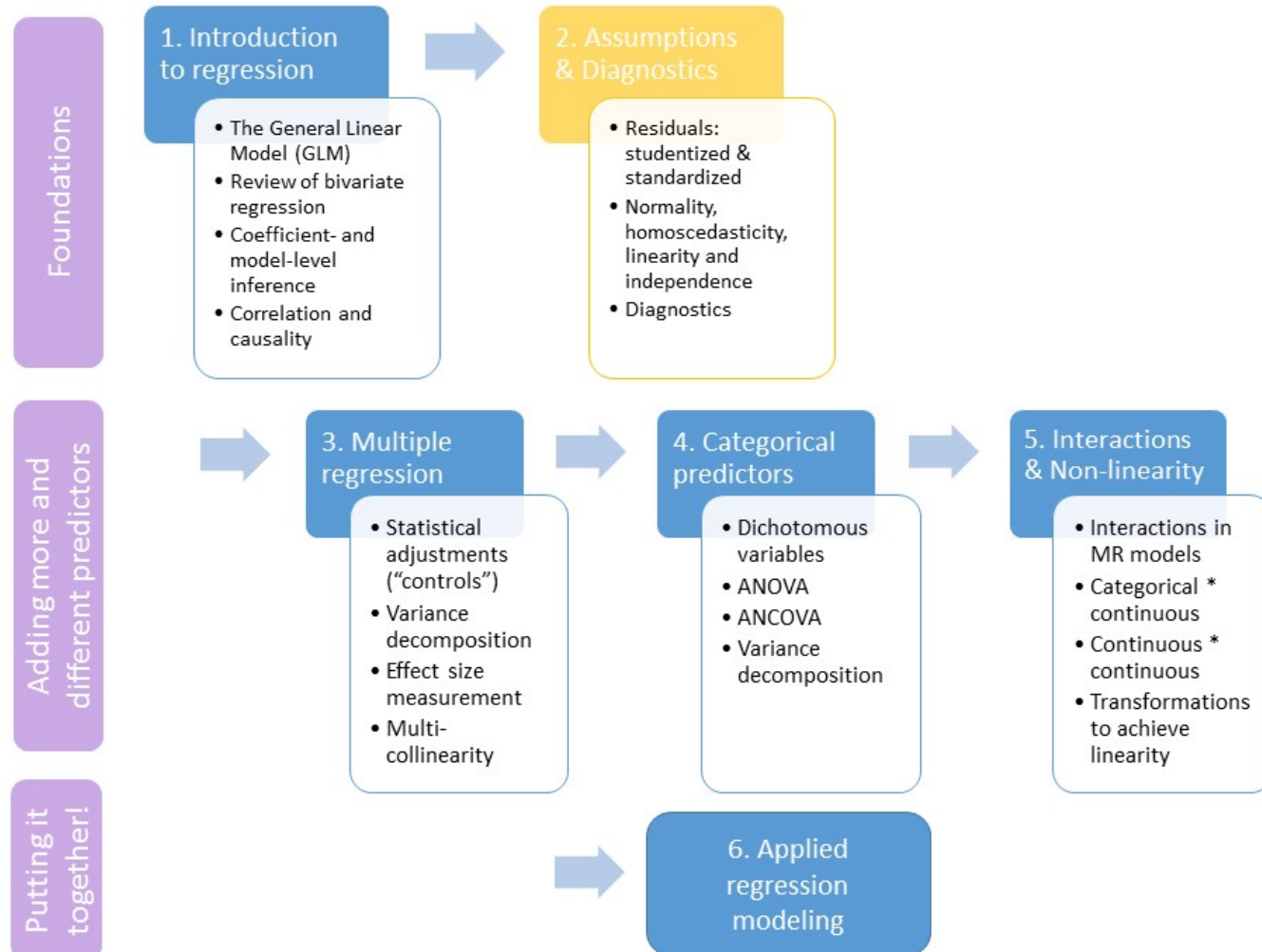


# Regression assumptions and diagnostics

EDUC 643: General Linear Model I

TBD

# Roadmap

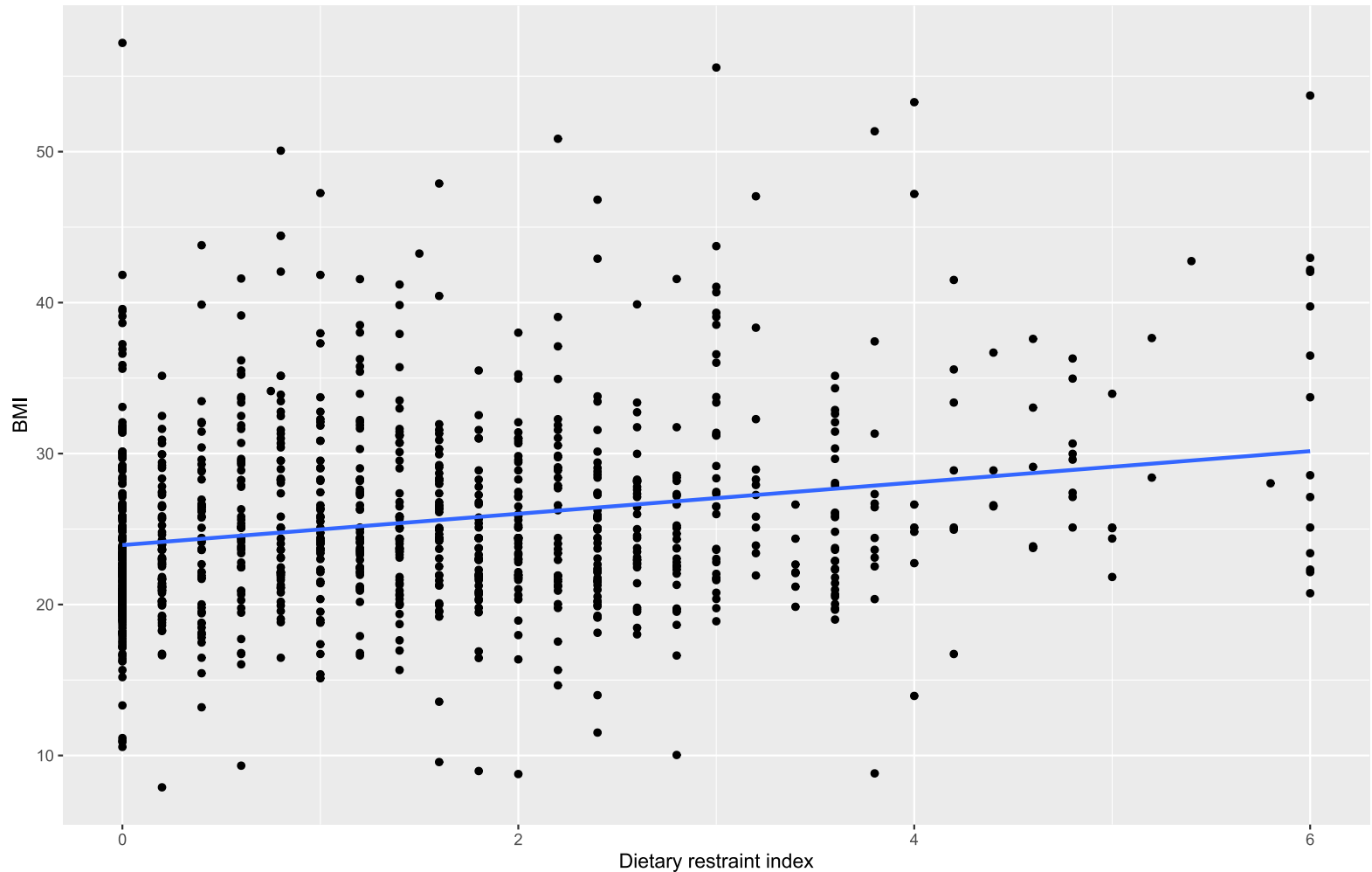


# A motivating question

Using real-world data about the dietary habits, health, and self-appraisals of younger males, we aim to answer the following RQ:

- **Does dietary restraint predict BMI among younger adult males?**

# Bivariate relationship



# Regression results

```
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       0.263     91.0      0
#> 2 EDEQ_restraint     1.04      0.136      7.64 4.80e-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 effect
- Effect 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

**So should we believe these results???**

# 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 three phenomenon you are interested in:

- **The estimand:** the underlying quantity you are seeking to learn about in the population
  - "the dropout rate", "disordered eating", etc.
- **The estimator:** the method used to obtain an approximation of this quantity
  - "Ordinary Least Squares regression", "t-test", "LASSO", etc.)
- **The estimate(s):** the value obtained from conducting a particular statistical test in a particular set of data
  - "38.1 scale points", "5.5 percentage points", "a log-odds ratio of 2.3", etc)

In estimating a model to data, you are seeking an estimator that will return a **BLUE** estimate:

- **B**est
- **L**inear
- **U**nbiased
- **E**stimate

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



# Key assumptions of regression

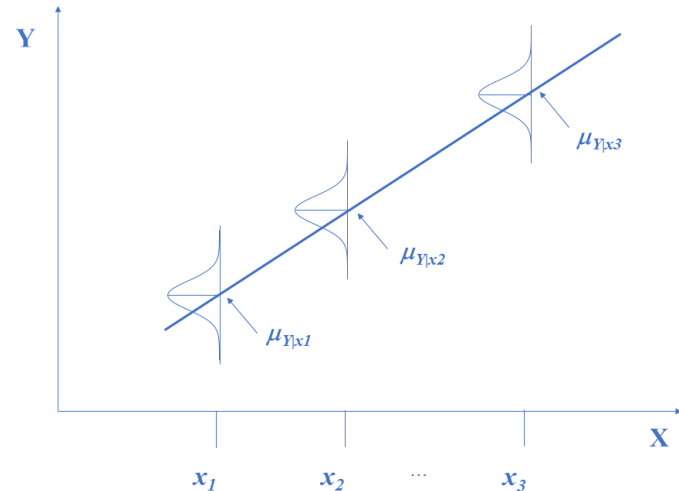
In fitting these models, we assume that the relationships are characterized by:

- 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  $\sigma_{Y|X}^2$

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 ( $\sigma_{Y|X}^2$ ) 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, but are about the individual error terms *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
#> -19.064  -3.962  -0.928   2.685  33.266
#>
#> Coefficients:
#>                Estimate Std. Error t value Pr(>|t|)
#> (Intercept)      23.9383     0.2631  90.973  < 2e-16 ***
#> EDEQ_restraint   1.0369     0.1358   7.638  4.8e-14 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 6.081 on 1092 degrees of freedom
#> Multiple R-squared:  0.05072,    Adjusted R-squared:  0.04985
```

# Adding 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
#> -19.351  -3.967  -0.969   2.778  32.410
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  24.1607      0.2546  94.909  < 2e-16 ***
#> EDEQ_noise    0.8717      0.1257   6.937 6.83e-12 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 6.108 on 1092 degrees of freedom
```

# Addressing measurement error

- Test for presence of measurement error with reliability statistics
- Solve by creating factors of underlying constructs, adjust standard errors post-hoc, SEM, get more data, improve design

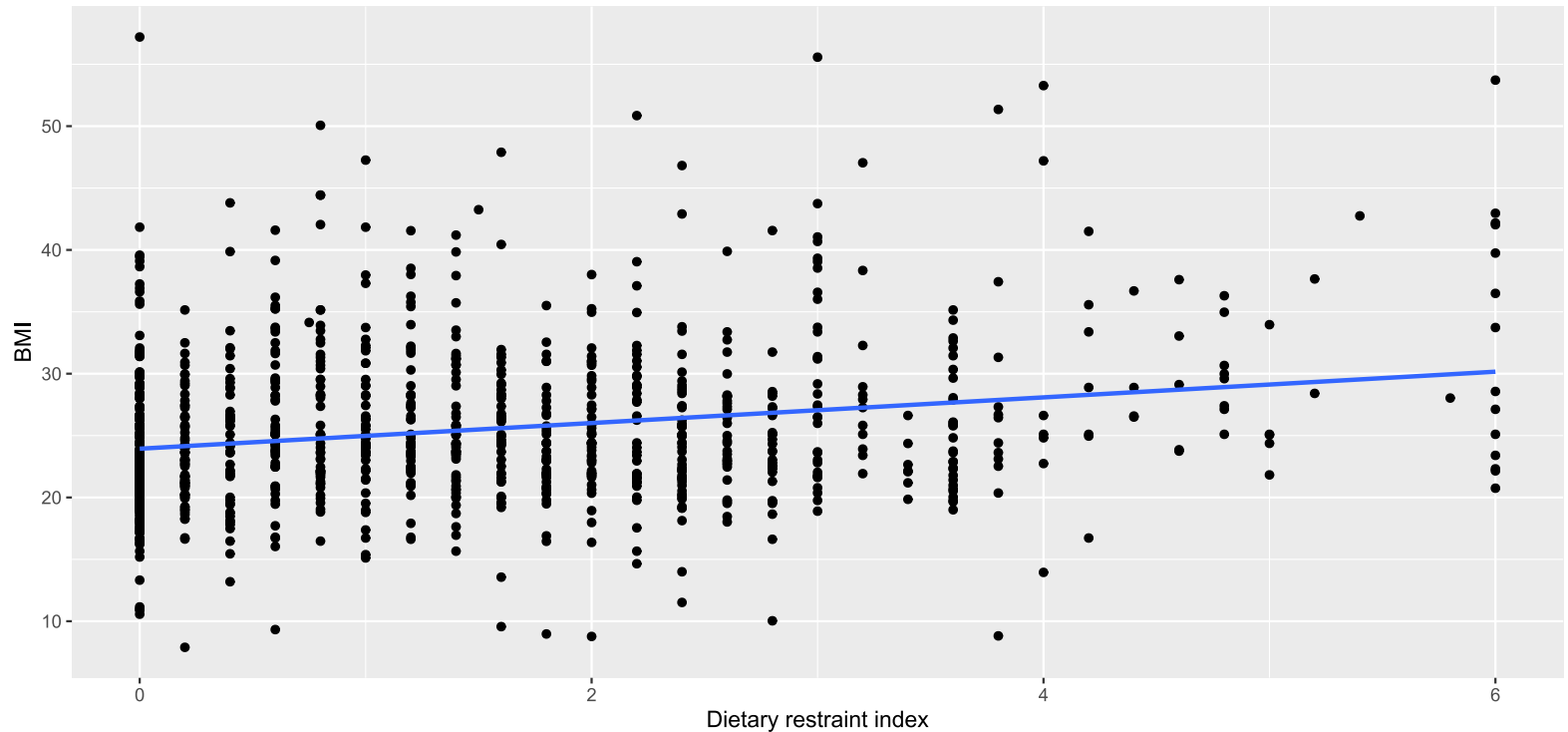
We'll learn about reliability in EDUC 645 as well as in more advanced courses such as EDLD 667

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

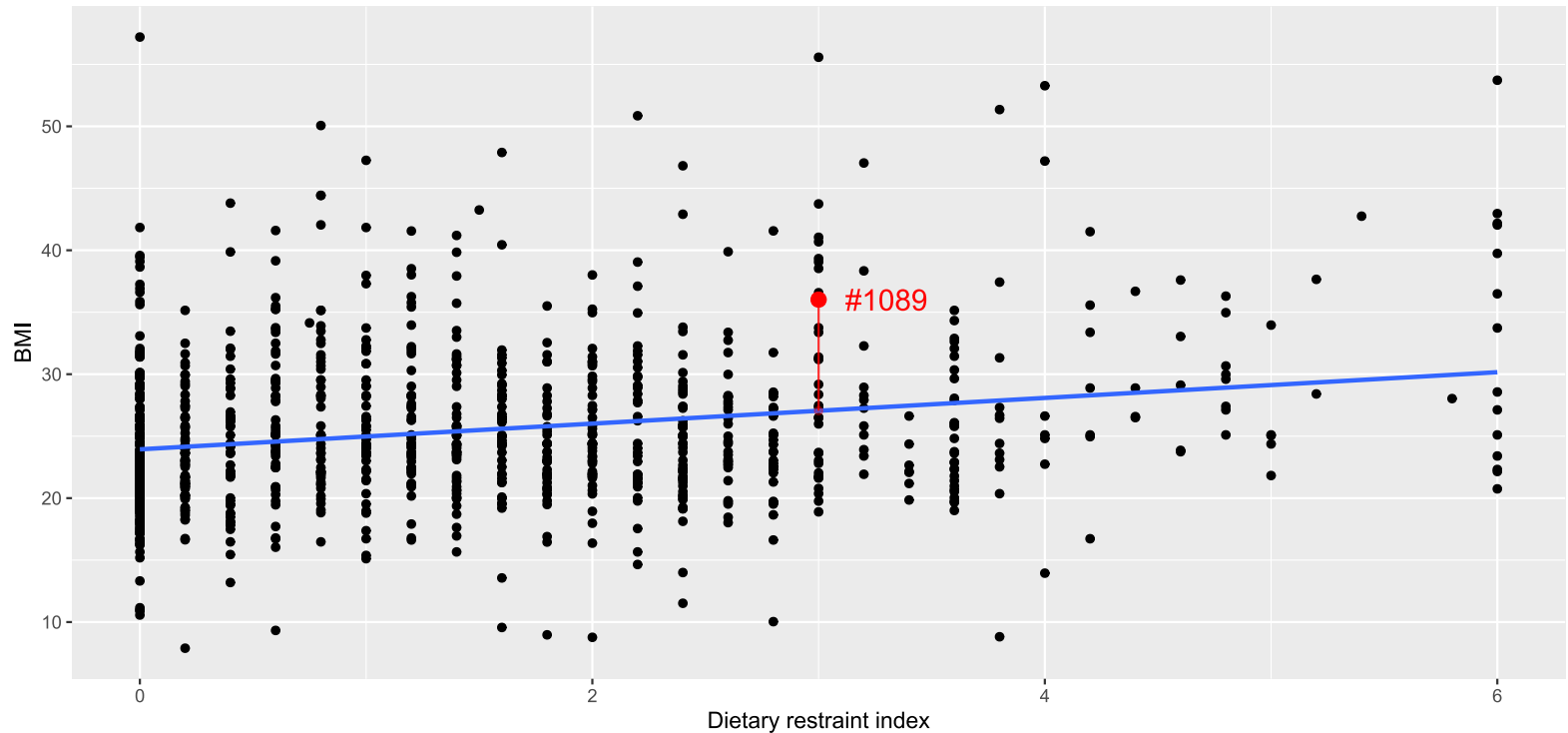
# Residuals



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 #1089



*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 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*. Sometimes it may be valuable to be able 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).<sup>1</sup>

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

- **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

# 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 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.
#> -19.064  -3.962  -0.928   0.000   2.685   33.266
```

