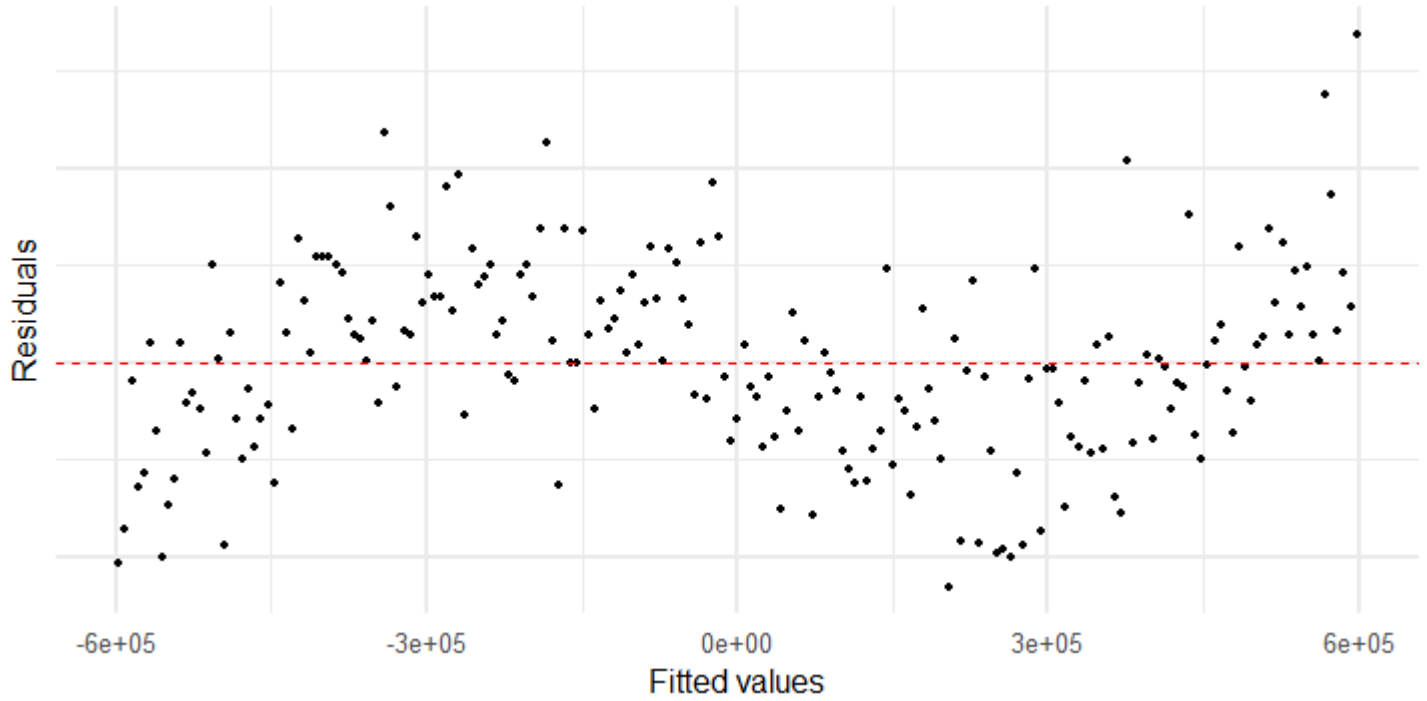
In addition to examining the original scatterplot to detect the presence of non-linearity...



We can also examine the residuals to detect non-linearity!

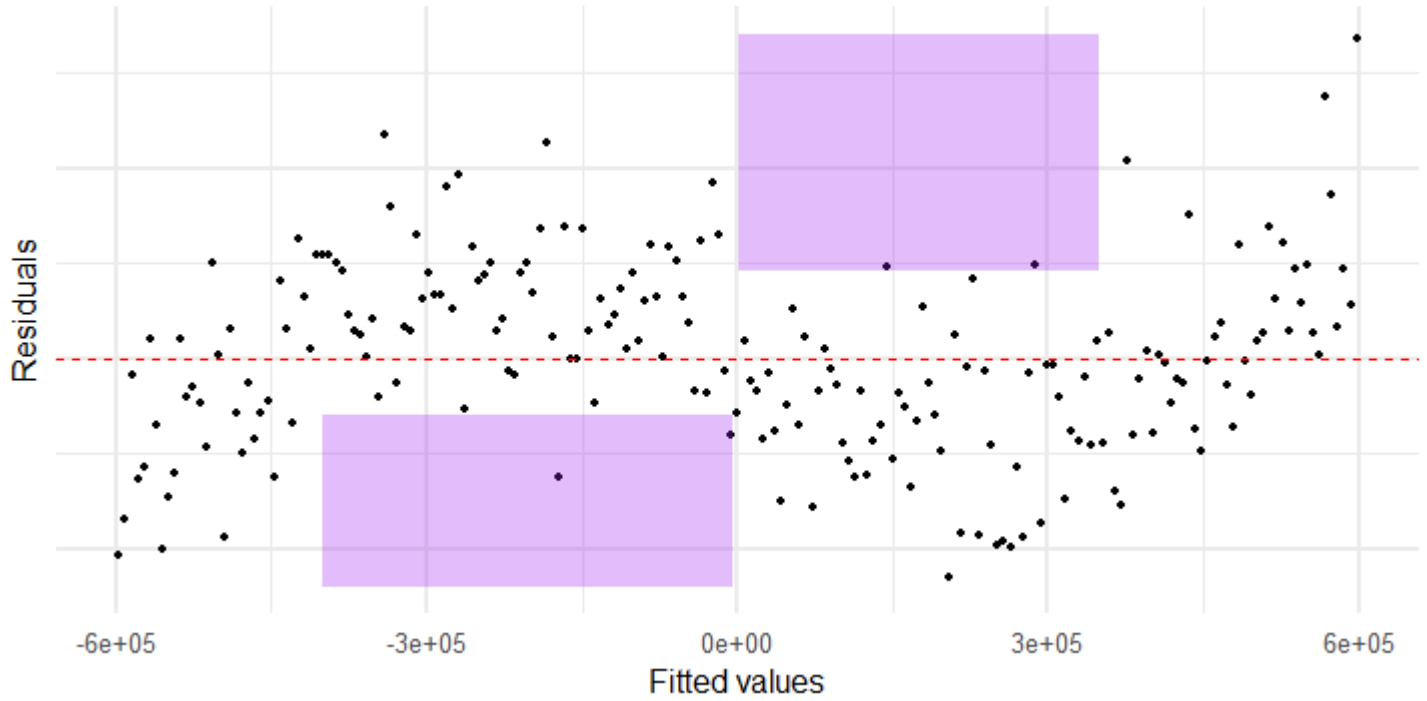
Non-linearity: residuals v. fitted plot

Plotting the residual values against their fitted values can provide informative insights to many of our regression assumptions.



