Regression Diagnostics

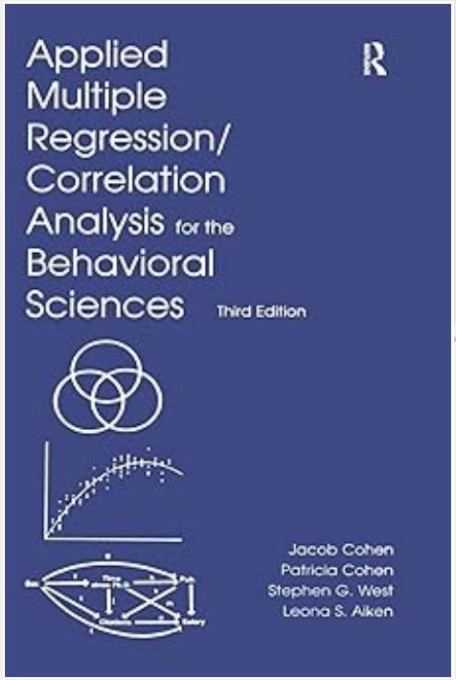
PSY 612

Regression Diagnostics

- **Regression diagnostics** refers to 1) screening the data for any potential issues that could lead an analysis to produce misleading or inaccurate results and to 2) evaluating whether model assumptions have been met.

Running Diagnostics

- ❑ Outliers
- ❑ Missing Data
- ❑ Multicollinearity
- ❑ Violations of Correctly Specified Form Assumption
- ❑ Violations of Normality Assumption
- ❑ Violations of Homogeneity of Variances Assumption
- ❑ Violations of Independence Assumption



Outliers

Outliers should be examined and handled prior to other potential issues with the data because they can be the cause of some of the other issues that we can run into.

Potential Causes of Outliers

- **Data Entry Errors**
  - Ex: Entering '77' instead of '7', or entering '1998' in instead of '26'
    - There's consensus that data entry errors should be fixed, if possible, or removed because they misrepresent how someone actually scored on a variable

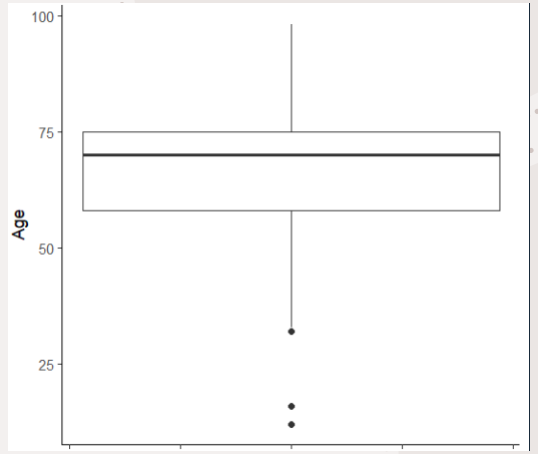
- **Inaccurate Measurements**
  - Ex: A participant falls asleep during a reaction time test and has a reaction time of 300 seconds.
    - These outliers can be justifiably removed on the basis of not measuring the intended construct.

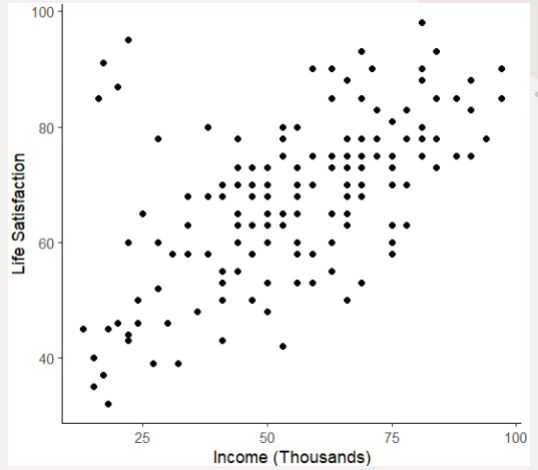
- **A real, but infrequent (in your data), occurrence**
  - Ex: A participant in your sample is 85 whereas the rest of the sample falls between the ages of 20-35.
    - There is less consensus on how to deal with these types of outliers that reflect a real, but infrequent (at least in your data), phenomenon.

Outliers

- **Univariate Outliers:** unusual values on a single variable, including either one of the predictors, X_K, or an outcome variable, Y_K

- **Multivariate Outliers:** unusual combination of values on a set of variables





Univariate Outliers

- **Univariate Outliers on a Predictor Variable (X_k)**
  - Recall the formula for calculating the slope of a regression model with a single predictor.

$$b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$$

  - The weight each score has on calculating the overall slope can be written as:
    - Scores on the predictor, X, further away form the mean contribute more to the overall slope of the model

$$b_1 = \Sigma w_i \left[\frac{Y_i - \bar{Y}}{X_i - \bar{X}}\right], \text{ where } w_i = \frac{(X_i - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}$$

- **Leverage** (or **hat** in R) is a measure of each observation's weight on determining the slope and intercept of the overall model

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$$

  - $\frac{1}{n}$ = each observation contributes equally to the model's y-intercept
  - $\frac{(X_i - \bar{X})^2}{\sum(X_i - \bar{X})^2}$ = observations with X values further from the mean of X are weighted more in calculating the model's slope

Rule of thumb: Inspect observations with a leverage larger than $\frac{3p}{n}$, where p is number of parameter estimates and n is sample size.

Univariate Outliers

Univariate Outliers on an Outcome Variable (Y_k)

- Outliers on Y cannot be determined by simply looking at each observation's residual, $e_i = (Y_i - Y'_i)$ because
  - They aren't in standardized values, and
  - If there are outliers, the predicted values on Y themselves have already been influenced by the presence of the outlier

- **Studentized residuals** assess outliers on Y by comparing actual scores on the outcome variable, Y_i, to values on Y predicted by a model that *excludes* that observation and expressing the difference in standard deviation units.

Deleted residual

$$d_i = y_i - \hat{y}_{(i)}$$

Studentized residual

$$\frac{d_i}{s(d_i)}$$

Rule of thumb: Inspect observations with a studentized residual larger than ±3.

Multivariate Outliers

- Multivariate outliers are outliers that have an unusual combination of scores on the set of variables included in the model.
  - One way of assessing whether an observation is a multivariate outlier is by assessing whether it has undue influence on the fit of the model.