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

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

```
#>      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#> -3.140908 -0.651970 -0.152739  0.000022  0.441871  5.475401
```

- Compare to the standardized residuals. *How similar are they?*<sup>1</sup>
- Note the minima and maxima 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 18)

# Examining the residuals

Standard deviation of the raw residuals can be quite useful in checking our assumptions

- *What assumption?*

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

```
#> [1] 6.07839
```

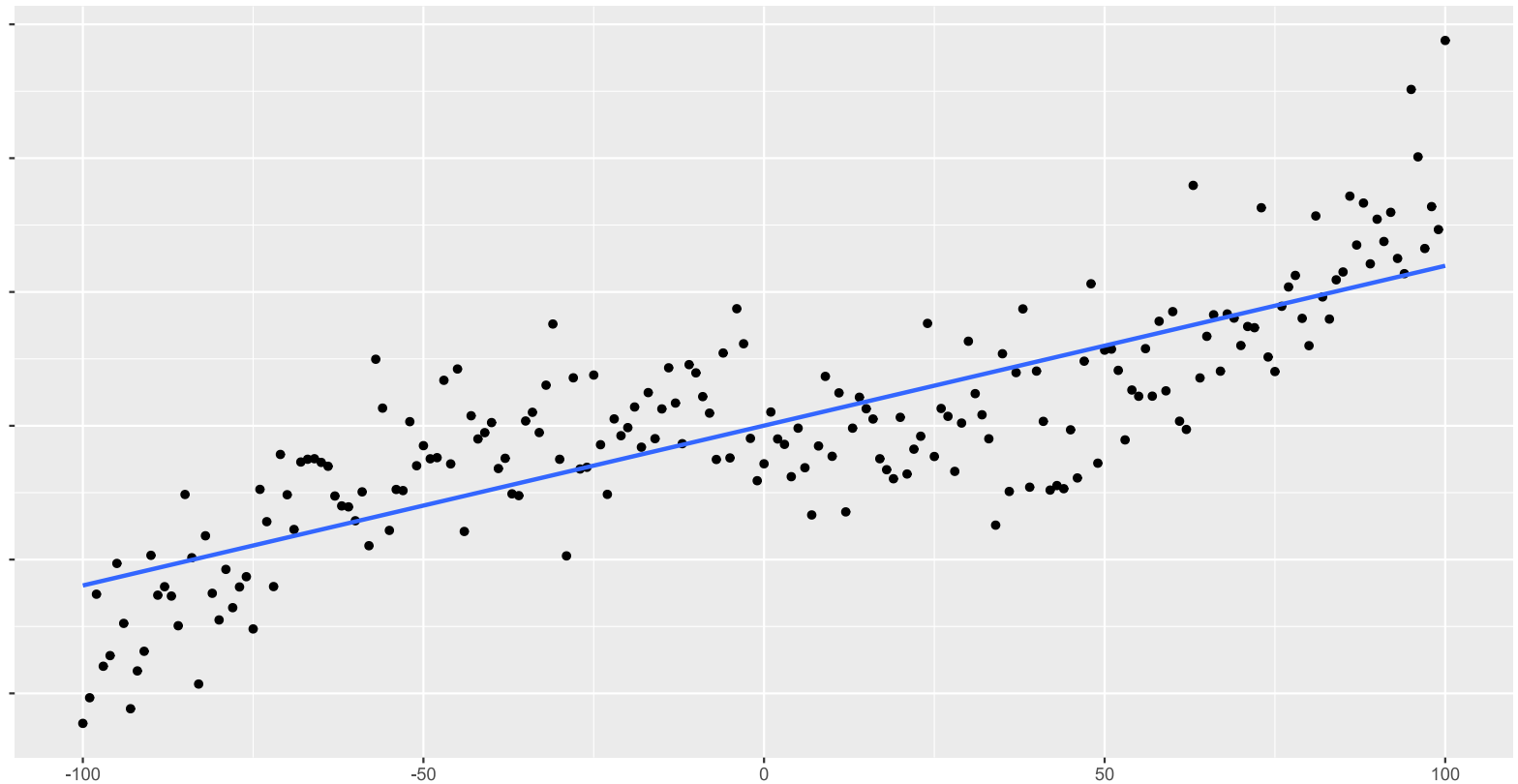
How does this compare to the standardized residual?

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

```
#> [1] 1.000621
```

# Non-linearity

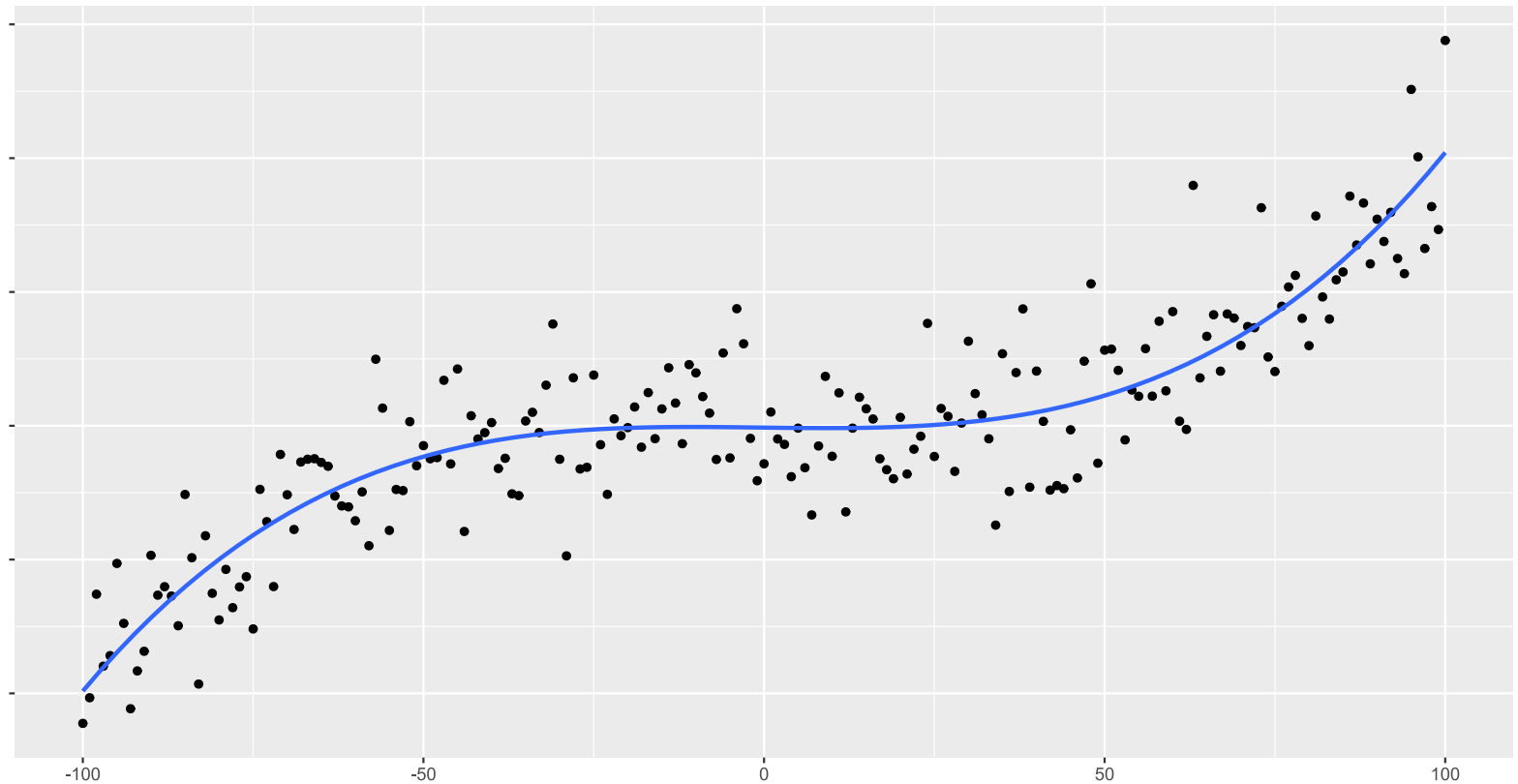
In addition to examining the original scatterplot to detect the presence of non-linearity...





