Course Business

• New dataset on Canvas

• R package you may want to install: mice

• Midterm grades & feedback posted
  • Excellent job! Mean: 96%

• Final project requirements to be posted by Monday
Week 10.2: Missing Data

- Rank Deficiency
- Linear Combinations
- Incomplete Designs
- Missing Data (NA values)
  - Types of Missingness
    - Non-Ignorable
    - Ignorable
    - Summary
  - Possible Solutions
    - Casewise Deletion
    - Listwise Deletion
    - Unconditional Imputation
    - Conditional Imputation
    - Multiple Imputation
    - Pattern-Mixture Models
**stress.csv**

- Longitudinal (5-week) study of stress
- Dependent measure: Concentration of cortisol, a stress hormone
  - Nanomoles per liter (nmol/L)
- Personality, cognitive, environmental, clinical variables

<table>
<thead>
<tr>
<th>Subject</th>
<th>Week</th>
<th>TempC</th>
<th>TempF</th>
<th>Gregariousness</th>
<th>Assertiveness</th>
<th>ExcitementSeeking</th>
</tr>
</thead>
<tbody>
<tr>
<td>S0001</td>
<td>5</td>
<td>Min. :1</td>
<td>Min. :5.00</td>
<td>Min. :23.00</td>
<td>Min. :1.0</td>
<td>Min. :1.00</td>
</tr>
<tr>
<td>S0002</td>
<td>5</td>
<td>1st Qu.:2</td>
<td>1st Qu.:9.00</td>
<td>1st Qu.:48.20</td>
<td>1st Qu.:2.0</td>
<td>1st Qu.:2.00</td>
</tr>
<tr>
<td>S0003</td>
<td>5</td>
<td>Median :3</td>
<td>Median :14.00</td>
<td>Median :57.20</td>
<td>Median :3.0</td>
<td>Median :4.00</td>
</tr>
<tr>
<td>S0004</td>
<td>5</td>
<td>Mean :3</td>
<td>Mean :13.74</td>
<td>Mean :56.74</td>
<td>Mean :3.4</td>
<td>Mean :3.51</td>
</tr>
<tr>
<td>S0005</td>
<td>5</td>
<td>3rd Qu.:4</td>
<td>3rd Qu.:18.00</td>
<td>3rd Qu.:64.40</td>
<td>3rd Qu.:5.0</td>
<td>3rd Qu.:5.00</td>
</tr>
<tr>
<td>S0006</td>
<td>5</td>
<td>Max. :5</td>
<td>Max. :30.00</td>
<td>Max. :86.00</td>
<td>Max. :6.0</td>
<td>Max. :6.00</td>
</tr>
</tbody>
</table>

(Other): 470

<table>
<thead>
<tr>
<th>Extraversion</th>
<th>Conscientiousness</th>
<th>Anxiety</th>
<th>Severity</th>
<th>ReadingSpan</th>
<th>OperationSpan</th>
<th>CortisolNMol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min. :4.00</td>
<td>Min. :1.000</td>
<td>No :390</td>
<td>No :440</td>
<td>Min. :1.50</td>
<td>Min. :0.50</td>
<td>Min. :13.32</td>
</tr>
<tr>
<td>1st Qu.:8.75</td>
<td>1st Qu.:2.000</td>
<td>Yes:110</td>
<td>Yes:60</td>
<td>1st Qu.:2.50</td>
<td>1st Qu.:2.00</td>
<td>1st Qu.:19.54</td>
</tr>
<tr>
<td>Median :11.00</td>
<td>Median :4.000</td>
<td></td>
<td></td>
<td>Median :3.50</td>
<td>Median :4.00</td>
<td>Median :22.01</td>
</tr>
<tr>
<td>Mean :10.40</td>
<td>Mean :3.604</td>
<td></td>
<td></td>
<td>Mean :3.97</td>
<td>Mean :3.79</td>
<td>Mean :21.74</td>
</tr>
<tr>
<td>3rd Qu.:12.00</td>
<td>3rd Qu.:5.000</td>
<td></td>
<td></td>
<td>3rd Qu.:5.50</td>
<td>3rd Qu.:5.00</td>
<td>3rd Qu.:23.90</td>
</tr>
<tr>
<td>Max. :16.00</td>
<td>Max. :6.000</td>
<td></td>
<td></td>
<td>Max. :6.50</td>
<td>Max. :9.00</td>
<td>Max. :31.35</td>
</tr>
</tbody>
</table>

NA's :45
NA's :19
Rank Deficiency

• Let’s examine cortisol as a function of both:
  • An external factor: Temperature in °C
  • An internal factor: Excitement seeking, on a scale of 1 to 7)

• `model1 <- lmer(CortisolNMol ~ 1 + TempC + ExcitementSeeking + (1|Subject), data=stress)`

<table>
<thead>
<tr>
<th>Random effects:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups Name</td>
</tr>
<tr>
<td>Subject (Intercept)</td>
</tr>
<tr>
<td>Residual</td>
</tr>
<tr>
<td>Number of obs: 500, groups: Subject, 100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fixed effects:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
</tr>
<tr>
<td>(Intercept)</td>
</tr>
<tr>
<td>TempC</td>
</tr>
<tr>
<td>ExcitementSeeking</td>
</tr>
</tbody>
</table>
Rank Deficiency

• We also have some other measures
• Let’s try adding temperature in Fahrenheit to the model (TempF)

• `model2 <- lmer(CortisolNMol ~ 1 + TempC + TempF + ExcitementSeeking + (1|Subject), data=stress)`