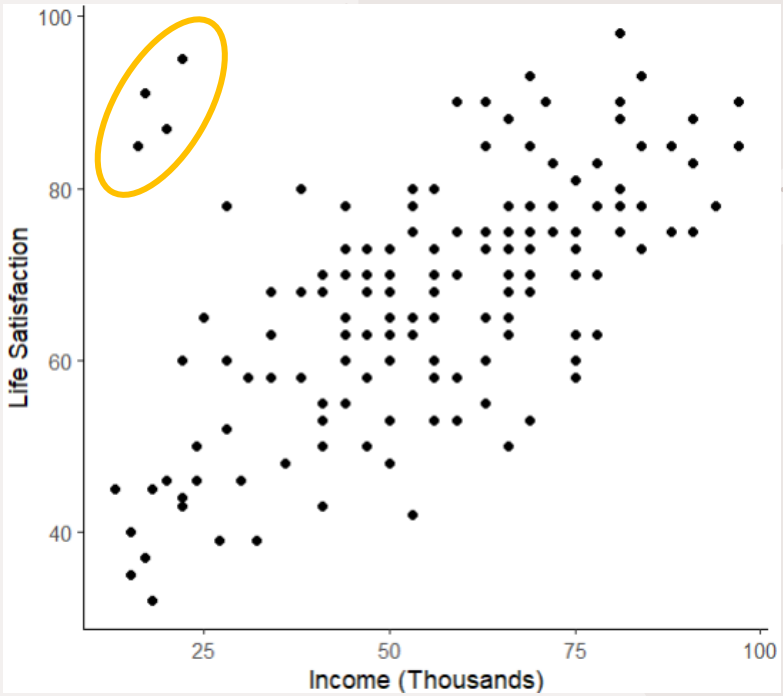
Cook's Distance (aka, **Cook's D**) measures how much the regression model would change if you removed a particular case. It examines how different the fitted values would be with versus without a particular observation.

- $\hat{Y}_i$ = original predicted value on Y for observation *i*
- $\hat{Y}_{i,[k]}$ = the predicted value on Y for observation *i* with the *k*th observation omitted

$$D_k = \frac{\sum_i (\hat{Y}_i - \hat{Y}_{i,[k]})^2}{\text{PA(MSE)}}$$

Rule of thumb: 1) Inspect observations with a Cook's D value larger than 0.5, or 2) inspect observations with a Cook's D value that is noticeable larger than the others'.

Handling Outliers



The options for handling an outlier are:

- Keep the outlier
- Remove the outlier
- Recode the outlier

The choice of how to handle an outlier depends on the reason for the outlier.

- As mentioned earlier, data entry errors and inaccurate measurements should be corrected (if possible) or removed.
  - *Q*: How do you think these outliers should be handled?
- One option for handling outliers that reflect a *real phenomenon* is to report the results of your analysis both **with and without** the outliers included.

Missing Data

- Another issue that researchers encounter when handling real data is **missing data**, which occurs when one or more values are missing for a participant across variables.

- Missing data can occur for various reasons:
  - Participant got bored and/or rushed through parts of the study
  - Participant dropped out of a long-term study (attrition)
  - Participant chooses not to answer a question for a particular reason

Handling Missing Data

- A classic way of handling missing data is called **listwise deletion**, which removes all participants from the analysis who have missing data on *any* of the variables included in the model
  - Also called **complete case analysis**
  - This is the default method used by lm( ) for handling missing data

- **Listwise deletion** is not the ideal way of handling missing data because it results in:
  - A ***loss of power*** – analysis is conducted using a smaller sample size
  - ***Biased parameter estimates*** – especially if there is something in common amongst the people who tend to have missing values
    - Ex: If people who identify as transgender are less likely to choose a response when asked to identify their gender (perhaps because an option that they identify with is not provided), and gender is included as a variable in the model, the model would estimate its parameter estimates while leaving out these individuals' responses completely.

Handling Missing Data

- Better methods for handling missing data include **multiple imputation** and **full information maximum likelihood** (you'll discuss FIML in 613 when you learn multilevel modeling).

- **Multiple imputation** uses a regression model based on associations observed between the variables in the model among other participants in the data set to predict a particular participant's missing value on a variable (plus random noise to reflect the uncertainty of these values).

Multiple Imputation

The steps of performing multiple imputation to handle missing data in R include:

1. Impute multiple data sets
   - Each will have unique random error added to predicted scores for missing values
   - A commonly-used number of imputations is 5 ($m = 5$), but increase the number of imputations as the fraction of missing data increases to prevent a loss of power (Graham, 2007)
   - Set seed before imputing data sets so results are replicable

2. Analyze model on each imputed data set
   - Each imputed data set is analyzed individually according to a common statistical model

3. Pool the results
   - Average the results across the individually conducted models

You'll get to perform multiple imputation using the `mice` package in lab.

Multiple Imputation

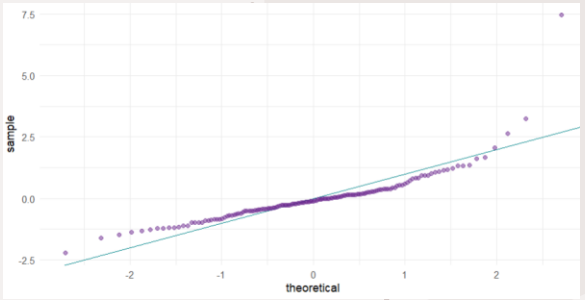
- See Woods et al. (2021) for an excellent discussion of using multiple imputation in social science research.

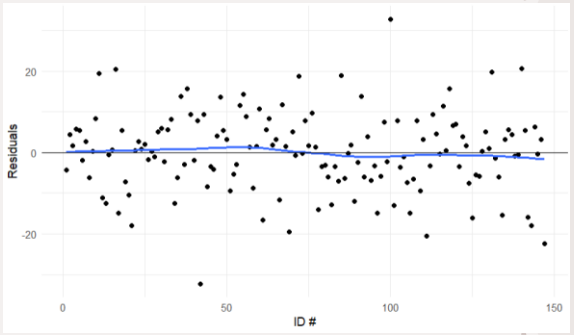
TABLE 1 Debunking misconceptions about multiple imputation.