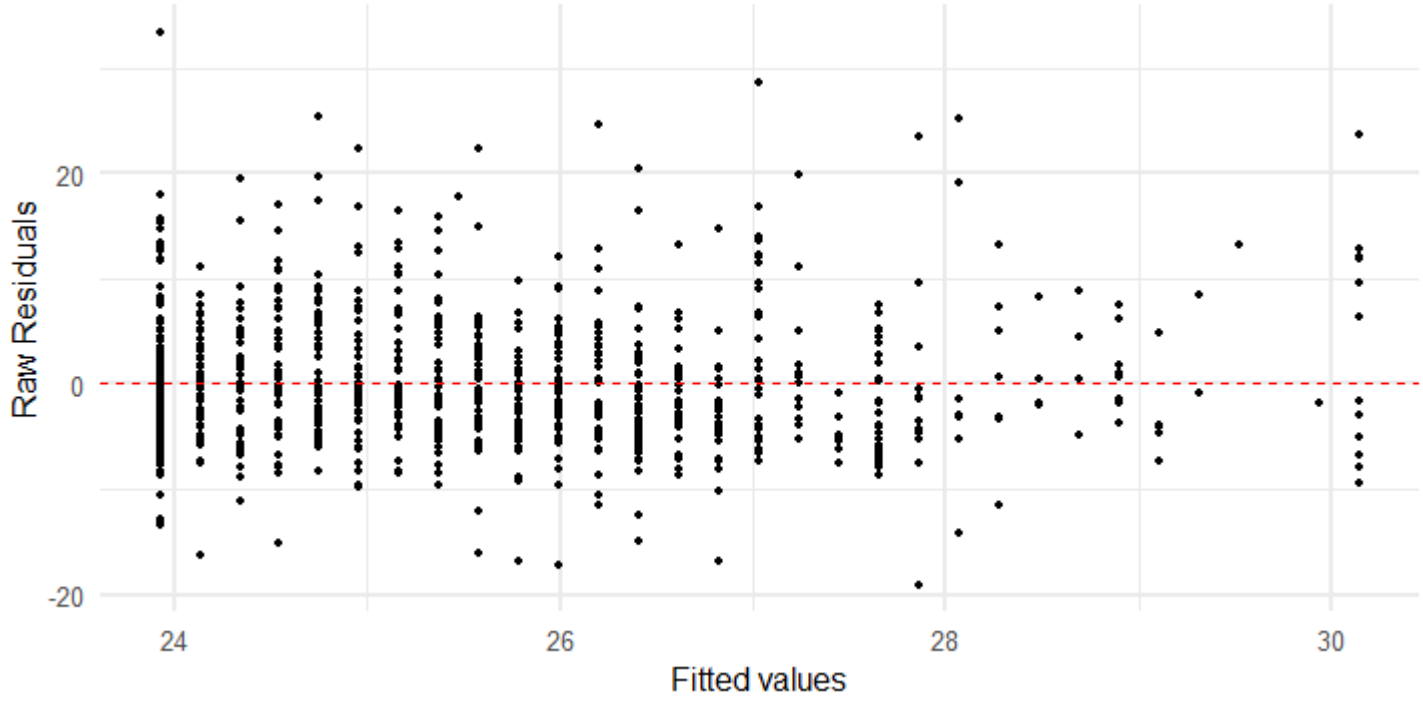
Non-linearity: residuals v. fitted plot

Plotting the residual values against their fitted values can provide informative insights to many of our regression assumptions.



Non-linearity: residuals v. fitted plot

Let's look at this for our estimates of disordered eating:

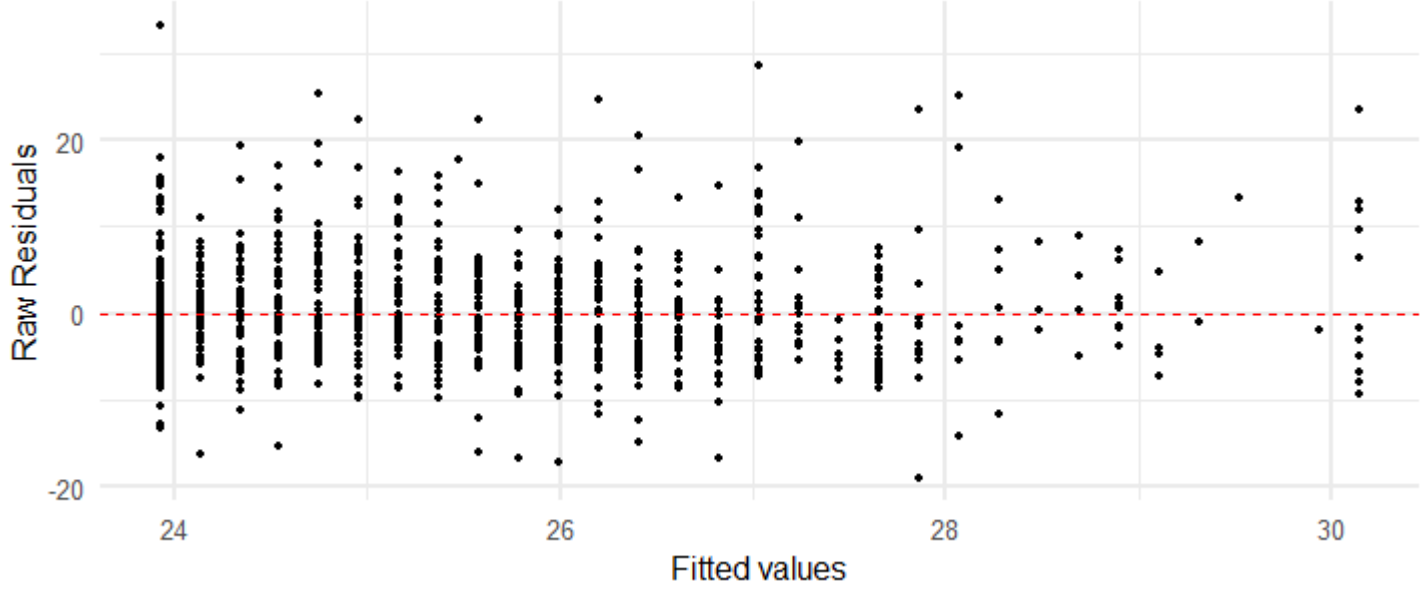


What evidence do you observe of linearity or non-linearity in the residuals for our disordered eating sample?

Non-linearity: residuals v. fitted plot

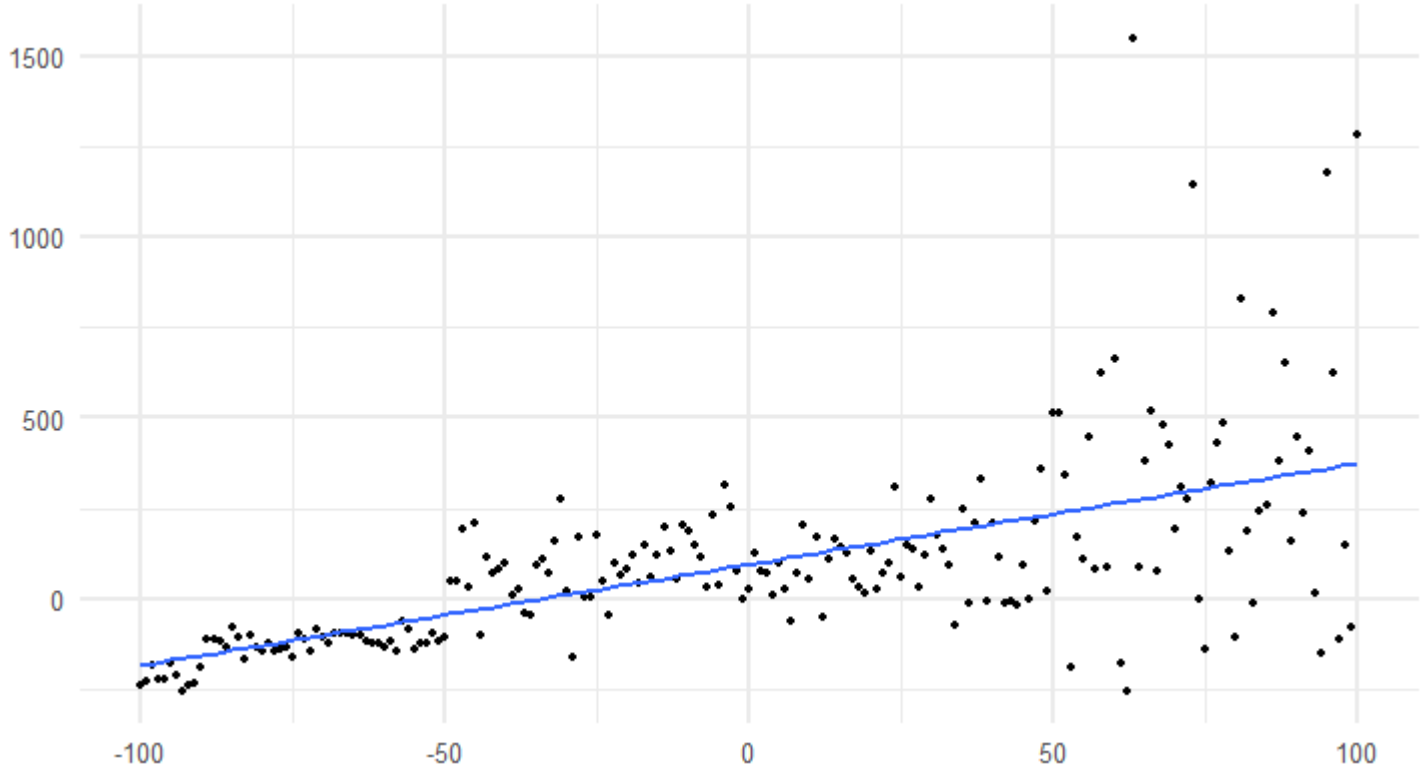
Let's look at this for our estimates of disordered eating:

```
ggplot(do, aes(x = predict, y = resid)) +
         geom_point() +
  geom_hline(yintercept = 0, color = "red", linetype="dashed") +
  ylab("Raw Residuals") + xlab("Fitted values") +
  theme_minimal(base_size = 16)
```