• This looks scary!
Rank Deficiency

• Our model:

\[ E(Y_{i(j)}) = Y_{000} + Y_{100}x_{1i(j)} + Y_{200}x_{2i(j)} + Y_{300}x_{3i(j)} \]

  Stress  Baseline  Temperature in Celsius  Temperature in Fahrenheit  Excitement seeking

• What does \( Y_{100} \) represent here?
  • The effect of a 1-unit change in degrees Celsius … while holding degrees Fahrenheit constant
  • This makes no sense—if °C changes, so does °F
  • In fact, one change perfectly predicts the other
    • °F = (9/5)°C + 32
  • Problem if one column is a linear combination of other(s)—it can be perfectly formed from other columns by adding, subtracting, multiplying, or dividing
Rank Deficiency

• Linear combinations result in a perfect correlation

• Here: Correlation between TempC and TempF
**Rank Deficiency**

Common response: “R won’t do what I want! I need to fit this model, but I can’t! This program is broken!”

Scott’s view: This really isn’t a coherent research question

- “Effect of changing °C while holding constant °F” doesn’t make sense to ask—can’t be answered
- We would have this same problem in any software package, or even if we computed the regression by hand
Rank Deficiency: Solutions

• Is it *unexpected* that one column perfectly predicts another?
  • Check your experiment script & data processing pipeline (e.g., is something saved in the wrong column?)

• Or is it *expected*? (*TempF* should be predictable from *TempC*):
  • Rethink analysis—would it ever be sensible to try to distinguish these effects?
    • Maybe this is just not a coherent research question
    • Or, a different design might make these independent
Rank Deficiency

• We have some other personality variables:
  • Gregariousness
  • Assertiveness
  • Extraversion, the sum of the gregariousness, assertiveness, and excitement seeking facets

• Our next model:
  • `modelExtra <- lmer(CortisolNMol ~ 1 + TempC + Gregariousness + Assertiveness + ExcitementSeeking + Extraversion + (1|Subject), data=stress)`

fixed-effect model matrix is rank deficient so dropping 1 column / coefficient

• Where is the rank deficiency in this case?
**Rank Deficiency**

- “Problem if one column is a **linear combination** of other(s)—it can be perfectly formed from other columns by adding, subtracting, multiplying, or dividing”
- Similar problem: One predictor variable is the *mean* or *sum* of other predictors in the model

<table>
<thead>
<tr>
<th>Gregariousness</th>
<th>Assertiveness</th>
<th>ExcitementSeeking</th>
<th>Extraversion</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

- We already know exactly what **Extraversion** is when we know the three facet scores
- Doesn’t make sense to ask about the effect of **Extraversion** while holding constant **Gregariousness**, **Assertiveness**, and **ExcitementSeeking**
Rank Deficiency

• “Problem if one column is a **linear combination** of other(s)—it can be perfectly formed from other columns by adding, subtracting, multiplying, or dividing”

• Similar problem: One predictor variable is the **mean** or **sum** of other predictors in the model

<table>
<thead>
<tr>
<th>Gregariousness</th>
<th>Assertiveness</th>
<th>ExcitementSeeking</th>
<th>Extraversion</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

• Can also get this with the **average** of other variables
  • Exam 1 score, Exam 2 score, Average score
  • Average = (Exam1 + Exam2) / 2
  • Again, can be perfectly predicted from the other columns
Week 10.2: Missing Data

- Rank Deficiency
- Linear Combinations
- Incomplete Designs
- Missing Data (NA values)
  - Types of Missingness
    - Non-Ignorable
    - Ignorable
    - Summary
  - Possible Solutions
    - Casewise Deletion
    - Listwise Deletion
    - Unconditional Imputation
    - Conditional Imputation
    - Multiple Imputation
    - Pattern-Mixture Models
Incomplete Designs

- The last variable that we’re interested in is anxiety
- Two relevant columns:
  - **Anxiety**: Does this person have a diagnosis of GAD (Generalized Anxiety Disorder) or not?
  - **Severity**: Is the anxiety severe or not?
- Let’s look at these two factors and their interaction:
  - `model4 <- lmer(CortisolNMol ~ 1 + Anxiety * Severity + (1|Subject), data=stress)`

fixed-effect model matrix is rank deficient so dropping 1 column / coefficient
Incomplete Designs

- Why is this model rank-deficient?
- Some cells of the interaction are not represented in our current design:
  - `xtabs(~ Anxiety + Severity, data=stress)`

No observations for people with No anxiety and Severe symptoms
Incomplete Designs

• Again, not a bug. Doesn’t make sense to ask about anxiety*severity interaction here

Anxiety: No  Severity: No

Anxiety: Yes  Severity: No

Anxiety: Yes  Severity: Yes

Anxiety effect

Severity effect
Incomplete Designs

• Again, not a bug. Doesn’t make sense to ask about anxiety*severity interaction here

• Interaction of Anxiety & Severity: “Effect of Anxiety and Severe anxiety over and above the effects of Anxiety alone and Severe anxiety alone.”
  • But, no effect of severe anxiety “alone.” Have to have anxiety to have severe anxiety.
**Incomplete Designs: Solutions**

- If missing cell is *intended*:
  - Often makes more sense to think of this as a single factor with > 2 categories
    - “No anxiety,” “Moderate anxiety,” “Severe anxiety”
    - Can use Contrast 1 to compare Moderate & Severe to Not Anxious, and Contrast 2 to compare Moderate to Severe

- If missing cell is *accidental*:
  - Might need to collect more data!
  - Check your experimental lists
Week 10.2: Missing Data

- Rank Deficiency
- Linear Combinations
- Incomplete Designs
- Missing Data (NA values)