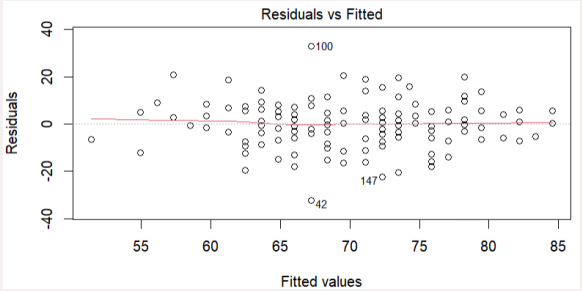
Misconceptions	Reality
Multiple imputation should only be used when the missingness is MAR	MAR is the least restrictive assumption for multiple imputation. Therefore, multiple imputation is also appropriate (and better than listwise deletion due to increased statistical power) under the more restrictive MCAR assumption. Even under MNAR, multiple imputation (used with sufficient auxiliary variables) can offer advantages over other approaches (e.g., deletion-based methods).
Multiple imputation should only be used when too few cases are left after listwise deletion	Multiple imputation has advantages even when the amount of missing data is low (i.e., because multiple imputation will eliminate bias under MAR and can partially eliminate bias under MNAR).
If results from statistical analyses obtained from multiple imputation differ from those of listwise deletion, the results of multiple imputations must be wrong	Results of multiple imputation have been shown to be more accurate and reduce bias in parameter estimates compared to deletion techniques when the multiple imputation model is correctly specified.
Certain variables must not be imputed (outcomes/predictors)	With the exception of special instances, most variables can be multiply imputed with benefits. Caution in using multiple imputations is, however, warranted for missing social identity data for ethical concerns (Randall et al., 2021).
Multiple imputation must not be used because it can produce several different outcomes in statistical analyses	Following the computation of multiply imputed data, point estimates from the analysis of each data set are pooled to provide one overall estimate. Generally, this is done using Rubin's (1987) rules. However, sometimes a pooling method is not available for certain commands in your software package of choice. In these instances, we recommend switching to another package. If this is not possible, transparently reporting an ad hoc solution is key.
Multiple imputation is making data up	Algorithms for imputing missing data use the available data to optimize the accuracy of missing values that are replaced. Sufficient multiple imputations allow researchers to estimate the most likely values for the variable and case while incorporating uncertainty.
Doing anything other than listwise or pairwise deletion is hard enough that it is not worth doing	With some training, researchers can develop skills to implement best practices for handling missingness such as multiple imputations, which can be completed in a reasonable amount of time and will ultimately provide knowledge producers and consumers with a more accurate understanding of the relations that are being examined. Researchers may also utilize the skills of a methodological consultant to help incorporate best practices for missing data analysis in their design and analysis.
The computational demands of multiple imputation are too intensive and/or will take too long to complete	Thanks to advances in computing power, only very complex analyses or 'big data' such as neuroimaging and genomics data sets are likely to have computational constraints. For most studies, multiple imputation

Assumptions underlying a linear regression model

- Errors are **normally distributed**
- Errors are **independent**
- Errors are **equally distributed** across the range of fitted values (i.e., homogeneity of variance)
- **Form** of the relationship between the IV(s) and DV is **correctly specified**



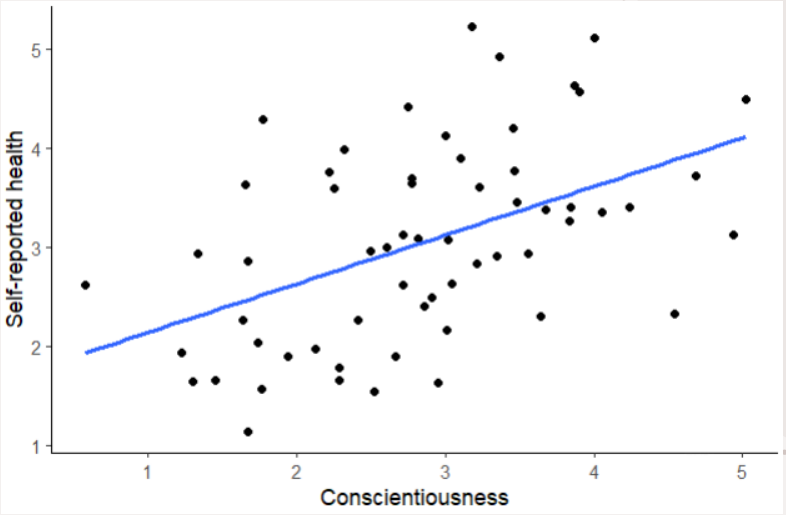




Correctly Specified Form Assumption

- An assumption of fitting a linear regression model to the relationship between a set of predictor(s) and outcome variable is that the **form of the relationship has been correctly identified**.

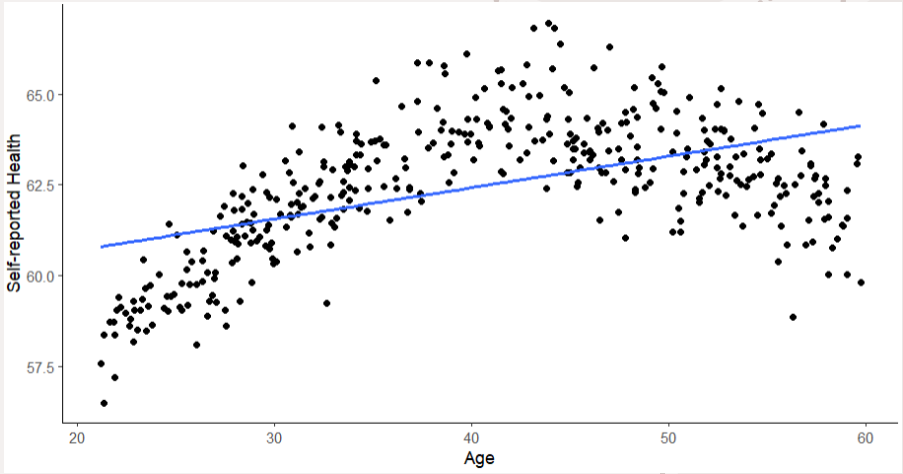
- Most often, we're using our model to fit a linear pattern to the data.