# Non-linearity

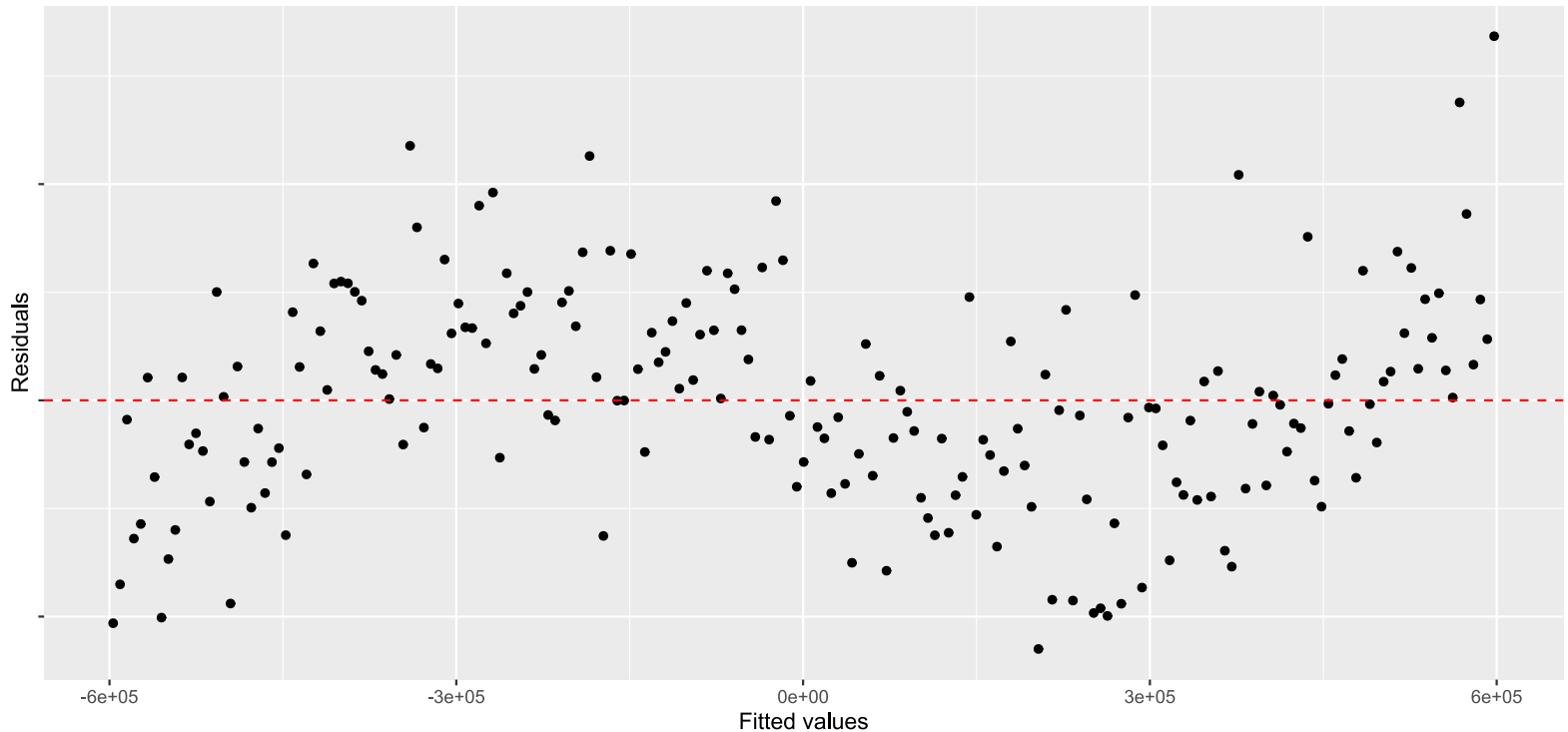
In addition to examining the original scatterplot to detect the presence of non-linearity...