Types of Missingness
- Non-Ignorable
- Ignorable
- Summary

Possible Solutions
- Casewise Deletion
- Listwise Deletion
- Unconditional Imputation
- Conditional Imputation
- Multiple Imputation
- Pattern-Mixture Models
So far, we’ve looked at cases where:

- Two predictor variables are *always* confounded
- A particular category of the dependent variable *never* (or *almost never*) appears in a particular cell

But lots of cases where some observations are missing more haphazardly
### Missing Data

- Lots of cases where a model makes sense in principle, but part of the dataset is missing
  - Computer crashes
  - Some people didn’t respond to all items or prompts
  - Participants dropped out (**loss to follow up**)
  - Implausible values excluded
  - Non-codable data (e.g., we’re looking at whether L2 learners produce the correct plural, but someone says *fish*)
  - Loss of signal (ERP, pitch, eye-tracking, etc.)
Missing Data

- Big issue: Is our sample still **representative**?
- Basic goal in inferential statistics is to generalize from limited sample to a population
  - If sample is truly **random**, this is justified
Missing Data

• Problem: If data from certain types of people *always* go missing, our sample will no longer be **representative** of the broader population
• It’s not a random sample if we systematically lose certain kinds of data
**Missing Data**

- In fact, it’s a problem even if certain types of people are *somewhat* more likely to have missing data
- Still not a fully random sample
Big Issue: WHY data is missing

- We will see several techniques for dealing with missing data

- The degree to which these techniques are appropriate depends on how the missing data relates to the other variables
  - Missingness of data may not be arbitrary (and often isn’t)

- Let’s first look at some hypothetical patterns of missing data
  - This is a conceptual distinction
  - In any given actual data set, we might not know the pattern for certain
Week 10.2: Missing Data

- Rank Deficiency
- Linear Combinations
- Incomplete Designs
- Missing Data (NA values)
- Types of Missingness
  - Non-Ignorable
    - Ignorable
    - Summary
- Possible Solutions
  - Casewise Deletion
  - Listwise Deletion
  - Unconditional Imputation
  - Conditional Imputation
  - Multiple Imputation
  - Pattern-Mixture Models
Hypothetical Scenario #1

- Sometimes, the fact that a data point is NA may be related to *what the value would have been ... if we’d been able to measure it*

"If only..."
**Hypothetical Scenario #1**

- Sometimes, the fact that a data point is NA may be related to *what the value would have been … if we’d been able to measure it*.
- A health psychologist is surveying high school students about their marijuana use.
  - Students who’ve tried marijuana may be more likely *to leave this question blank* than those who haven’t.
  - Remaining data is a biased sample.

<table>
<thead>
<tr>
<th>ACTUAL STATE OF WORLD</th>
<th>WHAT WE SEE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>No</td>
<td>NA</td>
</tr>
<tr>
<td>Yes</td>
<td>NA</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>NA</td>
</tr>
</tbody>
</table>

= 50% = 33%
Hypothetical Scenario #1

- Sometimes, the fact that a data point is NA may be related to *what the value would have been* … *if we’d been able to measure it*
- In other words, some values are more likely than others to end up as NAs
  - All that’s relevant here is that there is a statistical contingency
  - The actual causal chain might be more complex
    - e.g., marijuana use → fear of legal repercussions → omitted response
Hypothetical Scenario #1

- Further examples:
  - Clinical study where we’re measuring health outcome. People who are very ill might drop out of the study.
  - Experiment where you have to press a key within 3 seconds or the trial ends without a response time being recorded
  - People with low high school GPA decline to report it

- These are all examples of **nonignorable missingness**
Hypothetical Scenario #1

- Nonignorable missingness is bad 😞
  - Remaining observations (those without NAs) are not representative of the full population
  - We can’t fully account for what the missing data were, or why they’re missing
  - We simply don’t know what the missing RT what would have been if people had been allowed more time
Week 10.2: Missing Data

- Rank Deficiency
- Linear Combinations
- Incomplete Designs
- Missing Data (NA values)
- Types of Missingness
  - Non-Ignorable
  - Ignorable
    - Summary
- Possible Solutions
  - Casewise Deletion
  - Listwise Deletion
  - Unconditional Imputation
  - Conditional Imputation
  - Multiple Imputation
  - Pattern-Mixture Models
Hypothetical Scenario #2

- In other cases, data might go missing at “random” or for reasons completely unrelated to the study
  - Computer crash
  - Inclement weather
  - Experimenter error
  - Random subsampling of people for a follow-up
Hypothetical Scenario #2

- In other cases, data might go missing at “random” or for reasons completely unrelated to the study
  - Computer crash
  - Inclement weather
  - Experimenter error
  - Random subsampling of people for a follow-up

- In these cases, there is no reason to think that the missing data would look any different from the remaining data
  - **Ignorable missingness**
Hypothetical Scenario #2

- **Ignorable missingness isn’t nearly as bad! 😊**
  - It’s disappointing that we lost some data, but what’s left is still representative of the population
  - Our data is still a random sample of the population—it’s just a *smaller* random sample
  - Still valid to make inferences about the population
Hypothetical Scenario #3

- Another case of *ignorable* missingness is when the fact that the data is NA can be fully explained by other, *known* variables.