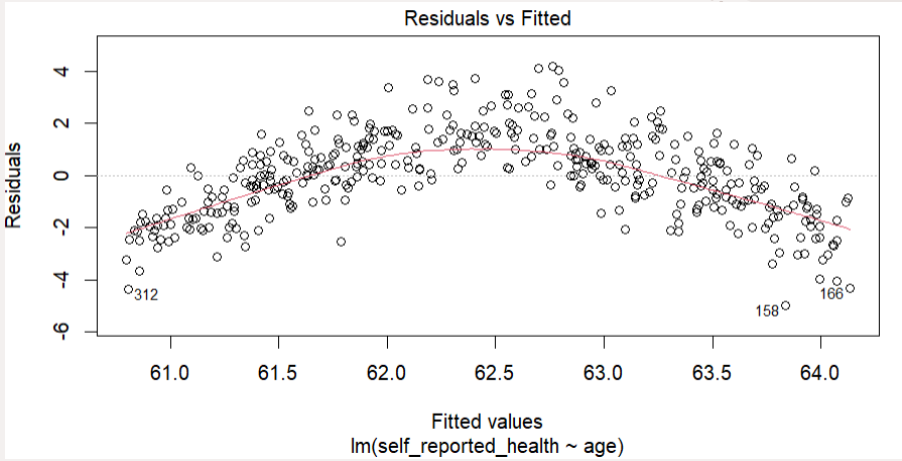


Identifying Incorrectly Specified Form

- The best way to assess whether the assumption of linearity is fair to make, look at a visualization of the relationship between your model's variables.

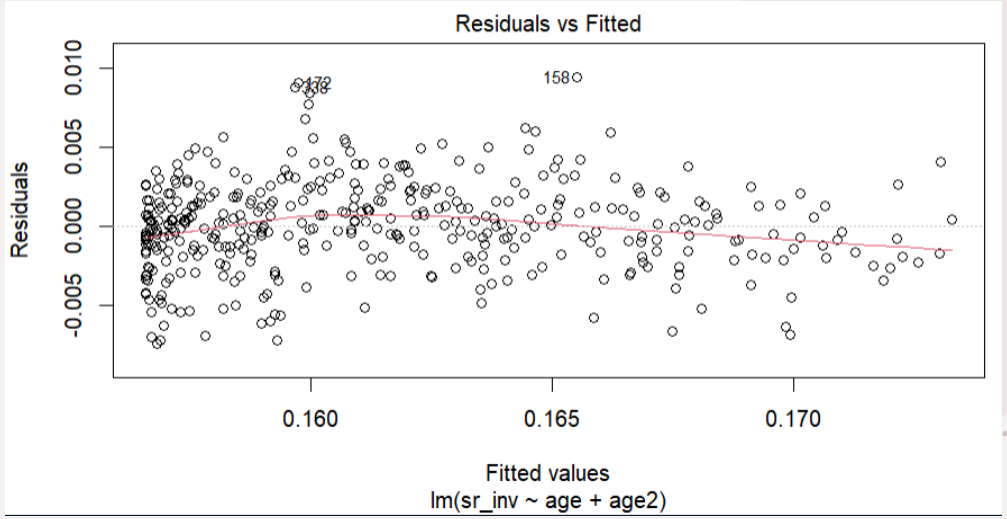
- After you've fit your model, a **residuals plot** can be used to diagnose whether you've correctly specified the form of the relationship between the predictor(s) and outcome.
  - If the form has been incorrectly specified, a systematic pattern should show up in the residuals plot.





Handling Incorrectly Specified Form

- To handle an incorrectly specified form, you can respecify a model that fits a curvilinear relationship between the predictor(s) and outcome that fits the pattern of the data.
  - On the right is the residuals plot produced from a model in which self-reported health scores were predicted from a quadratic relationship with age.

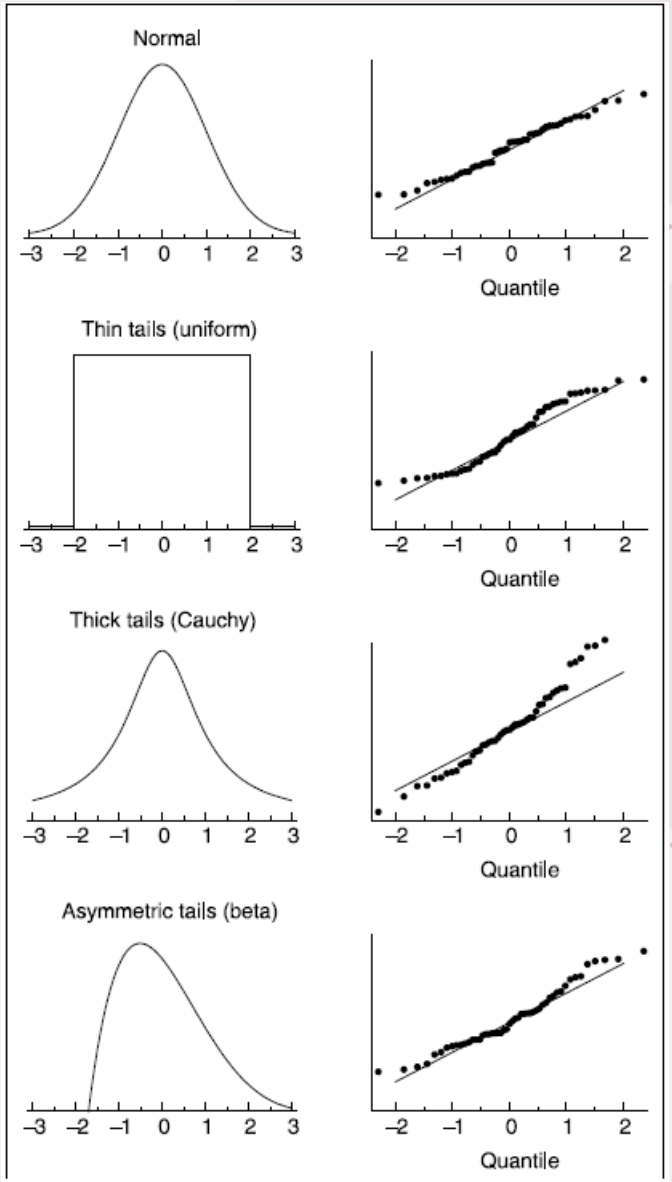


- We don't get to discuss many non-linear models in this course (luckily, the relationships between psychological phenomena often fit the linearity assumption), but you'll discuss one – logistic regression – in 613.