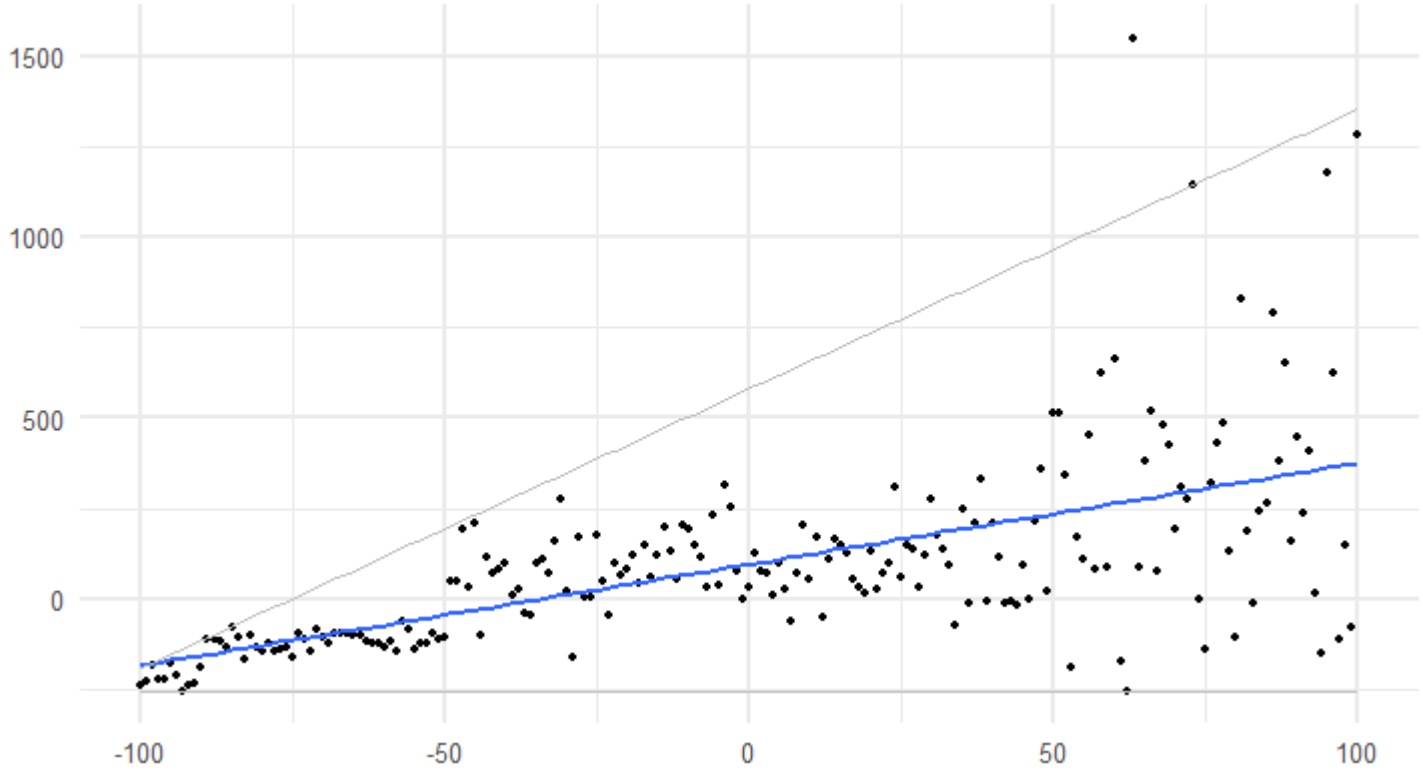
Heteroscedasticity

Heteroscedasticity: when variance of Y conditional on X ($\sigma^2_{Y|X}$) differs as a function of X

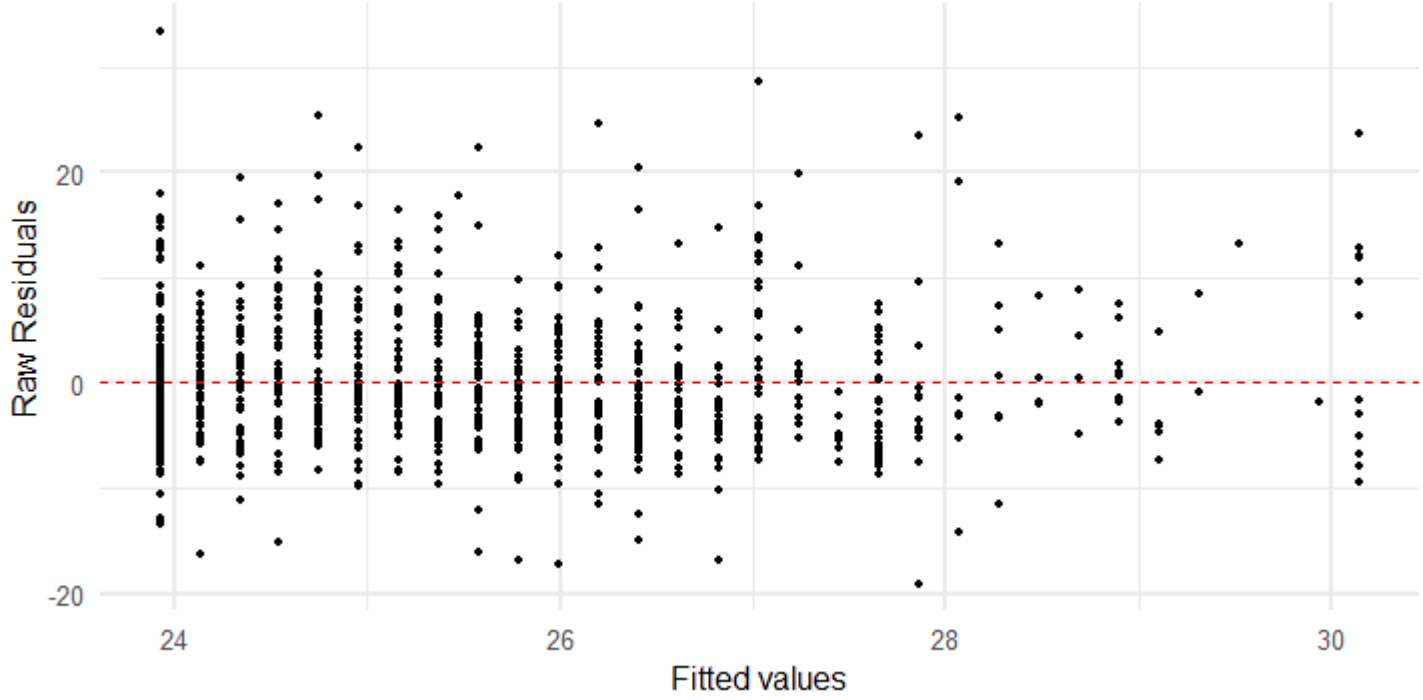


Heteroscedasticity

Heteroscedasticity: when variance of Y conditional on X ($\sigma^2_{Y|X}$) differs as a function of X



Residual heterodscedasticity



What evidence do you observe of homoscedasticity or heteroscedastictiy in the residuals for our disordered eating sample?

Solutions to heteroscedasticity

Heteroscedasticity biases our standard errors. The standard deviation of the estimator $\hat{\beta}_1$ is inconsistent for the true value of $\sigma^2_{\hat{\beta}_1}$ when there is heteroscedasticity.

If this were the case, the sample-based statistics do not follow a standard normal distribution (i.e., the CLT). This would invalidate our inferencing.

However, we generally can correct for the presence of heteroscedasticity by the computation of heteroscedasticity-robust standard errors. There are several approaches that rely on the variance-covariance matrix of the errors and a good deal of matrix algebra.

Solutions to heteroscedasticity

We won't spend time on these now, but know that if you are concerned about the presence of heteroscedasticity in your data, you will generally want to use heteroscedasticity- robust standard errors (sometimes called sandwich estimators or Eicker-Huber-White standard errors). Some R options include `estimatr::lm_robust` and `modelsummary`.