- **Examples:**
  - People who score high on a pretest are excluded from further participation in an intervention study.
  - We’re looking at child SES as a predictor of physical growth, but lower SES families are less likely to return for the post-test.
  - DV is whether people say a plural vs singular noun; we discard ambiguous words (e.g., “fish”). Rate of ambiguous words differs across conditions.
Hypothetical Scenario #3

- Another case of ignorable missingness is when the fact that the data is NA can be fully explained by other, known variables.
- This is also ignorable because there’s no mystery about why the data is missing, nor the values of the variables associated with NAness.
  - We know why the high-pretest people were excluded from the intervention. Has nothing to do with unobserved variables.
- Again, referring to statistical contingencies not direct causal links.
  - Low SES $\rightarrow$ Transportation less affordable $\rightarrow$ NA.
Week 10.2: Missing Data

- Rank Deficiency
  - Linear Combinations
  - Incomplete Designs
- Missing Data (NA values)
  - Types of Missingness
    - Non-Ignorable
    - Ignorable
- Summary

- Possible Solutions
  - Casewise Deletion
  - Listwise Deletion
  - Unconditional Imputation
  - Conditional Imputation
  - Multiple Imputation
  - Pattern-Mixture Models
Wrap-Up

- Ignorableness is more like a continuum

All of the missingness can be accounted for by known variables (IGNORABLE)

All of the missingness depends on unknown values (NONIGNORABLE)

- As long as we’re relatively ignorable, we’re OK
  - “We should expect departures from [ignorable missingness] … in many realistic cases … may often have only a minor impact on estimates and standard errors.” (Schafer & Graham, 2002, p. 152)
  - “In many psychological situations the departures … are probably not serious.” (Schafer & Graham, 2002, p. 154)
Ignorable or Non-ignorable?

- Where is my dataset on this continuum?
  - "In general, there is no way to test whether [ignorable missingness] holds in a data set." (Schafer & Graham, 2002, p. 152)
- Definitely ignorable if you used known variable(s) to decide to discard data or to stop data collection
  - Kids who get a low score on Task 1 aren’t given Task 2
  - People who are sufficiently healthy are excluded from a clinical study
  - We discarded a particular questionnaire item
- Other cases: Use your knowledge of the domain
  - Are certain values less likely to be measured & recorded; e.g., poor health in a clinical study? (non-ignorable)
  - Or, is the missingness basically happening at random? Can it be accounted for by things we measured? (ignorable)
- Big picture: Most departures from ignorable missingness aren’t terrible, but be aware of the possibility that certain values are systematically more likely to be missing
Ignorable or Non-ignorable?

- The post-experiment manipulation-check questionnaires for five participants were accidentally thrown away.

- In a 2-day memory experiment, people who know they would do poorly on the memory test are discouraged and don’t want to return for the second session.

- There was a problem with one of the auditory stimulus files in an ERP study of speech comprehension during noise, so we discarded any data from item #43.
**Ignorable or Non-ignorable?**

- The post-experiment manipulation-check questionnaires for five participants were accidentally thrown away.
  - Ignorable—not related to any variable

- In a 2-day memory experiment, people who know they would do poorly on the memory test are discouraged and don’t want to return for the second session.
  - Non-ignorable. Missingness depends on what your memory score would have been if we had observed it.

- There was a problem with one of the auditory stimulus files in an ERP study of speech comprehension during noise, so we discarded any data from item #43.
  - Ignorable; this depends on a known variable (item #)
Ignorable or Non-ignorable?

- We are comparing life satisfaction among a sample of students known to live on-campus vs. a sample of students known to live off-campus. But students off-campus are less likely to return their surveys because it’s more inconvenient for them to do so.
Ignorable or Non-ignorable?

- We are comparing life satisfaction among a sample of students known to live on-campus vs. a sample of students known to live off-campus. But students off-campus are less likely to return their surveys because it’s more inconvenient for them to do so.
  - Ignorable if missingness depends only on this known variable. Fewer off-campus students might return their surveys, but the off-campus students from whom we have data don’t differ from the off-campus students for whom we don’t have data.
  - If we think that there is also a relation to the unmeasured life-satisfaction variable (e.g., people unhappy with their lives don’t return the survey), although not mentioned above, then this would be non-ignorable.
  - Assumptions we are willing to make about the missing data (and why it’s missing) affect how we can then use it.
Week 10.2: Missing Data

- Rank Deficiency
- Linear Combinations
- Incomplete Designs
- Missing Data (NA values)
- Types of Missingness
  - Non-Ignorable
  - Ignorable
- Summary

Possible Solutions
- Casewise Deletion
- Listwise Deletion
- Unconditional Imputation
- Conditional Imputation
- Multiple Imputation
- Pattern-Mixture Models
### Missing Data Solutions