Normality Assumption

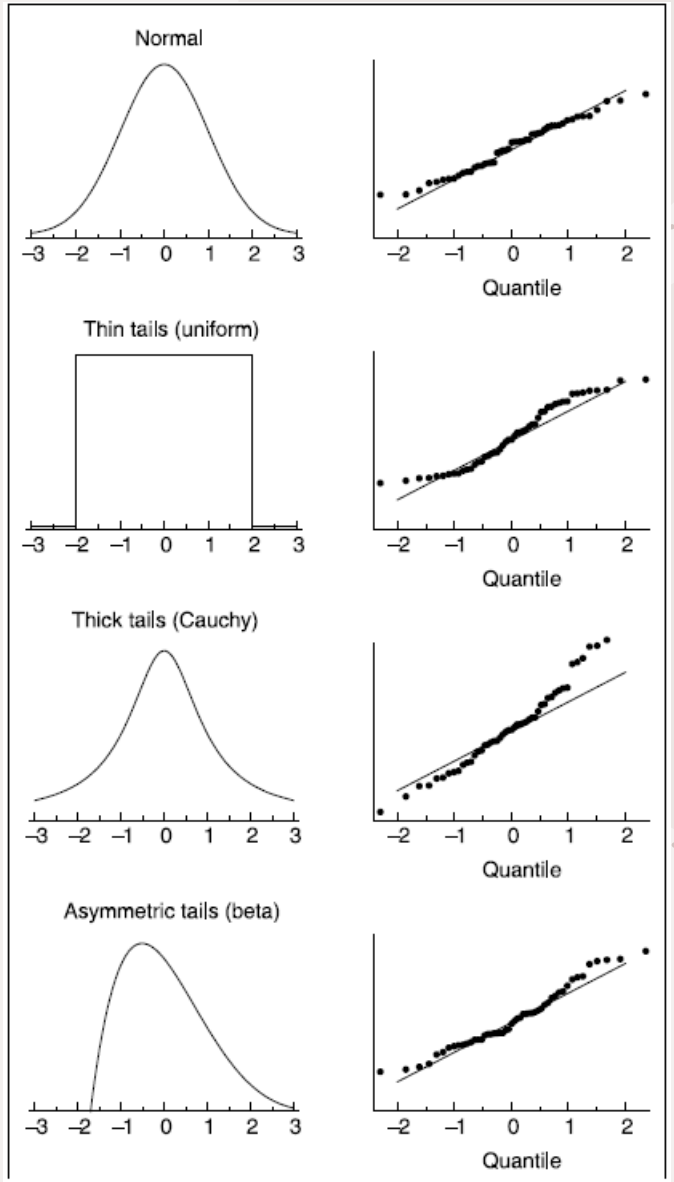
- Another assumption underlying the linear regression analysis is that **residuals are normally distributed**.

- We can diagnose **non-normality** by examining either 1) a distribution of the residuals, or 2) a QQ-plot that represents how the data is distributed compared to what would be expected for a normal distribution.
  - Example QQ-plots for different types of distributed variables are shown to the right.



Identifying Non-normality

- **Normal distribution**: The points are close to lying on the diagonal reference line on the QQ-plot
- **Thin tails:** The slope of the points on the QQ-plot is flat at either end.
- **Thick tails:** The slope of the points on the QQ-plot is steep at either end.
- **Skewed distribution:** The slope of the points on the QQ-plot is flat at one end and steep at the other.



Handling Non-normality

- The assumption of normality is **robust to violations**, meaning its presence doesn't severely bias the results unless the violation is extreme, especially if the sample size is large.
  - If you are still unsure after visual inspection, you can perform Shapiro-Wilk's normality test which examines whether the distribution of residuals is significantly different from normal (a non-significant *p*-value means the assumption is met).
  - shapiro.test(rstandard(model))

If you decide you need to handle the issue of non-normality:

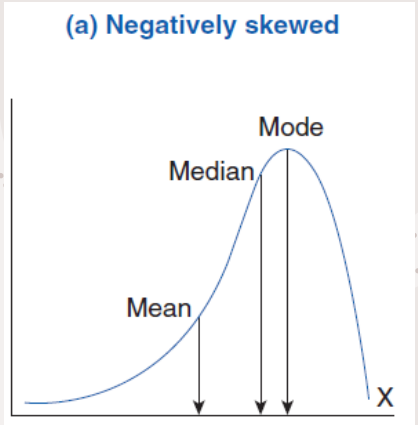
- Non-normality could be driven by the presence of outliers.
- Non-normality could be driven by incorrectly specifying the form of the relationship.
- If non-normality still poses an issue, you can consider performing a transformation on the scores.

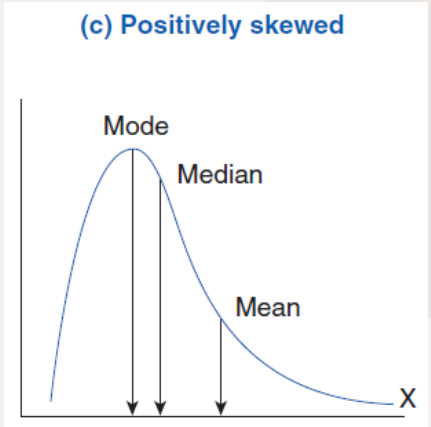
Transformations

Power transformations: transform a variable by raising it to a positive or negative power

- Positive powers spread out scores on the upper end of a variable relative to the lower end (larger values increase by more than smaller values)
  - Used to treat negative skew

- Negative powers spread out scores on the lower end of a variable relative to the upper end (larger values decrease by more than smaller values)
  - Used to treat positive skew
  - Log transformations can also be used to treat positive skew





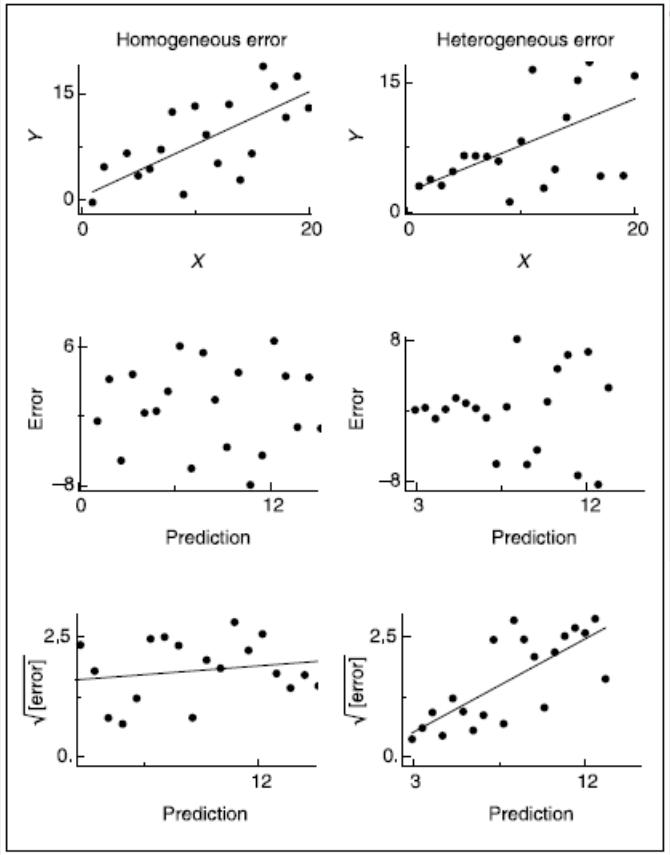
A word of caution about using transformations:

- Changing the nature of the variable that it is you transformed. Interpretation of the results is in light of this transformation. (Ex: A 1-unit increase in age-squared predicted a 0.2 unit increase in self-reported health scores.)