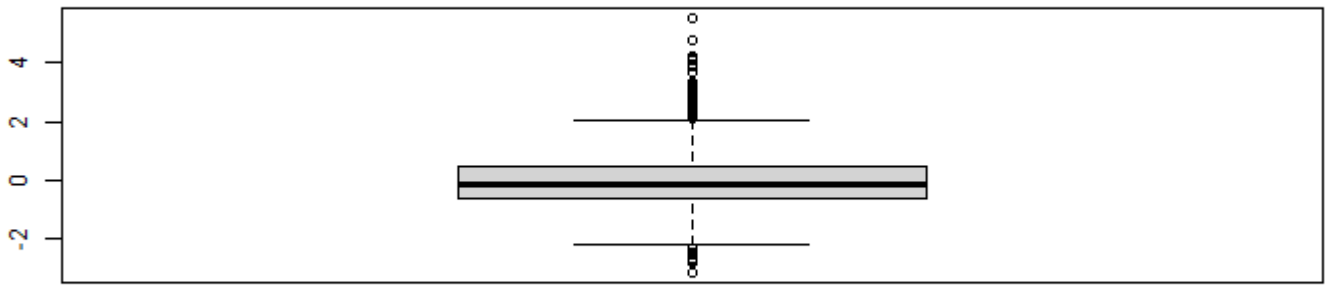
```
modelsummary(fit,
  stars=T,
  gof_omit = "Adj.|AIC|BIC|Log|RMSE",
  coef_rename = c("EDEQ_restraint" = "Dietary Restraint Index (0-6)"),
  vcov =  list("iid", "robust"))
```

	Model 1	Model 2
(Intercept)	23.922***	23.922***
	(0.265)	(0.258)
Dietary Restraint Index (0-6)	1.037***	1.037***
	(0.137)	(0.161)
Num.Obs.	1085	1085
R2	0.050	0.050

Residual normality: boxplot

For the standard errors (and associated inference tests) that we conduct in the regression analysis to be correct, the residuals **must be normally distributed.** Here it makes most sense to use our *studentized residuals*. Why?

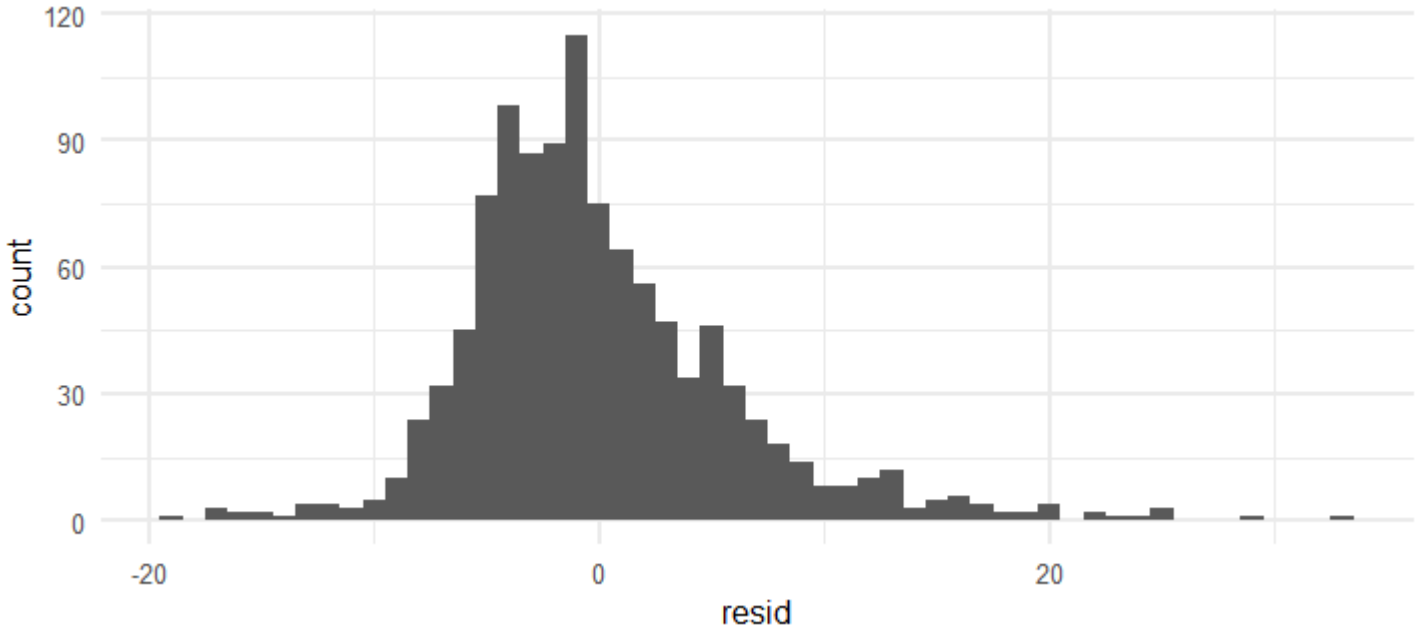
```
boxplot(rstudent(fit))
```



A few outliers, particularly in the positive direction... (means we are *under*-predicting the BMI of these folks)

Residual normality: histogram

For the standard errors (and associated inference tests) that we conduct in the regression analysis to be correct, the residuals **must be normally distributed**

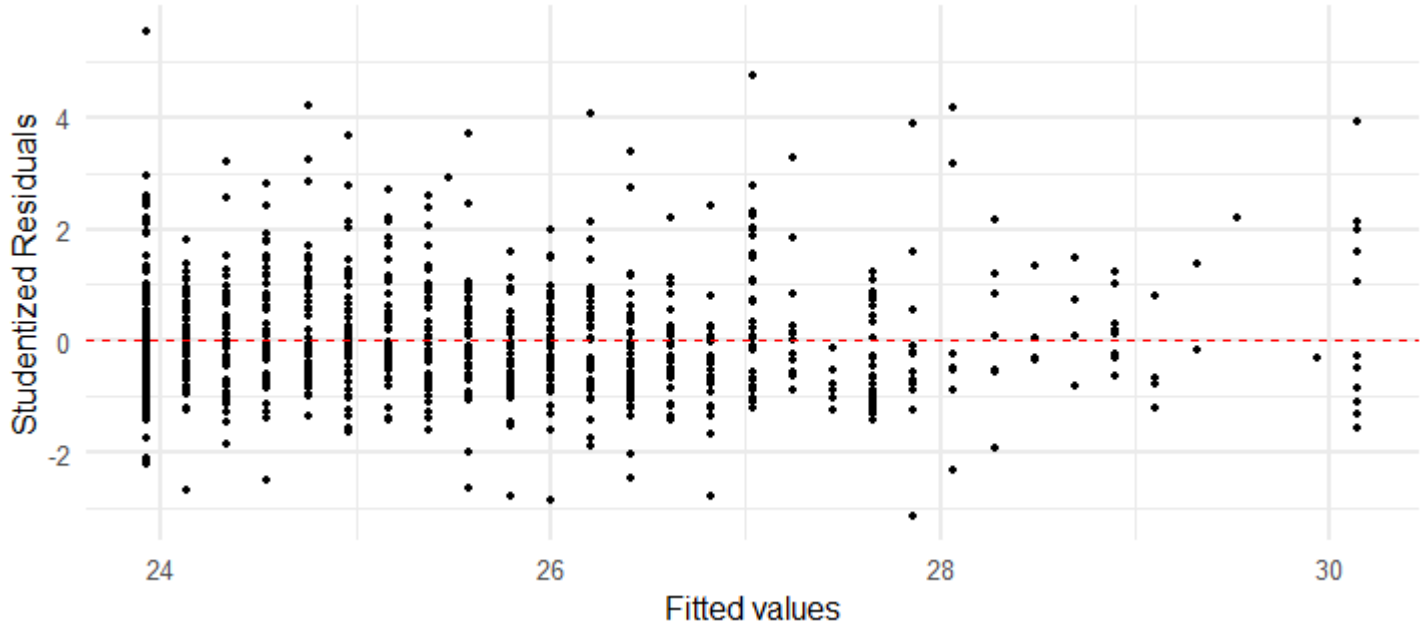


And this is confirmed in this histogram. What sort of skew is this distribution exhibiting?

All in all, though, these seem roughly normally distributed.