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

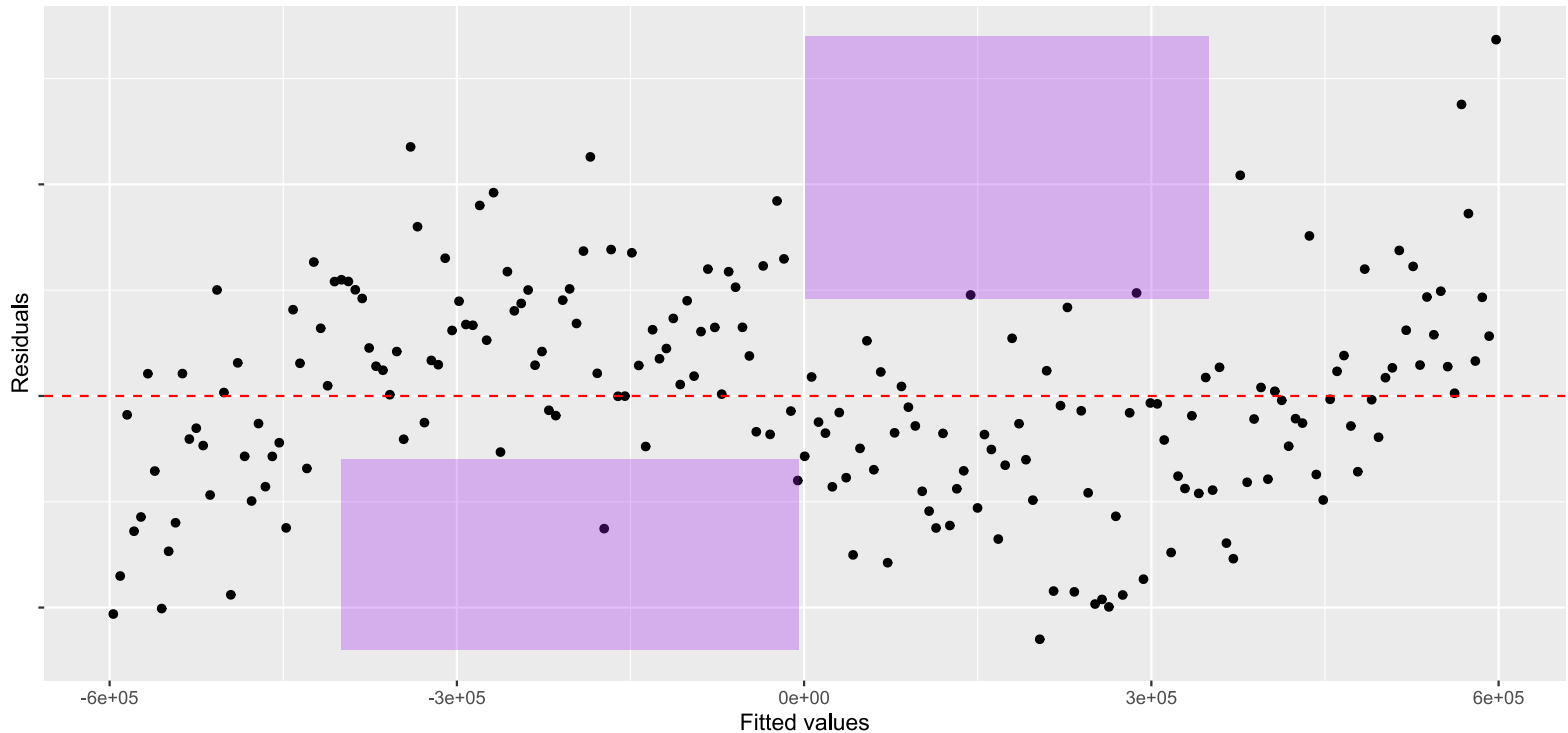
# Non-linearity: residuals v. fitted

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



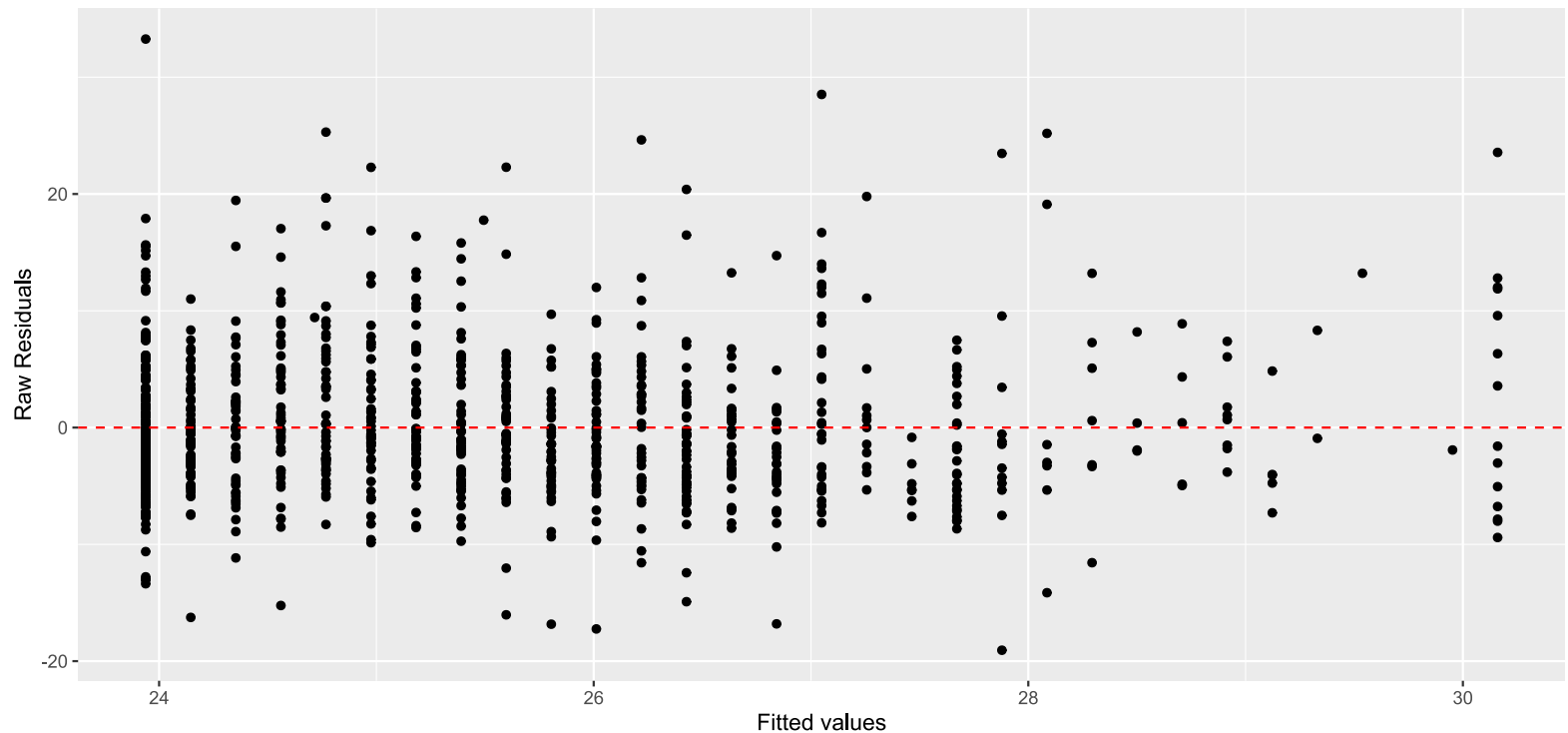
# Non-linearity: residuals v. fitted

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



# Non-linearity: residuals v. fitted

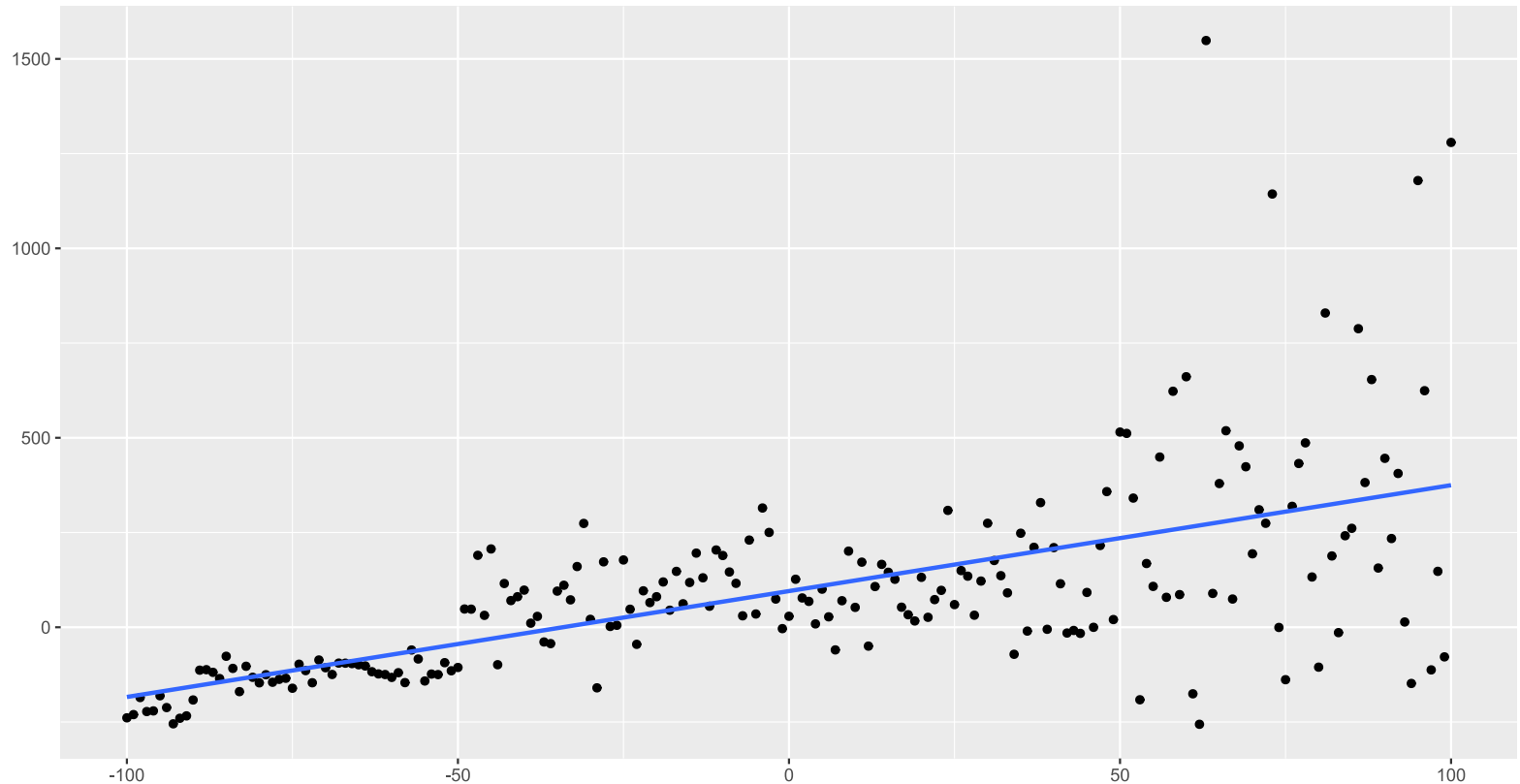
Let's look at this for our estimates of disordered eating:



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

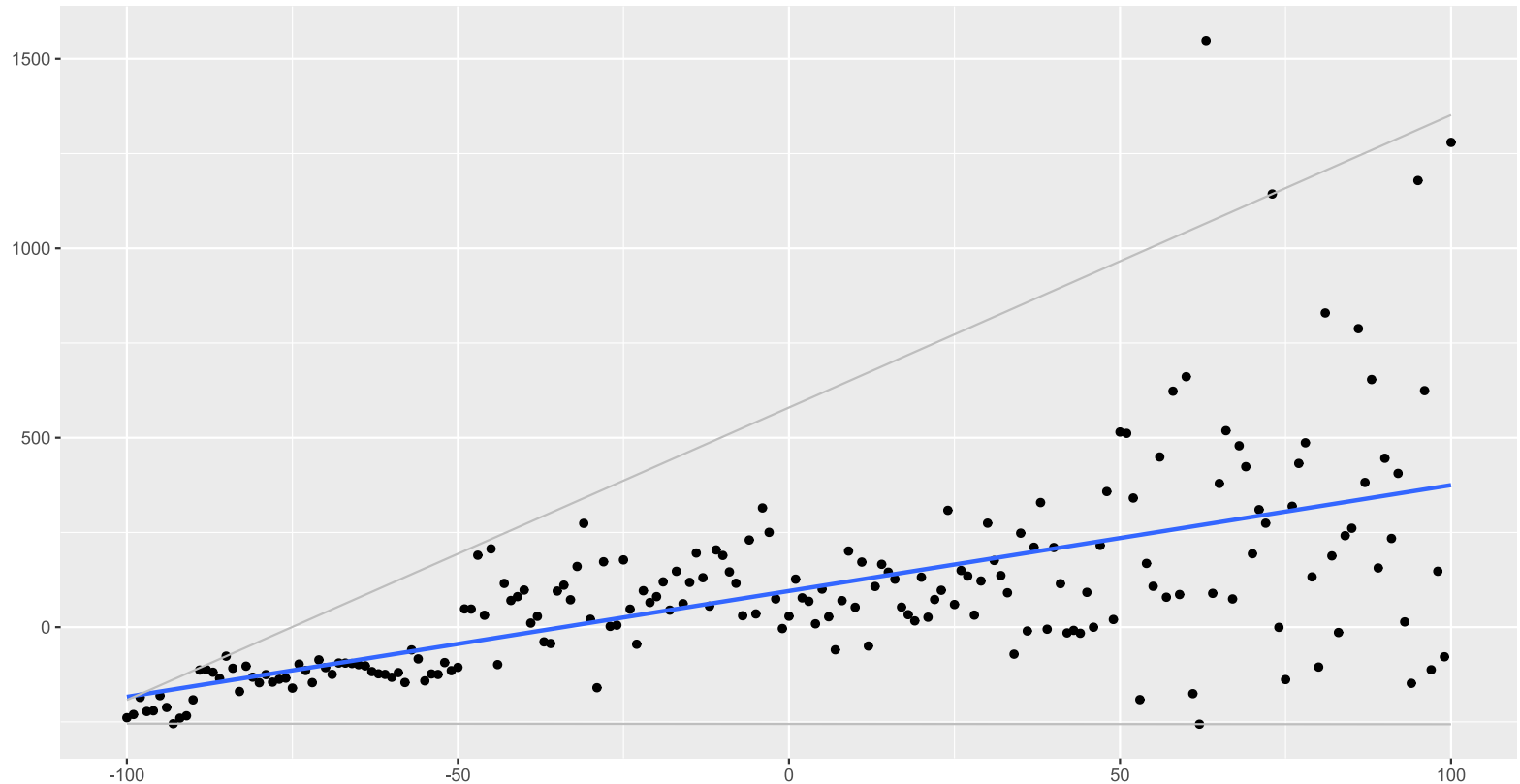
# Heteroscedasticity

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

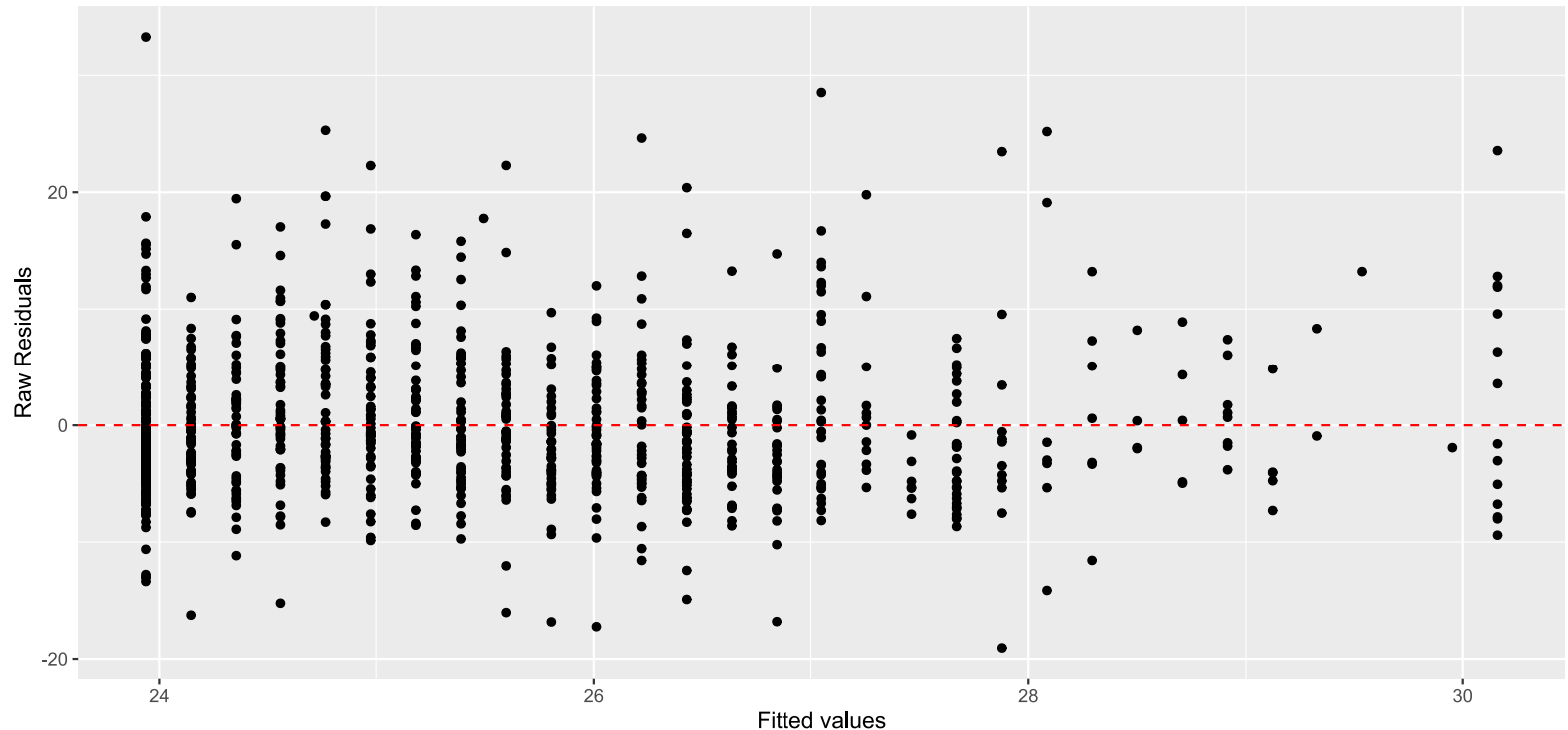


# Heteroscedasticity

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



# Residual heteroscedasticity



**What evidence do you observe of homoscedasticity or heteroscedasticity 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  $\sigma^2$  when there is heteroscedasticity. The sample-based statistic do not follow a standard normal distribution (following 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 the `estimatr::lm_robust` command.

```
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      0.263     91.0     0
#> 2 EDEQ_restraint  1.04      0.136     7.64 4.80e-14
```

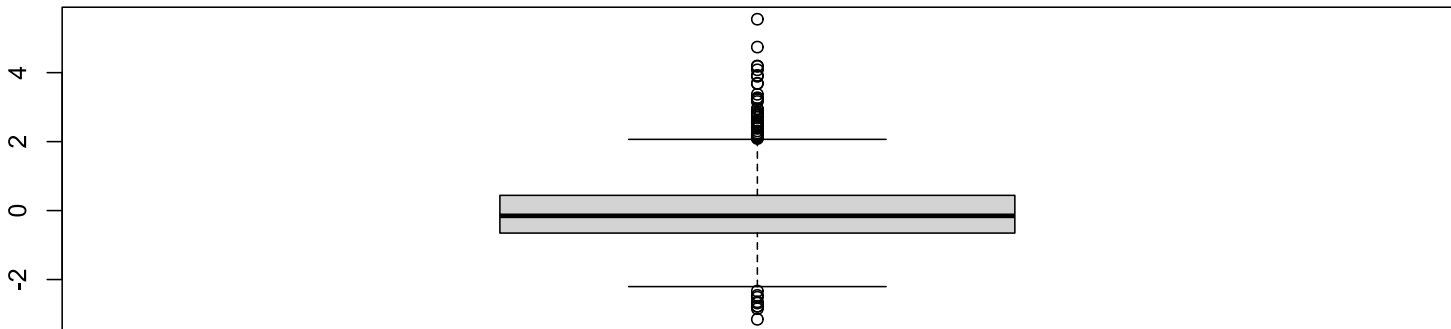
```
tidy(estimatr::lm_robust(BMI ~ EDEQ_restraint, se_type="HC1", data=do))
```

```
#>           term estimate std.error statistic      p.value  conf.low
#> 1 (Intercept) 23.938280 0.2556127 93.650586 0.000000e+00 23.4367326
#> 2 EDEQ_restraint 1.036911 0.1578113  6.570577 7.743309e-11  0.7272636
#>   conf.high  df outcome
#> 1 24.439828 1092     BMI
#> 2  1.346559 1092     BMI
```