Homogeneity of Variances Assumption

- Another assumption underlying the linear regression analysis is **homogeneity of variances**, which assumes that residuals are evenly distributed across the range of possible values.

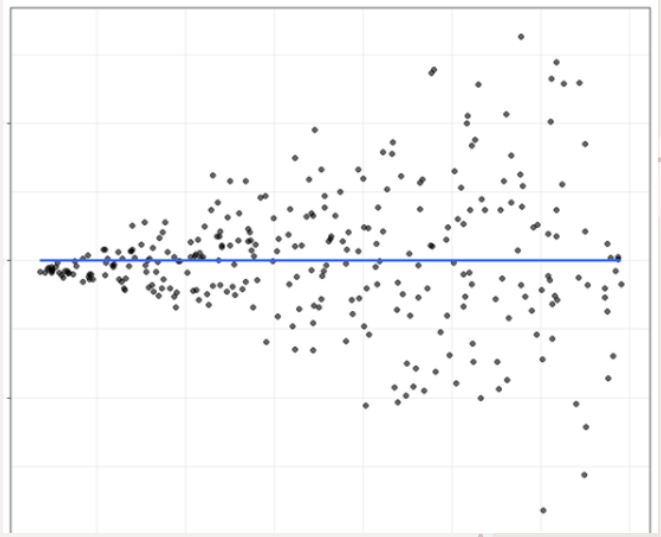
- **Homoscedasticity:** evenly distributed residuals
- **Heteroskedasticity:** unevenly distributed residuals

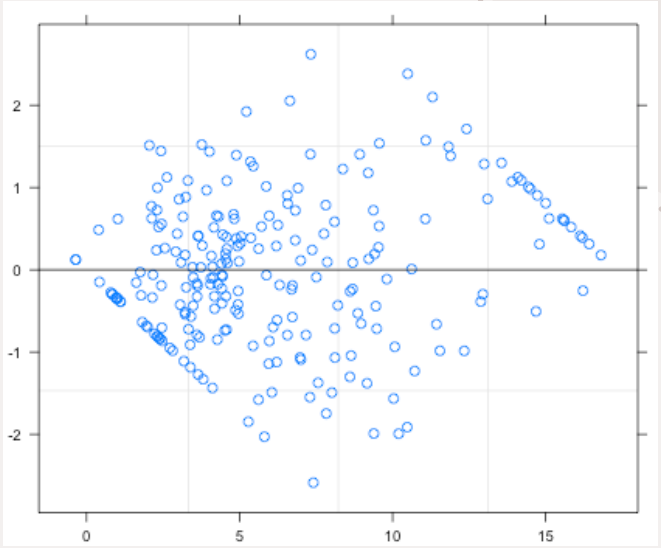


Identifying Heteroskedasticity

- Violation of the **homogeneity of variances** assumption can also be diagnosed using a **residuals plot.**

- Common patterns of **heteroskedasticity** that arise in residuals plots are:
  - Funnel-shaped residuals
    - Can occur when it's easier to give accurate answers to a variable at lower values than at higher values (Ex: guessing the number of objects in a container)
  - Diamond-shaped residuals
    - Can occur when it's easier to give accurate answers at the extreme ends of a variable than in the middle (Ex: being extremely high or low on a personality trait)



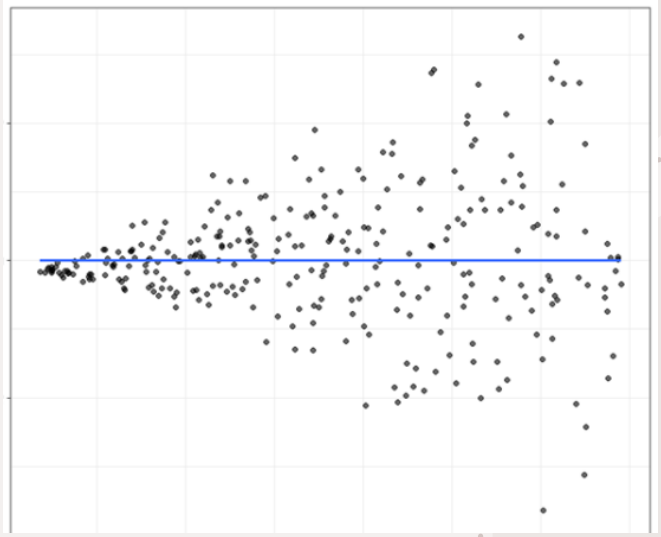


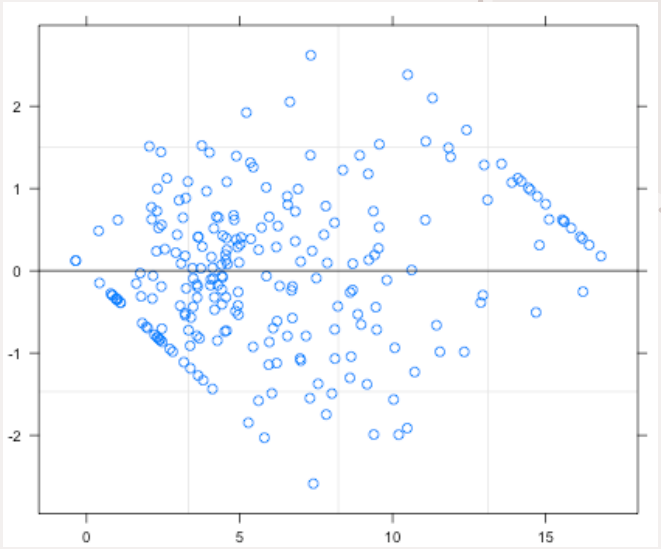
Handling Heteroskedasticity

- Sometimes, handling other issue with the data (e.g., outliers, non-normality, form) can also address the issue of heteroskedasticity.

- If heteroskedasticity continues to be an issue, a common option for handling it is by using **weighted least squares**.
  - When fitting the model using lm(), add a weighting factor that will reduce the weight given to observations with high residuals and increase the weight given to observations with low residuals.