Residual normality: resid v. fitted

For the standard errors (and associated inference tests) that we conduct in the regression analysis to be correct, the residuals **must be normally distributed**

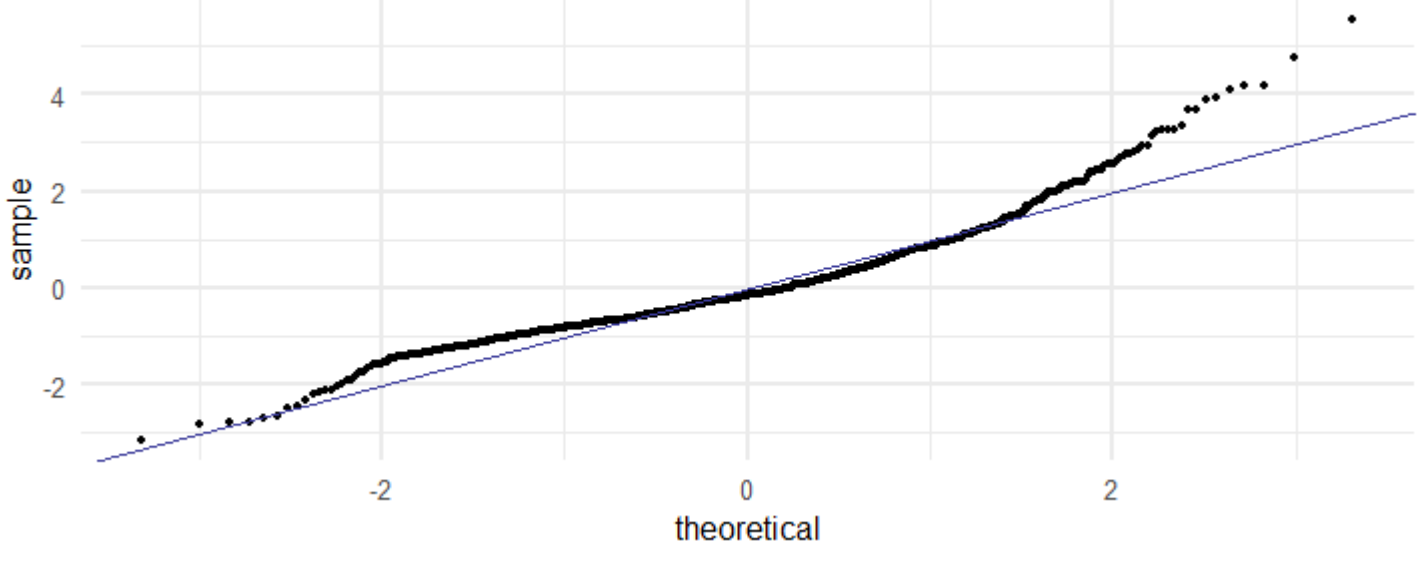


What would you be looking for here to assess for **normality**?

Residual normality: Q-Q plot

For the standard errors (and associated inference tests) that we conduct in the regression analysis to be correct, the residuals **must be normally distributed**. A quantile-quantile (Q-Q) plot compares where observed values in the sample--at specified percentiles of the distribution--fall in relation to where those same percentile values would fall in a normal distribution. A straight line indicates what a precisely normal distribution would look like.

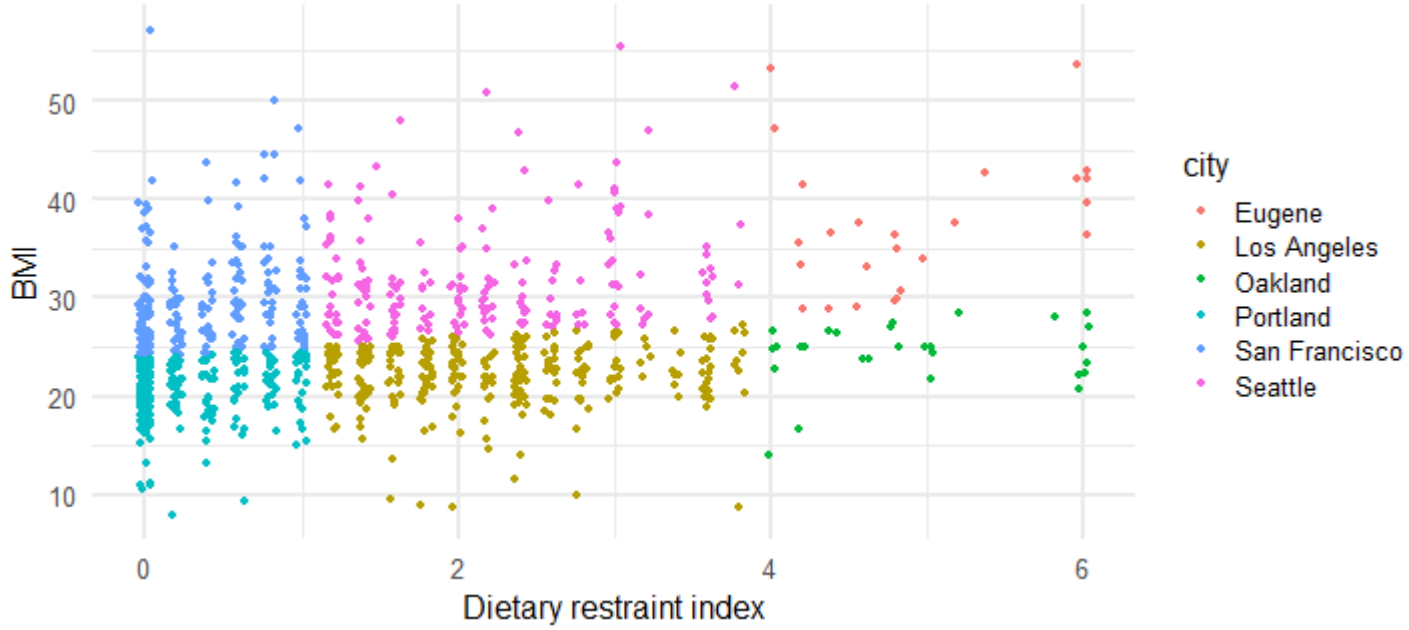
```
qq <- ggplot(do) +
        stat_qq(aes(sample=stu_resid)) +
        geom_abline(color=blue)
```



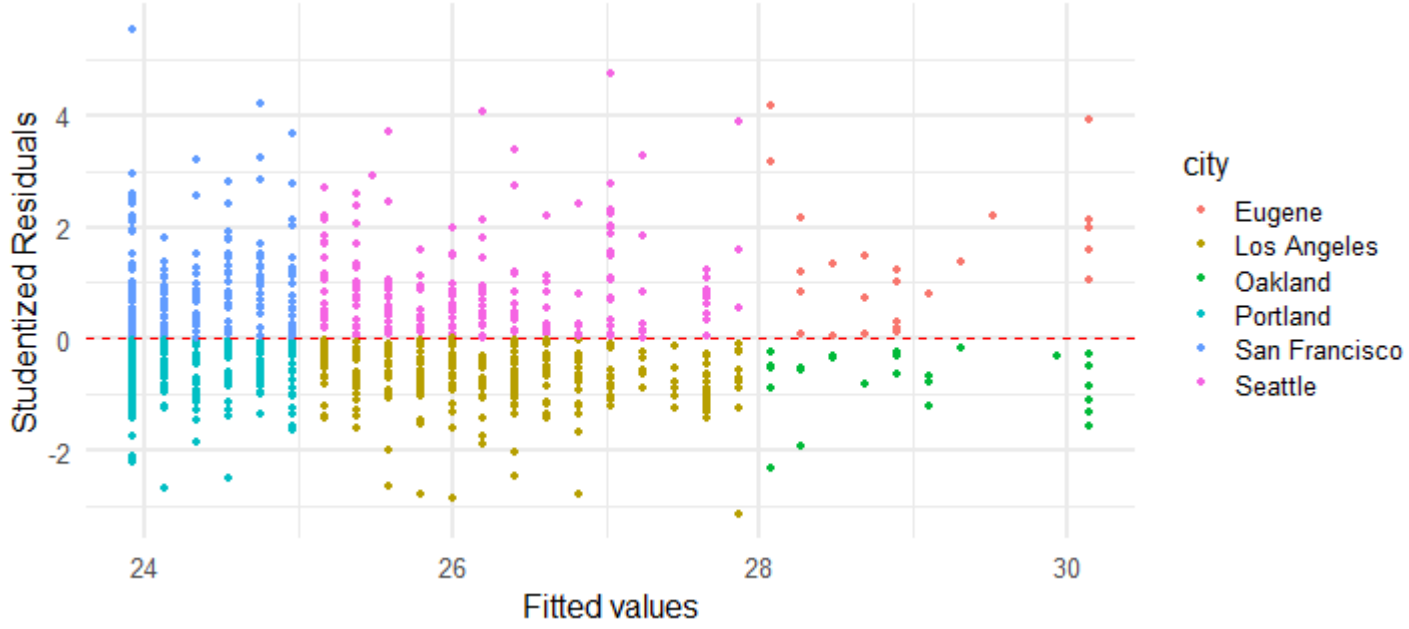
Residual independence

If observations (and residuals) are correlated (i.e., there is some feature in our data that makes some observations have correlated outcomes), then our model is mis-specified and we will have biased standard errors.