# 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?**

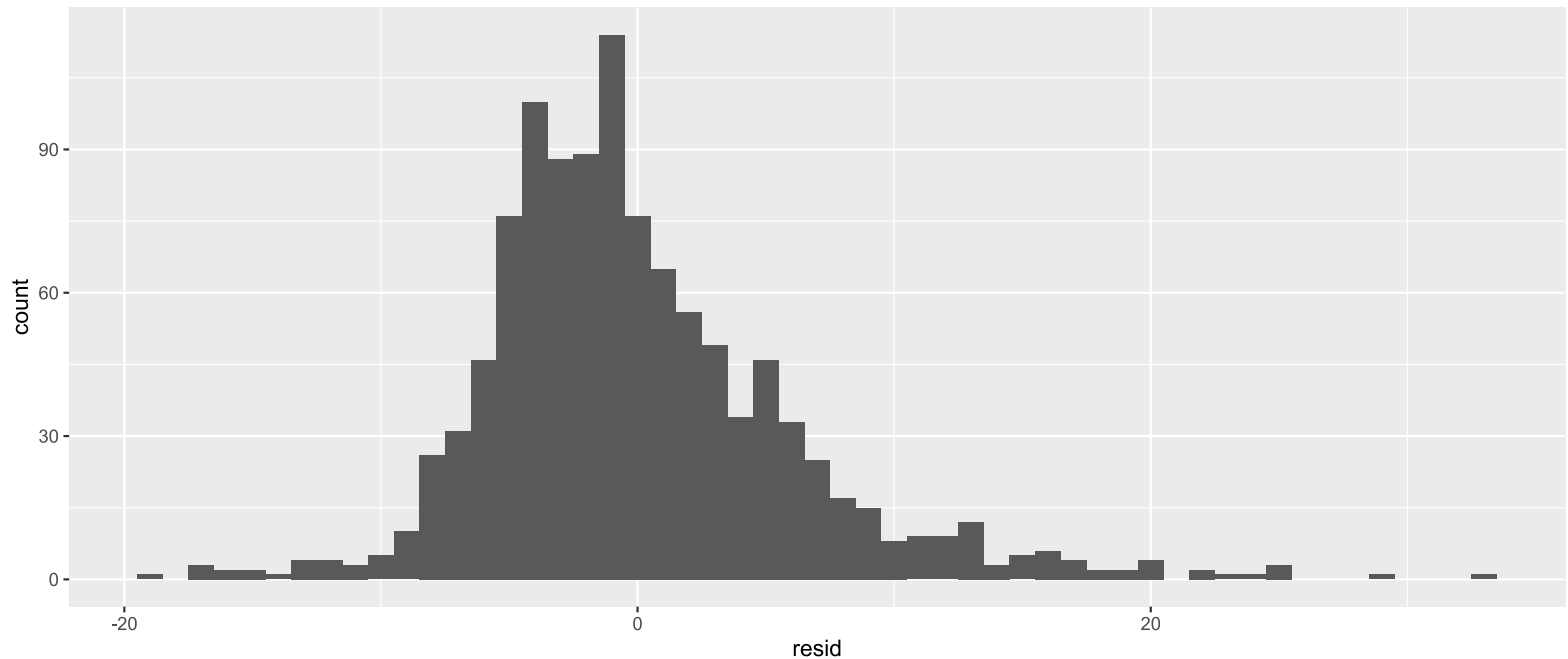
```
boxplot(rstudent(fit))
```



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**

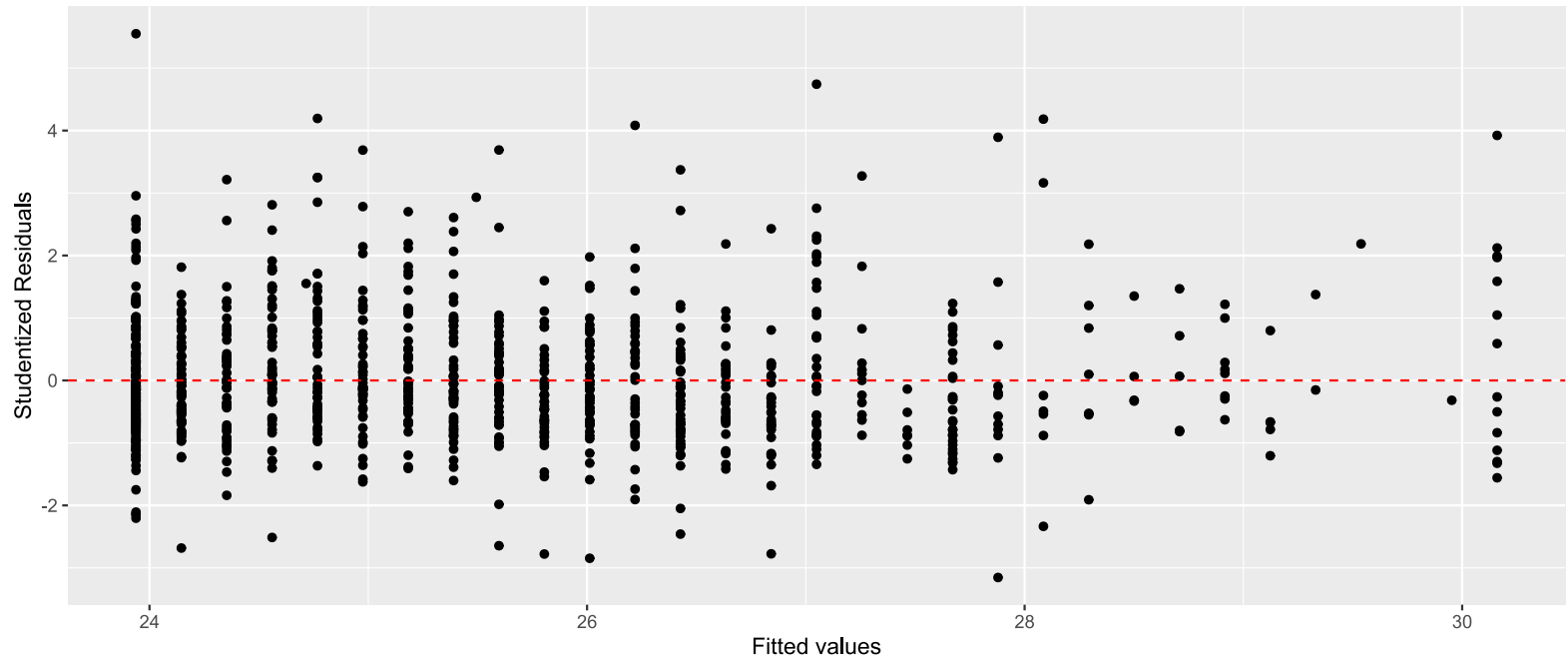


And this is confirmed in this histogram. [What sort of skew is this distribution exhibiting?](#)

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

# Residual normality: resid v. fitted

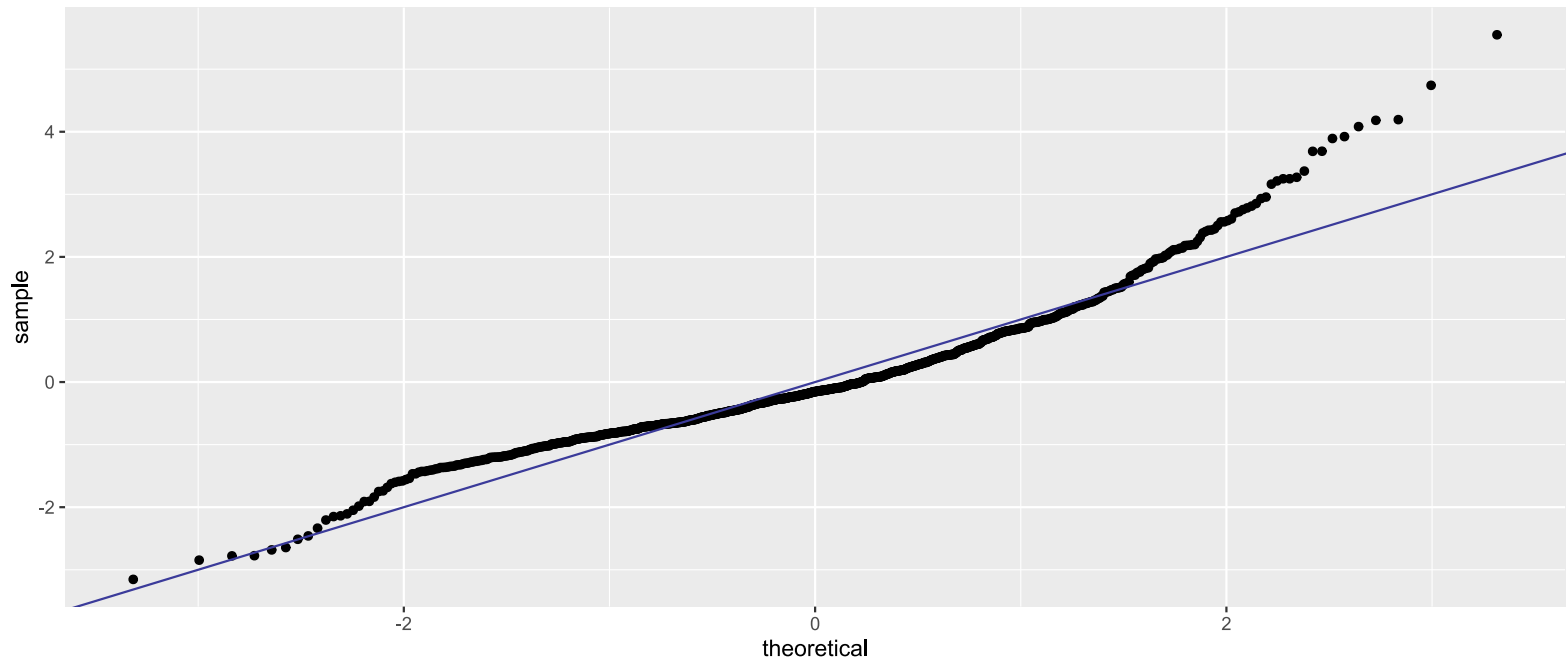
For the standard errors (and associated inference tests) that we conduct in the regression analysis to be correct, the residuals **must be normally distributed**



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

# Residual normality: QQ plot

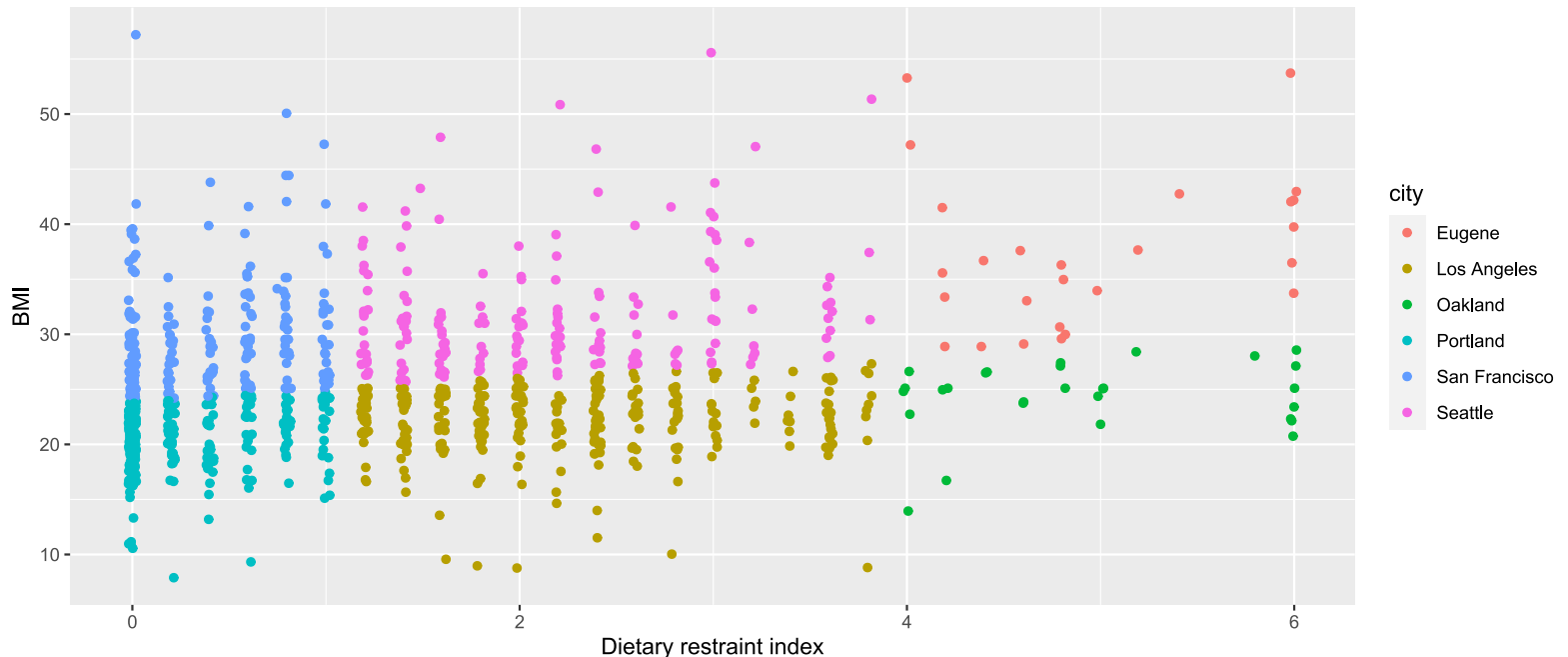
For the standard errors (and associated inference tests) that we conduct in the regression analysis to be correct, the residuals **must be normally distributed**