- Our data has some **NAs**

<table>
<thead>
<tr>
<th>Subject</th>
<th>Week</th>
<th>TempC</th>
<th>Extraversion</th>
<th>Conscientiousness</th>
<th>Anxiety Severity</th>
<th>Reading Span</th>
<th>Operation Span</th>
<th>Cortisol NMol</th>
</tr>
</thead>
<tbody>
<tr>
<td>350 S0070</td>
<td>5</td>
<td>4</td>
<td>12</td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>4.5</td>
</tr>
<tr>
<td>351 S0071</td>
<td>1</td>
<td>9</td>
<td>7</td>
<td></td>
<td></td>
<td>No</td>
<td>No</td>
<td>6.5</td>
</tr>
<tr>
<td>352 S0071</td>
<td>2</td>
<td>4</td>
<td>7</td>
<td></td>
<td></td>
<td>No</td>
<td>No</td>
<td>5.5</td>
</tr>
<tr>
<td>353 S0071</td>
<td>3</td>
<td>19</td>
<td>7</td>
<td></td>
<td></td>
<td>No</td>
<td>No</td>
<td>6.0</td>
</tr>
<tr>
<td>354 S0071</td>
<td>4</td>
<td>13</td>
<td>7</td>
<td></td>
<td></td>
<td>No</td>
<td>No</td>
<td>6.5</td>
</tr>
<tr>
<td>355 S0071</td>
<td>5</td>
<td>2</td>
<td>7</td>
<td></td>
<td></td>
<td>No</td>
<td>No</td>
<td>6.0</td>
</tr>
<tr>
<td>356 S0072</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td>No</td>
<td>No</td>
<td>5.5</td>
</tr>
<tr>
<td>357 S0072</td>
<td>2</td>
<td>7</td>
<td>3</td>
<td></td>
<td></td>
<td>No</td>
<td>No</td>
<td>6.5</td>
</tr>
<tr>
<td>358 S0072</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td>No</td>
<td>No</td>
<td>NA</td>
</tr>
<tr>
<td>359 S0072</td>
<td>4</td>
<td>27</td>
<td>3</td>
<td></td>
<td></td>
<td>No</td>
<td>No</td>
<td>6.0</td>
</tr>
<tr>
<td>360 S0072</td>
<td>5</td>
<td>15</td>
<td>3</td>
<td></td>
<td></td>
<td>No</td>
<td>No</td>
<td>5.5</td>
</tr>
<tr>
<td>361 S0073</td>
<td>1</td>
<td>11</td>
<td>4</td>
<td></td>
<td></td>
<td>No</td>
<td>No</td>
<td>5.5</td>
</tr>
<tr>
<td>362 S0073</td>
<td>2</td>
<td>21</td>
<td>4</td>
<td></td>
<td></td>
<td>No</td>
<td>No</td>
<td>6.5</td>
</tr>
<tr>
<td>363 S0073</td>
<td>3</td>
<td>17</td>
<td>4</td>
<td></td>
<td></td>
<td>No</td>
<td>No</td>
<td>5.5</td>
</tr>
<tr>
<td>364 S0073</td>
<td>4</td>
<td>11</td>
<td>4</td>
<td></td>
<td></td>
<td>No</td>
<td>No</td>
<td>6.0</td>
</tr>
<tr>
<td>365 S0073</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td>No</td>
<td>No</td>
<td>6.0</td>
</tr>
</tbody>
</table>
Missing Data Solutions

- `library(mice)`
- `stress %>% md.pattern()`
- Each row = one pattern of missingness

Number of people who have this pattern
### Missing Data Solutions

- Our data has some **NA**s
- What are some possible solutions?

<table>
<thead>
<tr>
<th>Subject</th>
<th>Week</th>
<th>TempC</th>
<th>Extraversion</th>
<th>Conscientiousness</th>
<th>Anxiety Severity</th>
<th>Span</th>
<th>Operation</th>
<th>CortisolN Mol</th>
</tr>
</thead>
<tbody>
<tr>
<td>350</td>
<td>S0070</td>
<td>5</td>
<td>4</td>
<td>12</td>
<td>6</td>
<td>Yes</td>
<td>No</td>
<td>4.5</td>
</tr>
<tr>
<td>351</td>
<td>S0071</td>
<td>1</td>
<td>9</td>
<td>7</td>
<td>NA</td>
<td>No</td>
<td>No</td>
<td>6.5</td>
</tr>
<tr>
<td>352</td>
<td>S0071</td>
<td>2</td>
<td>4</td>
<td>7</td>
<td>NA</td>
<td>No</td>
<td>No</td>
<td>5.5</td>
</tr>
<tr>
<td>353</td>
<td>S0071</td>
<td>3</td>
<td>19</td>
<td>7</td>
<td>NA</td>
<td>No</td>
<td>No</td>
<td>6.0</td>
</tr>
<tr>
<td>354</td>
<td>S0071</td>
<td>4</td>
<td>13</td>
<td>7</td>
<td>NA</td>
<td>No</td>
<td>No</td>
<td>6.5</td>
</tr>
<tr>
<td>355</td>
<td>S0071</td>
<td>5</td>
<td>2</td>
<td>7</td>
<td>NA</td>
<td>No</td>
<td>No</td>
<td>6.0</td>
</tr>
<tr>
<td>356</td>
<td>S0072</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>3</td>
<td>No</td>
<td>No</td>
<td>5.5</td>
</tr>
<tr>
<td>357</td>
<td>S0072</td>
<td>2</td>
<td>7</td>
<td>9</td>
<td>3</td>
<td>No</td>
<td>No</td>
<td>6.5</td>
</tr>
<tr>
<td>358</td>
<td>S0072</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>3</td>
<td>No</td>
<td>No</td>
<td>4.5</td>
</tr>
<tr>
<td>359</td>
<td>S0072</td>
<td>4</td>
<td>27</td>
<td>9</td>
<td>3</td>
<td>No</td>
<td>No</td>
<td>6.0</td>
</tr>
<tr>
<td>360</td>
<td>S0072</td>
<td>5</td>
<td>15</td>
<td>9</td>
<td>3</td>
<td>No</td>
<td>No</td>
<td>5.5</td>
</tr>
<tr>
<td>361</td>
<td>S0073</td>
<td>1</td>
<td>11</td>
<td>9</td>
<td>4</td>
<td>No</td>
<td>No</td>
<td>5.5</td>
</tr>
<tr>
<td>362</td>
<td>S0073</td>
<td>2</td>
<td>21</td>
<td>9</td>
<td>4</td>
<td>No</td>
<td>No</td>
<td>6.5</td>
</tr>
<tr>
<td>363</td>
<td>S0073</td>
<td>3</td>
<td>17</td>
<td>9</td>
<td>4</td>
<td>No</td>
<td>No</td>
<td>5.5</td>
</tr>
<tr>
<td>364</td>
<td>S0073</td>
<td>4</td>
<td>11</td>
<td>9</td>
<td>4</td>
<td>No</td>
<td>No</td>
<td>6.0</td>
</tr>
<tr>
<td>365</td>
<td>S0073</td>
<td>5</td>
<td>3</td>
<td>9</td>
<td>4</td>
<td>No</td>
<td>No</td>
<td>6.0</td>
</tr>
</tbody>
</table>
Week 10.2: Missing Data