Imagine that most of the young men in our disordered eating sample lived in one of six cities.



Residual independence



It might seem like we have 1083 degrees of freedom, but actually there is far less independent variation in our data because one's membership in a particular group (in this case city) informs (some) of one's outcome value. Thus, our naïve inference will be incorrect.

Depending on disciplinary perspective, we can address these concerns through multi-level modeling or standard error adjustment. More on this in EDUC 645 and EDLD 628/629 (HLM)

Solutions to residual independence

We won't spend time here, but if you are concerned about the presence of clusters in your data--*because you are worried it will affect your inference*--consider using cluster-robust standard errors, which inflate standard errors for the k^{th} regressor by τ_k , where $\tau_k = 1 + \rho_{kx} + \rho_{\mu} * (\text{avg. cluster size} - 1)$.

```
modelsummary(lm_city,
  stars=T,
  coef_omit = "city",
  gof_omit = "Adj.|AIC|BIC|Log|RMSE",
  coef_rename = c("EDEQ_restraint" = "Dietary Restraint Index (0-6)"),
  vcov =  list("iid", ~ city))
```

	Model 1	Model 2
(Intercept)	32.476***	32.476***
	(1.316)	(2.269)
Dietary Restraint Index (0-6)	1.026***	1.026*
	(0.212)	(0.461)
Num.Obs.	1085	1085

Outliers

Broadly defined as highly atypical and/or influential data point(s).

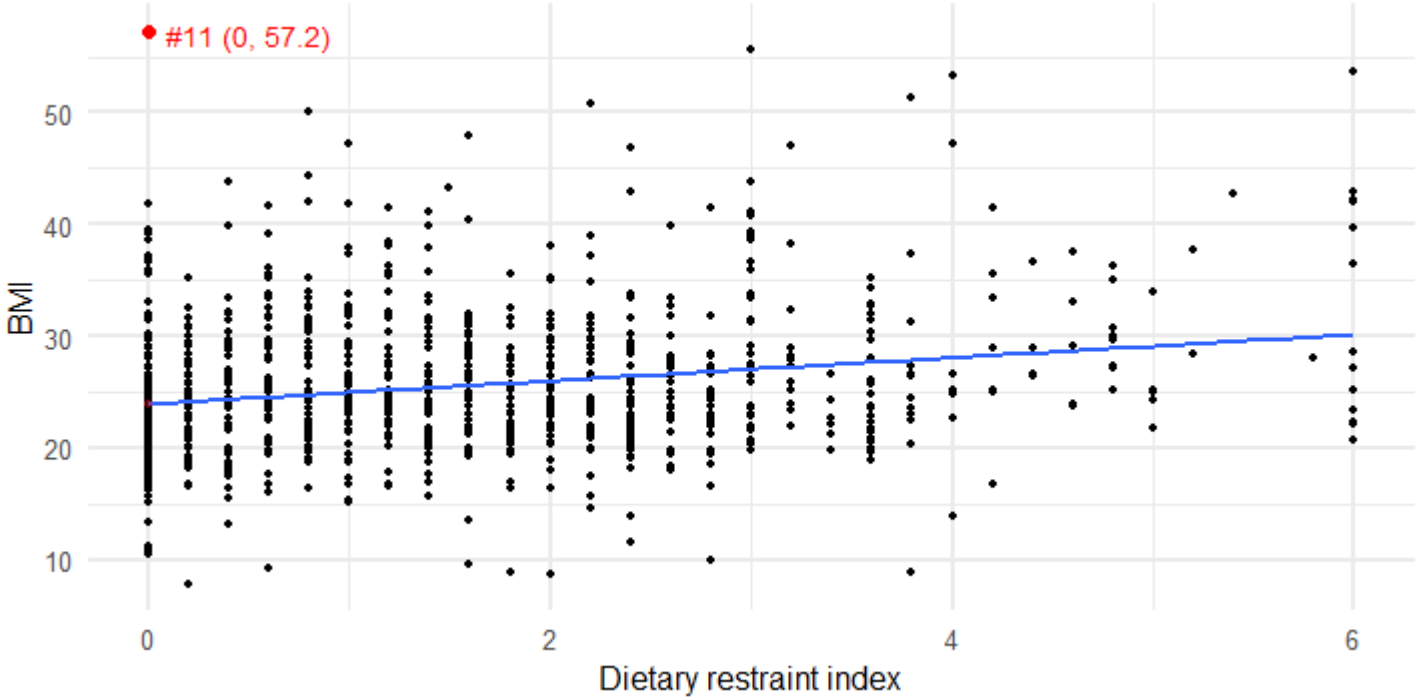
These could be a result of:

- Coding (recording) error
- Accurate observations of a rare case(s)
- Observation of an individual from a different population

We can characterize outliers on three dimensions:

1. **Leverage**
    - How unusual is the case in the 'X' direction?
2. **Distance**
    - How unusual is the case in the 'Y' direction?
3. **Influence**
    - How much does the relationship (regression coefficient) change in the absence of the case?
    - Influence is roughly Leverage * Distance

Outliers



How would you characterize participant #11 on leverage and distance?

Outliers

```
do_out <- filter(do, id!=11)
tidy(lm(BMI ~ EDEQ_restraint, data=do))
```

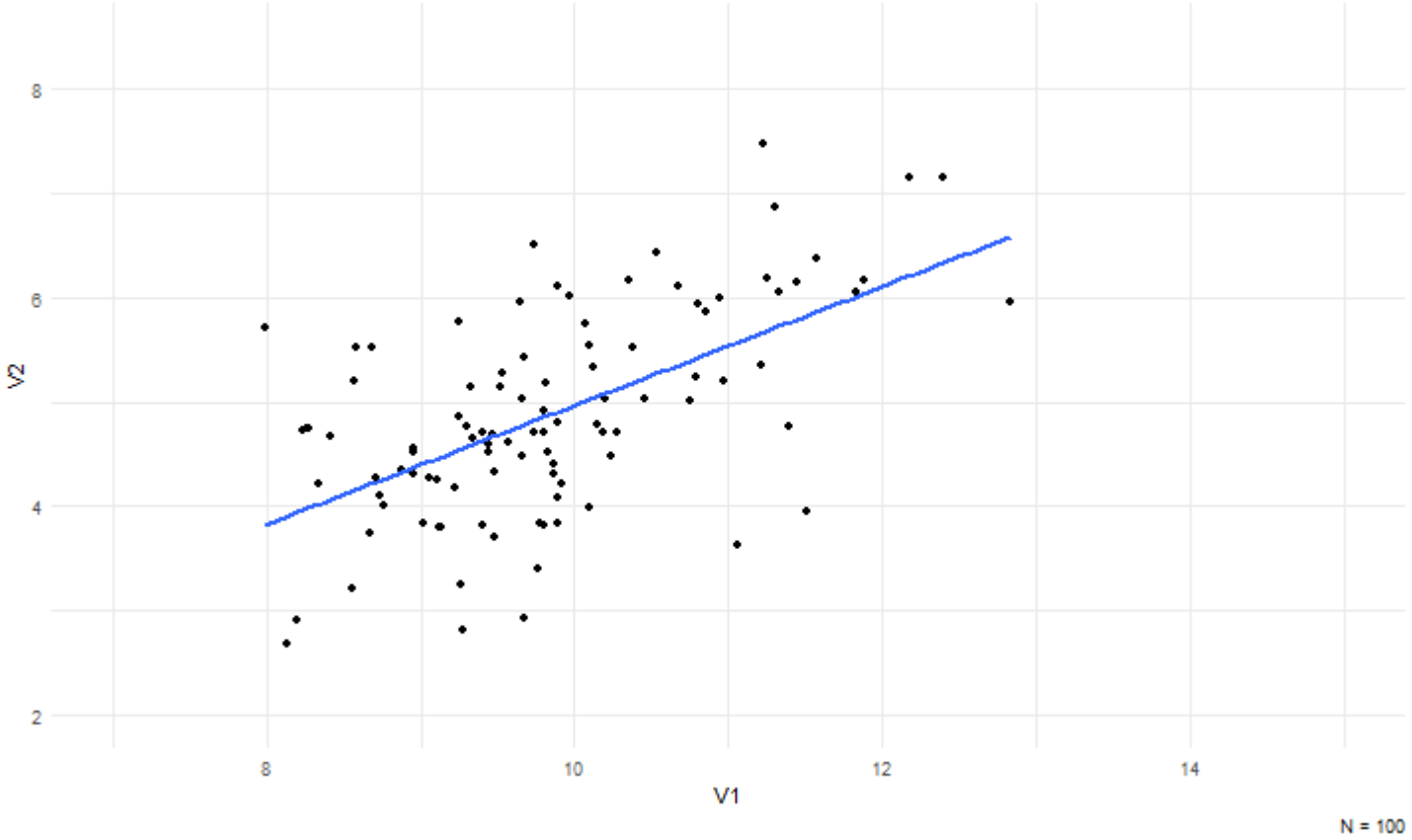
```
## # A tibble: 2 x 5
##   term           estimate std.error statistic  p.value
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       23.9      0.265     90.4  0
## 2 EDEQ_restraint     1.04     0.137      7.57 8.18e-14
```

```
tidy(lm(BMI ~ EDEQ_restraint, data=do_out))
```