  - May have to trial-and-error to find correct weighting factor
    - For funnel-shaped heteroskedasticity, one option is: weights = $\frac{1}{X^2}$





Independence Assumption

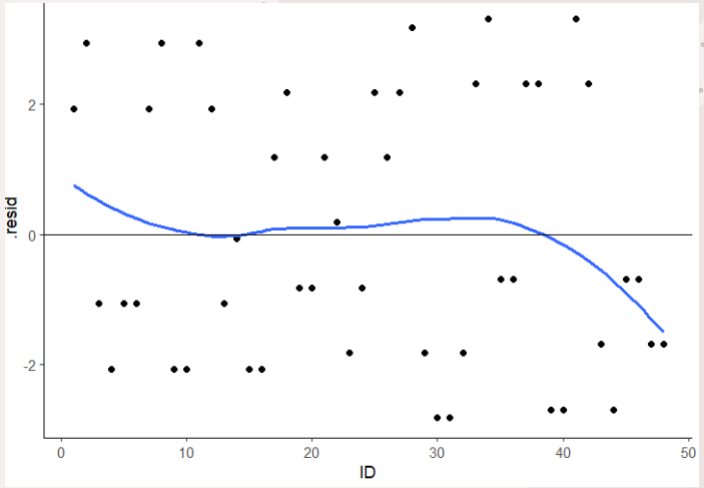
- Another assumption underlying the linear regression analysis is that each participant's error is **independent** of every other participant's error

- The most direct way of knowing whether you have violated the independence assumption is by being aware of the **design used to collect the data**

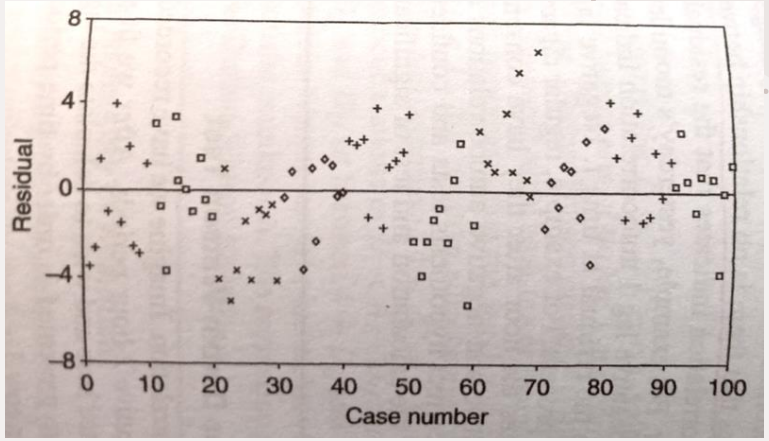
- The errors will not be independent if:
  - Multiple scores on the outcome were taken from the same participants, or if
  - There is a relationship between the participants in your sample

Identifying Non-independence

- If you're unsure of whether the research design has accomplished maintaining independence among participants, a visual way of inspecting whether residuals are non-independent is by **plotting the residuals by ID** (or another variable in the data set that should have no relationship to the residuals).



- Sources of non-independence:
  - Adjacent errors from same participants
  - Adjacent errors from clustered participants



Handling Non-independence

You'll learn two analyses that are intended for non-independent samples of participants in 613.

- Non-independence is okay as long as you are performing an analysis that accounts for it.
  - If you don't account for it, your standard errors will underestimate the actual standard errors (which makes Type I errors more likely).

Longitudinal Analysis (or repeated measures)

- Analyzing scores from participants taken at multiple time points

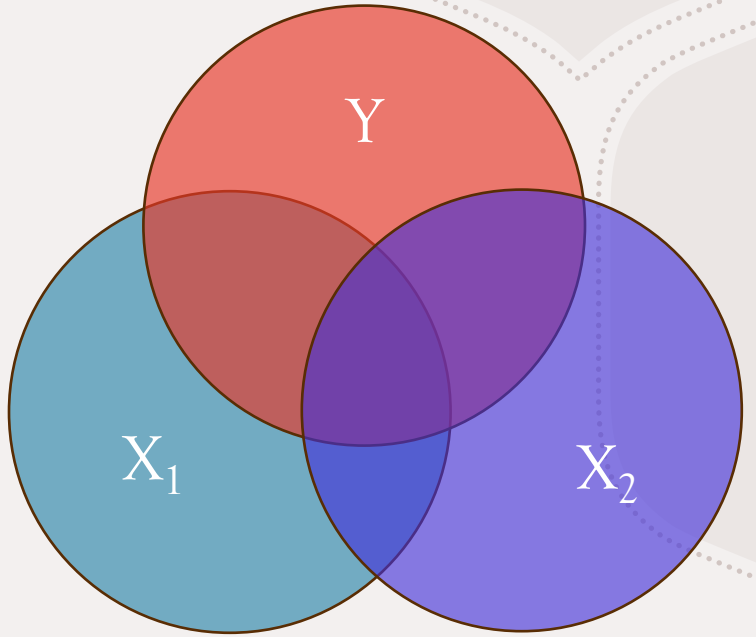
Multilevel Modeling

- Analyzing scores from clustered participants
  - Ex: Comparing how effective an educational intervention is on group students at different elementary schools.

Multicollinearity

- In a linear regression model with **multiple predictors**, the parameter estimate corresponding to any individual predictor is the relationship between that predictor and the outcome variable **controlling for the other predictor(s) in the model**

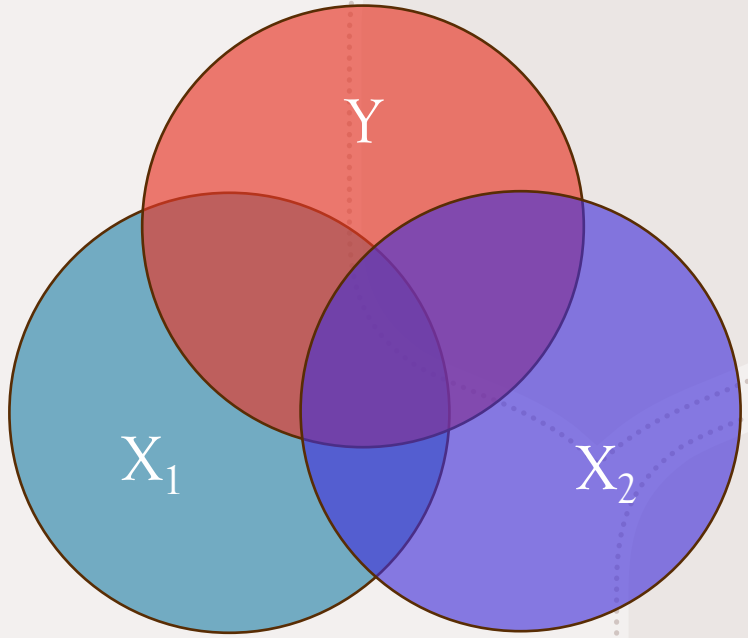
- Each individual predictor is only able to explain variation in the outcome variable *that has been left unexplained by the other predictors in the model.*