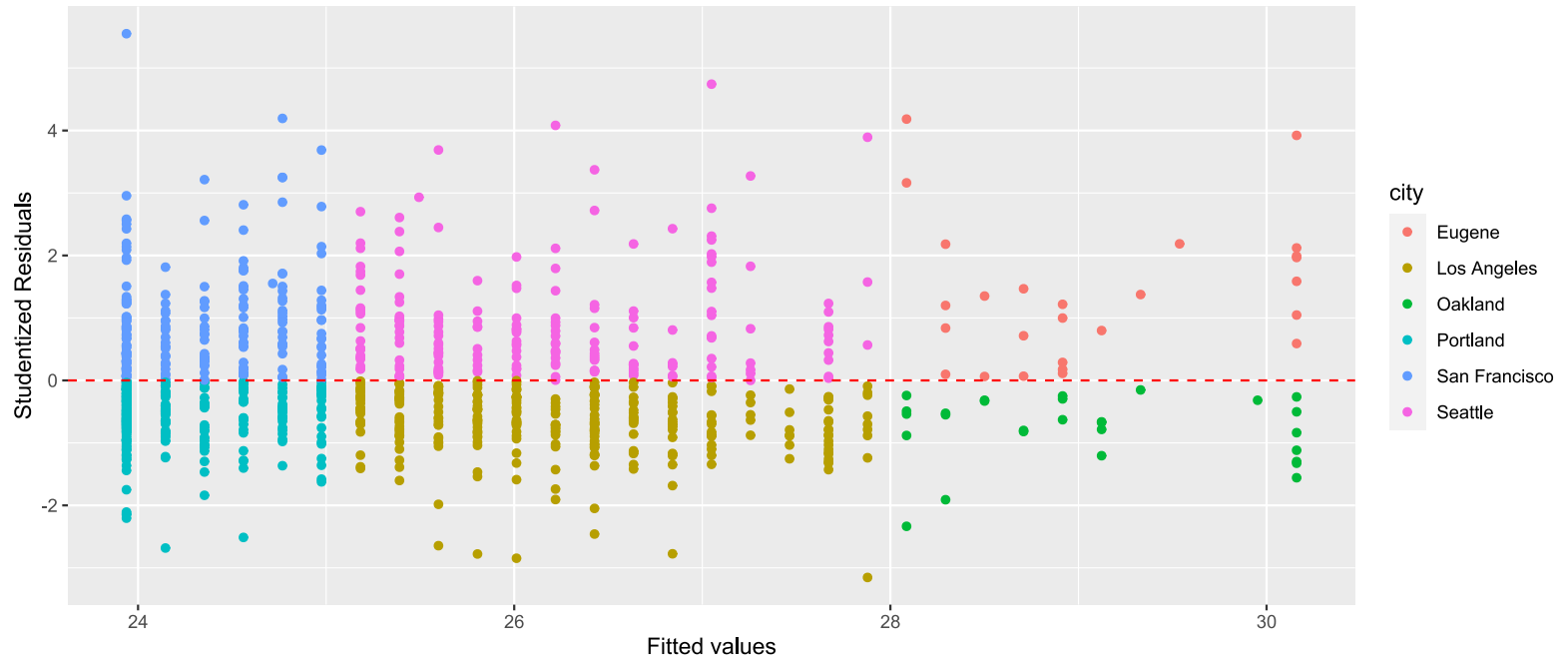
# Residual independence

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



**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)**

# Outliers

Broadly defined as highly atypical and/or influential data point(s). These could be a result of:

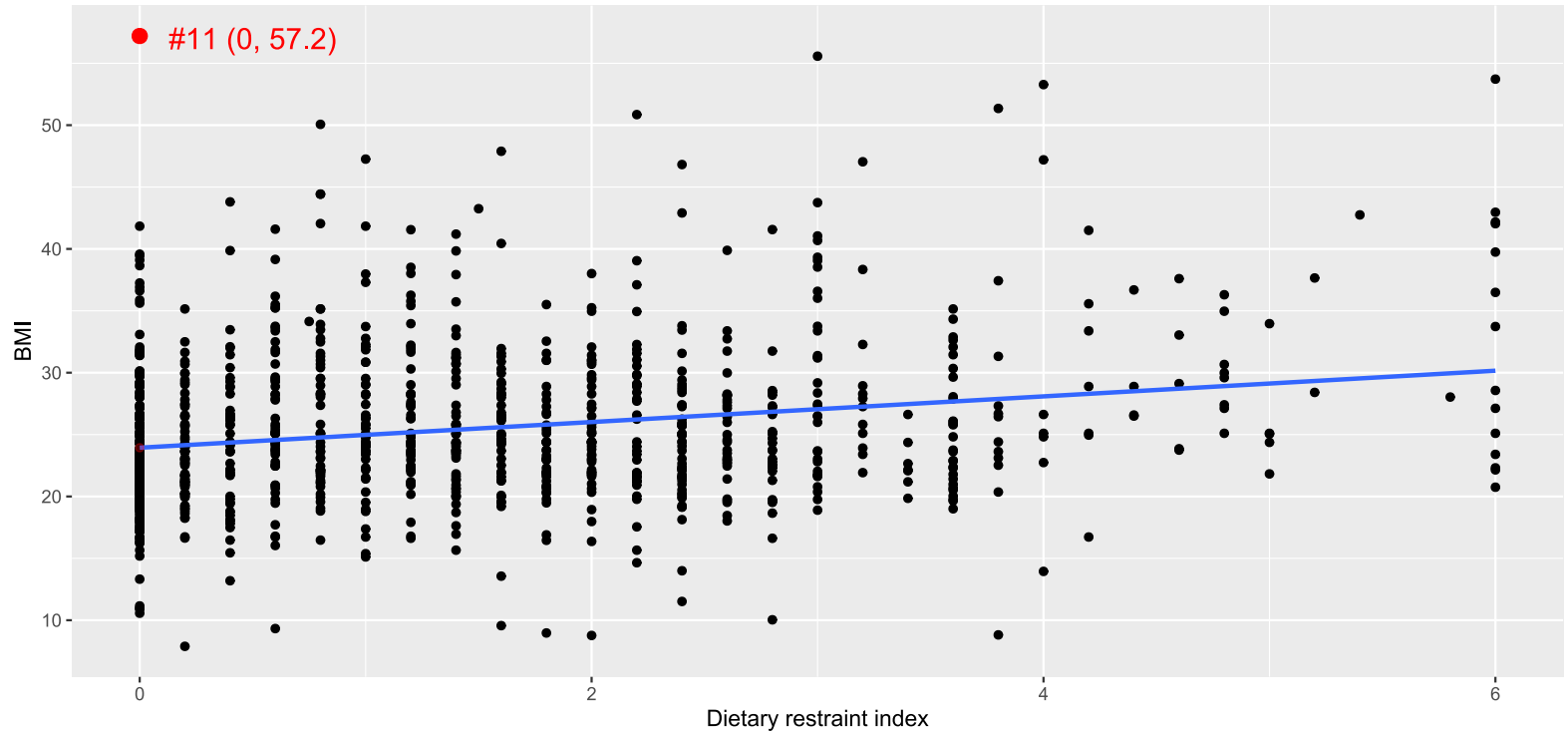
- Coding (recording) error
- Accurate observations of a rare case
- Observation of an individual from a different population

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  $\text{Leverage} * \text{Distance}$



# 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      0.263     91.0      0
#> 2 EDEQ_restraint  1.04      0.136      7.64 4.80e-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      0.260     91.9      0
#> 2 EDEQ_restraint  1.06      0.134      7.91 6.29e-15
```

**Things don't change much in this case!**

# 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)

# 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
- Graph residuals by a few other variable to think about independence
- Plot residual distribution and Q-Q plot to check for normality

# 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*.

# Synthesis and wrap-up