- Rank Deficiency
- Linear Combinations
- Incomplete Designs
- Missing Data (NA values)
- Types of Missingness
  - Non-Ignorable
  - Ignorable
- Summary
- Possible Solutions
  - Casewise Deletion
    - Listwise Deletion
    - Unconditional Imputation
    - Conditional Imputation
    - Multiple Imputation
    - Pattern-Mixture Models
**Casewise Deletion**

- In each comparison, delete only observations if the missing data is relevant to *this comparison*.

```
<table>
<thead>
<tr>
<th>Subject</th>
<th>Week</th>
<th>TempC</th>
<th>Extraversion</th>
<th>Conscientiousness</th>
<th>Anxiety</th>
<th>Severity</th>
<th>ReadingSpan</th>
<th>OperationSpan</th>
<th>Cortisol</th>
<th>N Mol</th>
</tr>
</thead>
<tbody>
<tr>
<td>350</td>
<td>S0070</td>
<td>5</td>
<td>4</td>
<td>12</td>
<td>6</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>4.5</td>
<td>3.5</td>
</tr>
<tr>
<td>351</td>
<td>S0071</td>
<td>1</td>
<td>9</td>
<td>7</td>
<td>NA</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>6.5</td>
<td>5.0</td>
</tr>
<tr>
<td>352</td>
<td>S0071</td>
<td>2</td>
<td>4</td>
<td>7</td>
<td>NA</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>5.5</td>
<td>3.0</td>
</tr>
<tr>
<td>353</td>
<td>S0071</td>
<td>3</td>
<td>19</td>
<td>7</td>
<td>NA</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>6.0</td>
<td>6.0</td>
</tr>
<tr>
<td>354</td>
<td>S0071</td>
<td>4</td>
<td>13</td>
<td>7</td>
<td>NA</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>6.5</td>
<td>7.0</td>
</tr>
<tr>
<td>355</td>
<td>S0071</td>
<td>5</td>
<td>2</td>
<td>7</td>
<td>NA</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>6.0</td>
<td>2.0</td>
</tr>
<tr>
<td>356</td>
<td>S0072</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>3</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>5.5</td>
<td>5.5</td>
</tr>
<tr>
<td>357</td>
<td>S0072</td>
<td>2</td>
<td>7</td>
<td>9</td>
<td>3</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>6.5</td>
<td>4.5</td>
</tr>
<tr>
<td>358</td>
<td>S0072</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>3</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>NA</td>
<td>4.5</td>
</tr>
<tr>
<td>359</td>
<td>S0072</td>
<td>4</td>
<td>27</td>
<td>9</td>
<td>3</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>6.0</td>
<td>5.5</td>
</tr>
<tr>
<td>360</td>
<td>S0072</td>
<td>5</td>
<td>15</td>
<td>9</td>
<td>3</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>5.5</td>
<td>5.5</td>
</tr>
<tr>
<td>361</td>
<td>S0073</td>
<td>1</td>
<td>11</td>
<td>9</td>
<td>4</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>5.5</td>
<td>3.5</td>
</tr>
<tr>
<td>362</td>
<td>S0073</td>
<td>2</td>
<td>21</td>
<td>9</td>
<td>4</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>6.5</td>
<td>5.0</td>
</tr>
<tr>
<td>363</td>
<td>S0073</td>
<td>3</td>
<td>17</td>
<td>9</td>
<td>4</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>5.5</td>
<td>4.5</td>
</tr>
<tr>
<td>364</td>
<td>S0073</td>
<td>4</td>
<td>11</td>
<td>9</td>
<td>4</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>6.0</td>
<td>4.0</td>
</tr>
<tr>
<td>365</td>
<td>S0073</td>
<td>5</td>
<td>3</td>
<td>9</td>
<td>4</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>6.0</td>
<td>8.5</td>
</tr>
</tbody>
</table>
```

- Correlating Extraversion & Conscientiousness
  - delete/ignore the red rows

**Casewise Deletion**

- In each comparison, delete only observations if the missing data is relevant to *this comparison*