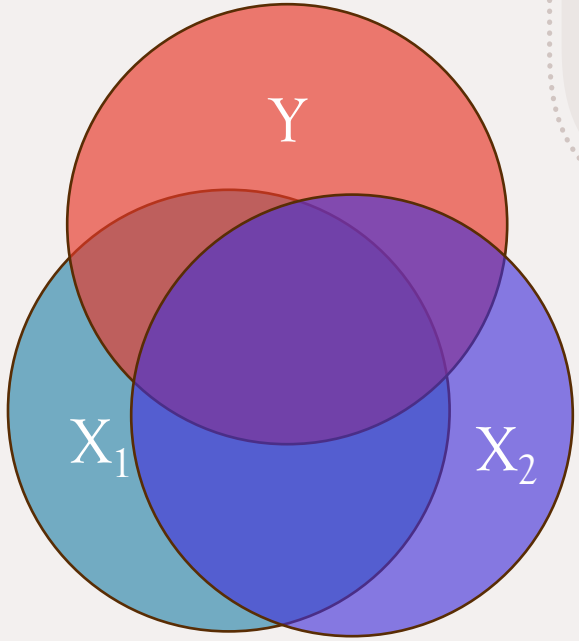


Multicollinearity

- As the correlation between the predictors increases (e.g., r_{12}), the unique variation in that each predictor can explain in Y decreases

- Makes it difficult to assess the unique relationship between each predictor and the outcome variable





Problems with Multicollinearity

- **Parameter estimates are unstable**
  - Could dramatically vary across different models depending on the other predictors in the model
    - Can see this reflected in the formula for the standardized regression coefficient for an individual predictor

- **Standard errors are larger and confidence intervals are wider**
  - Which decreases chances of finding a significant relationship between a predictor and outcome (aka, reduces power)

Standardized regression coefficient for X1 predicting Y:

$$b^*_1 = \frac{r_{Y1} - r_{Y2}r_{12}}{1 - r^2_{12}}$$

Standard error of a regression coefficient, b_X

$$SE_{bx} = \sqrt{\frac{MSE}{(SS_X)(1 - R_p{}^2)}}$$

95%CI for a regression coefficient, b_X

95%CI = $b_X \pm t_{CV}$*SE

Identifying Multicollinearity

Tolerance: a measure of how much a predictor's variance is unique from the other predictors in the model

$$\text{Tolerance} = 1 - R^2_p$$

- where R^2_p is the R-squared value resulting from a model in which a particular predictor, p, is predicted by all the other p-1 predictors in the model

$$\text{or VIF} = \frac{1}{Tolerance}$$

Thresholds for problematic multicollinearity

- Obtain using ols_vif_tol(model) in R
- A **low tolerance** (below 0.20) or a **high VIF** (above 5 or 10) indicates a problem with multicollinearity

Example of Multicollinearity

- Say we're interested in predicting `happiness` from how `extraverted` people are and their levels of `social_engagement`.

- Always a good idea to examine a **correlation matrix** demonstrating the nature of the zero-order correlations

```
> data_multicoll %>%
+   dplyr::select(happiness, extraversion, social_engagement) %>%
+   cor()
                happiness extraversion social_engagement
happiness       1.0000000    0.5591066         0.5341360
extraversion    0.5591066    1.0000000         0.9619373
social_engagement 0.5341360    0.9619373         1.0000000
```

Example of Multicollinearity

- Say we're interested in predicting `happiness` from how `extraverted` people are and their levels of `social_engagement`.

- Fitting a model and examining the multicollinearity diagnostics:

```
model <- lm(happiness ~ extraversion + social_engagement, data = data_multicoll)
```

```
> ols_vif_tol(model)
          Variables  Tolerance      VIF
1      extraversion 0.07467659 13.39108
2 social_engagement 0.07467659 13.39108
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        46.5389     3.2641  14.258   <2e-16 ***
extraversion        0.4280     0.1784   2.400   0.0177 *
social_engagement  -0.0322     0.1648  -0.195   0.8453
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.855 on 144 degrees of freedom
Multiple R-squared:  0.3128,    Adjusted R-squared:  0.3032
F-statistic: 32.77 on 2 and 144 DF,  p-value: 1.865e-12
```

Tolerance is too low (below 0.20) and VIF is too high (above 5).

Handling Multicollinearity

Some options for handling multicollinearity include:

- **Option 1:** Combine highly correlated predictors into a single variable
  - Give the new variable a name that captures the conceptual similarities between the individual predictors

- **Option 2:** Remove one of the predictors from the model
  - Keep the predictor in the model that is more crucial for testing your research question
  - Not an ideal option since we lose information, and *all* of the predictors may be important to your research question

Option 1: Combining highly correlated predictors into a single variable

Example:

```
data_multicoll <- data_multicoll %>%
  mutate(ext_std = scale(extraversion, center = TRUE, scale = TRUE),
         soc_eng_std = scale(social_engagement, center = TRUE, scale = TRUE),
         sociality = ((ext_std + soc_eng_std)/2))
```

```
model_sociality <- lm(happiness ~ sociality, data = data_multicoll)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  69.7143     0.8148  85.558  < 2e-16 ***
sociality     6.5787     0.8255   7.969 4.31e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.879 on 145 degrees of freedom
Multiple R-squared:  0.3046,    Adjusted R-squared:  0.2998
F-statistic: 63.51 on 1 and 145 DF,  p-value: 4.311e-13
```

* May wish to standardize scores on each predictor before combining them if they are measured on different scales.

Option 2: Removing the less important predictor

Example:

```
model_remove <- lm(happiness ~ extraversion, data = data_multicoll)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  46.29039    2.99619   15.45  < 2e-16 ***
extraversion  0.39451    0.04858    8.12 1.84e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.822 on 145 degrees of freedom
Multiple R-squared:  0.3126,    Adjusted R-squared:  0.3079
F-statistic: 65.94 on 1 and 145 DF,  p-value: 1.838e-13
```