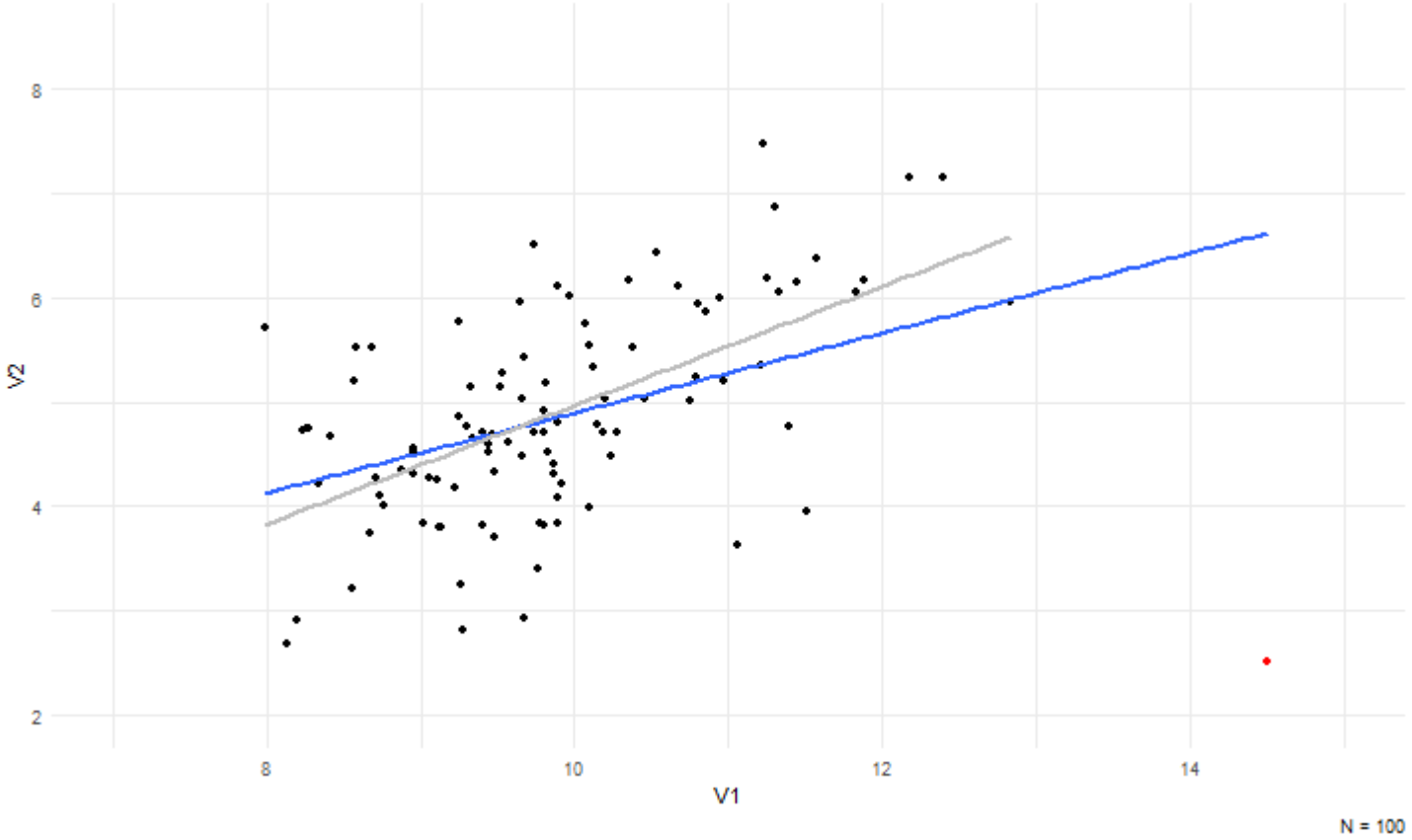
```
## # A tibble: 2 x 5
##   term           estimate std.error statistic  p.value
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       23.9      0.261     91.3  0
## 2 EDEQ_restraint     1.06     0.135      7.84 1.09e-14
```

Things don't change much in this case!