### Data Table

<table>
<thead>
<tr>
<th>Subject</th>
<th>Week</th>
<th>TempC</th>
<th>Extraversion</th>
<th>Conscientiousness</th>
<th>Anxiety Severity</th>
<th>ReadingSpan</th>
<th>OperationSpan</th>
<th>CortisolNMol</th>
</tr>
</thead>
<tbody>
<tr>
<td>350</td>
<td>S0070</td>
<td>5</td>
<td>4</td>
<td>12</td>
<td>6 Yes</td>
<td>No</td>
<td>No</td>
<td>4.5</td>
</tr>
<tr>
<td>351</td>
<td>S0071</td>
<td>1</td>
<td>9</td>
<td>7</td>
<td>NA No</td>
<td>No</td>
<td>No</td>
<td>6.5</td>
</tr>
<tr>
<td>352</td>
<td>S0071</td>
<td>2</td>
<td>4</td>
<td>7</td>
<td>NA No</td>
<td>No</td>
<td>No</td>
<td>5.5</td>
</tr>
<tr>
<td>353</td>
<td>S0071</td>
<td>3</td>
<td>19</td>
<td>7</td>
<td>NA No</td>
<td>No</td>
<td>No</td>
<td>6.0</td>
</tr>
<tr>
<td>354</td>
<td>S0071</td>
<td>4</td>
<td>13</td>
<td>7</td>
<td>NA No</td>
<td>No</td>
<td>No</td>
<td>6.5</td>
</tr>
<tr>
<td>355</td>
<td>S0071</td>
<td>5</td>
<td>2</td>
<td>7</td>
<td>NA No</td>
<td>No</td>
<td>No</td>
<td>6.0</td>
</tr>
<tr>
<td>356</td>
<td>S0072</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>3 No</td>
<td>No</td>
<td>No</td>
<td>5.5</td>
</tr>
<tr>
<td>357</td>
<td>S0072</td>
<td>2</td>
<td>7</td>
<td>9</td>
<td>3 No</td>
<td>No</td>
<td>No</td>
<td>6.5</td>
</tr>
<tr>
<td>358</td>
<td>S0072</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>3 No</td>
<td>No</td>
<td>No</td>
<td>NA</td>
</tr>
<tr>
<td>359</td>
<td>S0072</td>
<td>4</td>
<td>27</td>
<td>9</td>
<td>3 No</td>
<td>No</td>
<td>No</td>
<td>6.0</td>
</tr>
<tr>
<td>360</td>
<td>S0072</td>
<td>5</td>
<td>15</td>
<td>9</td>
<td>3 No</td>
<td>No</td>
<td>No</td>
<td>5.5</td>
</tr>
<tr>
<td>361</td>
<td>S0073</td>
<td>1</td>
<td>11</td>
<td>9</td>
<td>4 No</td>
<td>No</td>
<td>No</td>
<td>5.5</td>
</tr>
<tr>
<td>362</td>
<td>S0073</td>
<td>2</td>
<td>21</td>
<td>9</td>
<td>4 No</td>
<td>No</td>
<td>No</td>
<td>6.5</td>
</tr>
<tr>
<td>363</td>
<td>S0073</td>
<td>3</td>
<td>17</td>
<td>9</td>
<td>4 No</td>
<td>No</td>
<td>No</td>
<td>5.5</td>
</tr>
<tr>
<td>364</td>
<td>S0073</td>
<td>4</td>
<td>11</td>
<td>9</td>
<td>4 No</td>
<td>No</td>
<td>No</td>
<td>6.0</td>
</tr>
<tr>
<td>365</td>
<td>S0073</td>
<td>5</td>
<td>3</td>
<td>9</td>
<td>4 No</td>
<td>No</td>
<td>No</td>
<td>6.0</td>
</tr>
</tbody>
</table>

- Correlating Extraversion & ReadingSpan → delete/ignore the blue row
Casewise Deletion

- Avoids data loss

- But, results not completely consistent or comparable—based on different observations
Week 10.2: Missing Data

- Rank Deficiency
- Linear Combinations
- Incomplete Designs
- Missing Data (NA values)
- Types of Missingness
  - Non-Ignorable
  - Ignorable
- Summary
- Possible Solutions
  - Case-wise Deletion
  - List-wise Deletion
    - Unconditional Imputation
    - Conditional Imputation
    - Multiple Imputation
    - Pattern-Mixture Models
**Listwise Deletion**

- Delete any observation where data is missing *anywhere*
- e.g., `stress %>% na.omit() -> stress2`
- Default in `lmer()` and many other analyses
Listwise Deletion

- Avoids inconsistency
- In some cases, could result in a lot of data loss
  - However, mixed effects models do well even with moderate data loss (25%; Quene & van den Bergh, 2004)
  - Unlike ANOVA, MEMs properly account for some subjects or conditions having fewer observations

<table>
<thead>
<tr>
<th>Subject</th>
<th>Week</th>
<th>TempC</th>
<th>Extraversion</th>
<th>Conscientiousness</th>
<th>Anxiety</th>
<th>Severity</th>
<th>Reading Span</th>
<th>Operation Span</th>
<th>Cortisol nmol</th>
</tr>
</thead>
<tbody>
<tr>
<td>350</td>
<td>5</td>
<td>4</td>
<td>12</td>
<td>NA</td>
<td>Yes</td>
<td>No</td>
<td>6.5</td>
<td>3.5</td>
<td>18.95714</td>
</tr>
<tr>
<td>351</td>
<td>6</td>
<td>7</td>
<td>No</td>
<td>NA</td>
<td>No</td>
<td>No</td>
<td>6.5</td>
<td>5.0</td>
<td>21.62344</td>
</tr>
<tr>
<td>352</td>
<td>2</td>
<td>4</td>
<td>7</td>
<td>NA</td>
<td>No</td>
<td>No</td>
<td>5.5</td>
<td>3.0</td>
<td>19.11777</td>
</tr>
<tr>
<td>353</td>
<td>3</td>
<td>19</td>
<td>7</td>
<td>NA</td>
<td>No</td>
<td>No</td>
<td>6.0</td>
<td>6.0</td>
<td>23.07907</td>
</tr>
<tr>
<td>354</td>
<td>4</td>
<td>13</td>
<td>7</td>
<td>NA</td>
<td>No</td>
<td>No</td>
<td>6.5</td>
<td>7.0</td>
<td>22.10193</td>
</tr>
<tr>
<td>355</td>
<td>5</td>
<td>2</td>
<td>7</td>
<td>NA</td>
<td>No</td>
<td>No</td>
<td>6.0</td>
<td>2.0</td>
<td>19.63036</td>
</tr>
<tr>
<td>356</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>5.5</td>
<td>5.5</td>
<td>14.40984</td>
</tr>
<tr>
<td>357</td>
<td>2</td>
<td>7</td>
<td>9</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>6.5</td>
<td>4.5</td>
<td>15.45432</td>
</tr>
<tr>
<td>358</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>No</td>
<td>No</td>
<td>NA</td>
<td>4.5</td>
<td>4.5</td>
<td>15.02257</td>
</tr>
<tr>
<td>359</td>
<td>4</td>
<td>27</td>
<td>9</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>6.0</td>
<td>5.5</td>
<td>21.86012</td>
</tr>
<tr>
<td>360</td>
<td>5</td>
<td>15</td>
<td>9</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>5.5</td>
<td>5.5</td>
<td>15.86730</td>
</tr>
<tr>
<td>361</td>
<td>1</td>
<td>11</td>
<td>9</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>5.5</td>
<td>3.5</td>
<td>21.54821</td>
</tr>
<tr>
<td>362</td>
<td>2</td>
<td>21</td>
<td>9</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>6.5</td>
<td>5.0</td>
<td>25.52504</td>
</tr>
<tr>
<td>363</td>
<td>3</td>
<td>17</td>
<td>9</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>5.5</td>
<td>4.5</td>
<td>21.84911</td>
</tr>
<tr>
<td>364</td>
<td>4</td>
<td>11</td>
<td>9</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>6.0</td>
<td>4.0</td>
<td>21.03723</td>
</tr>
<tr>
<td>365</td>
<td>5</td>
<td>3</td>
<td>9</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>6.0</td>
<td>8.5</td>
<td>18.97753</td>
</tr>
</tbody>
</table>
**Listwise Deletion**