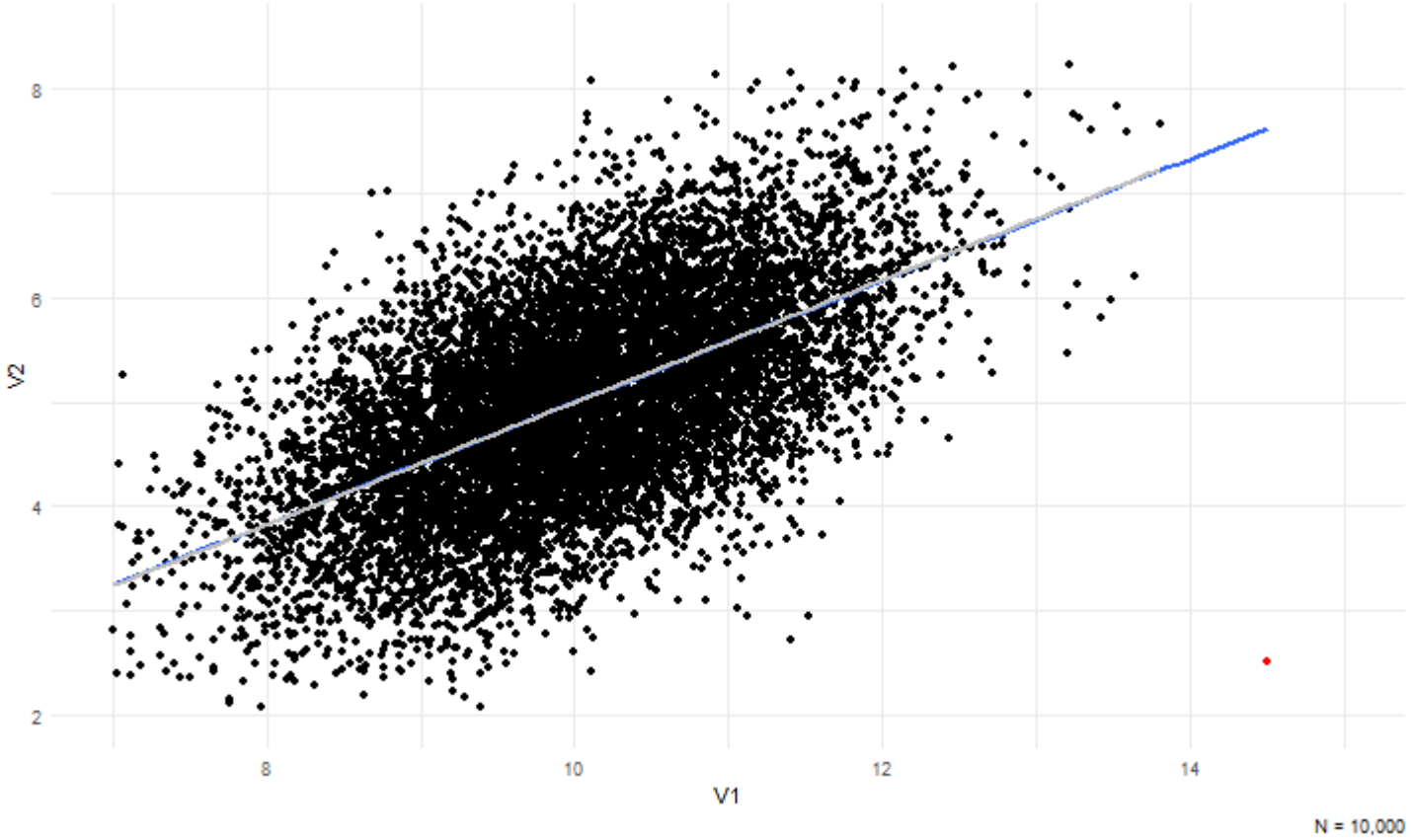
Depends on context



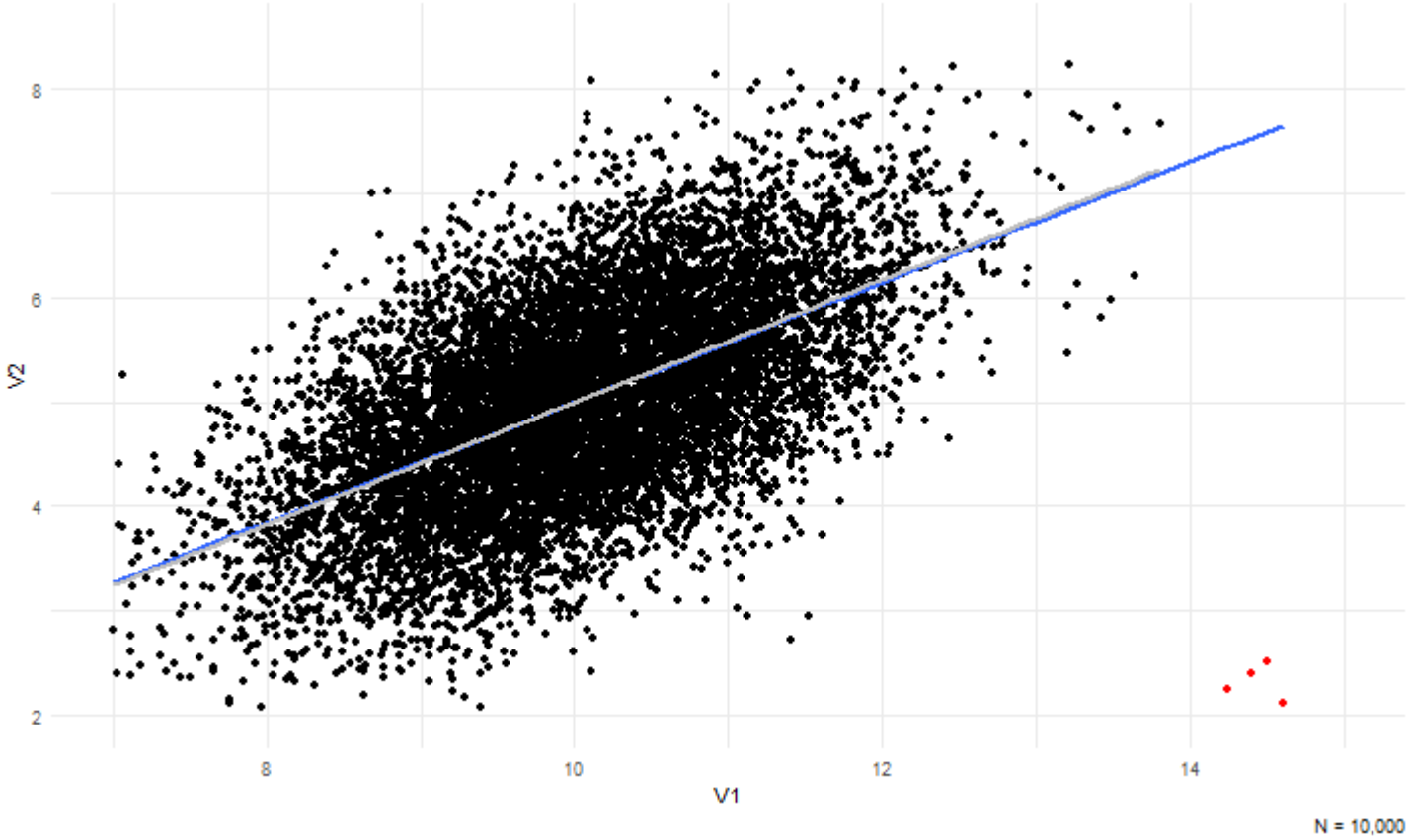
Depends on context



Depends on context



Depends on context



Outlier best practices

There are various tests and common "rules of thumb" you may hear to assess each of distance, leverage and influence:

- Mahalanobis distance, studentized residuals, Cook's distance, etc.

Our recommendation: address outliers substantively rather than with a particular statistic or "rule of thumb." Some recommendations:

- The more data you have, the less relevant outliers are
- Analyze data with/without outliers to see whether results change
- Use outlier identification to identify mis-coded data and/or individuals who are substantively different than the population of interest (e.g, a student who has 24 schools of enrollment in middle school is likely from a fundamentally different population than students who have between 1 and 6 schools)
- Consider processes such as top-coding (aka, "Winsorizing")
- Do not simply exclude observations because they are above a certain threshold

Synthesis and wrap-up

Screening steps

1. **Calculate univariate and bivariate descriptive stats**
    - Check max/min and examine for outlying observations
    - Verify the type of variable (factor or numeric?)
    - Verify the observation counts match the expected numbers
2. **Create boxplots, histograms, density and scatter plots**
    - Examine for outliers and floor/ceiling effects
    - Check for linearity, skew, and normality
3. **Test assumptions using residuals**
    - Graph residuals by predicted value. Check for linearity and heteroscedasticity
    - We'll address non-linearity in unit 4
    - Graph residuals by a few other variables to think about independence
    - Plot residual distribution and Q-Q plot to check for normality
    - Do not throw out data simply because it is a set number of SD away from mean
4. **Apply corrections in presence of assumption violations**
    - Generally a good idea to use heteroscedasticity-robust standard errors
    - In clustered settings (schools, facilities, hospitals), generally a good idea to use cluster-robust inference

Putting it all together

Ordinary Least Squares (OLS) estimators return the Best Linear Unbiased Estimate (BLUE) *as long as certain assumptions are met*.

- Before accepting a set of results, examine these assumptions in turn to make sure they are tenable. **You can't be sure the substantive interpretation of the results is correct until you have evaluated your assumptions!**
- Residuals can be key tools in evaluating these assumptions

Linear regression models assume that:

- There is no measurement error, the relationship is linear, the variance of Y at each value of X is homoscedastic, residuals are normally distributed, errors are independent and there are no unduly influential outliers.
- Violations of these assumptions mean that either our estimates will be biased, our standard errors will be biased, or both will be.

We can conduct a series of diagnostic steps and (sometimes) corrections to ensure our assumptions are met

- Develop a set of regular best practices that you implement every time prior to fitting models to ensure the basic assumptions of regression are satisfied
- Some assumption violations can be addressed via data management and statistical tools. Others are a feature of our research design that cannot be solved *post-hoc*.

Goals of the unit

- Articulate the assumptions of the General Linear Model broadly and least squares estimation and inference particularly
- Describe sources of assumption violation in the regression model including: measurement error, non-linearity, heteroscedasticity, non-normally distributed residuals, correlated errors, and outliers.
- Articulate properties of residuals and describe their centrality in understanding the regression model assumptions
- Conduct diagnostic tests on regression model assumption violations
- Implement a consistent screening protocol to identify regression model assumption violations
- Implement solutions to regression model assumption violations, when appropriate

To-Dos

Reading:

- **Finish by Jan. 31**: LSWR Chapter 15.8 - 15.9

Quiz 2:

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

Assignment 2:

- Due Jan. 31, 11:59pm

Next week: Multiple regression