- Avoids inconsistency
- In some cases, could result in a lot of data loss
  - However, mixed effects models do well even with moderate data loss (25%; Quene & van den Bergh, 2004)
  - Unlike ANOVA, MEMs properly account for some subjects or conditions having fewer observations
- Produces the correct parameter estimates if missingness is *ignorable*
  - Although some other things ($R^2$) may be incorrect
- Estimates will be wrong if missingness is *non-ignorable*
Week 10.2: Missing Data

- Rank Deficiency
  - Linear Combinations
  - Incomplete Designs
- Missing Data (NA values)
  - Types of Missingness
    - Non-Ignorable
    - Ignorable
  - Summary
- Possible Solutions
  - Casewise Deletion
  - Listwise Deletion
  - Unconditional Imputation
    - Conditional Imputation
    - Multiple Imputation
    - Pattern-Mixture Models
Unconditional Imputation

- Replace missing values with the mean of the observed values

5, 8, 3, ?, ?

- $M = 5.33$
- $S^2 = 12.5$

5, 8, 3, 5.33, 5.33

- $M = 5.33$
- $S^2 = 3.17$

- Imputing the mean reduces the variance
  - This increases chance of detecting spurious effects
  - Also distorts the correlations with other variables
- Bad. Don’t do this!
Week 10.2: Missing Data

- Rank Deficiency
- Linear Combinations
- Incomplete Designs
- Missing Data (NA values)
  - Types of Missingness
    - Non-Ignorable
    - Ignorable
    - Summary
- Possible Solutions
  - Casewise Deletion
  - Listwise Deletion
  - Unconditional Imputation
  - Conditional Imputation
    - Multiple Imputation
    - Pattern-Mixture Models
**Conditional Imputation**

- Replace missing values with the values predicted by a model using known variable(s).

If *ignorable* missingness, get the correct parameter estimates.

And, standard errors not as distorted as unconditional imputation.
  - Especially if we add some noise to the fitted values.

```
Reading Span ~ 1 + Operation Span
```

Reading Span = NA
**Conditional Imputation**

- Replace missing values with the values predicted by a model using known variable(s)

Where this is useful?
- Many observations have a small amount of missing data, but which column it is varies
- Listwise deletion would wipe out every row with a NA anywhere
Week 10.2: Missing Data

- Rank Deficiency
  - Linear Combinations
  - Incomplete Designs
- Missing Data (NA values)
  - Types of Missingness
    - Non-Ignorable
    - Ignorable
    - Summary
- Possible Solutions
  - Casewise Deletion
  - Listwise Deletion
  - Unconditional Imputation
  - Conditional Imputation
  - Multiple Imputation
- Pattern-Mixture Models
Multiple Imputation

- Like doing conditional imputation several times
  - Replace missing data with one possible set of values
  - Run the model
  - Repeat it
- Final result *averages* these
  - Reflects the *uncertainty* about the missing data
Multiple Imputation with `mice`

- **Process:**
  - `imp <- mice(stress)`
    - Creates several sets of imputed data
    - Need to indicate level 1 vs level 2 variables
    - Check to make sure imputed data looks reasonable (e.g., in same range as original data)
  - `miModel <- with(imp, lmer(…) )`
    - Fit the model to each set of imputed data
  - `result <- pool(miModel)`
    - Combine the model results
  - `summary(result)`

- **Limitations:**
  - Limited to two nested levels (with current software)
  - Only gives you fixed effect estimates (not estimates of random effect variance)
  - Can be time-consuming
Week 10.2: Missing Data

- Rank Deficiency
  - Linear Combinations
  - Incomplete Designs
- Missing Data (NA values)
  - Types of Missingness
    - Non-Ignorable
    - Ignorable
  - Summary
- Possible Solutions
  - Casewise Deletion
  - Listwise Deletion
  - Unconditional Imputation
  - Conditional Imputation
  - Multiple Imputation
- Pattern-Mixture Models
Pattern Mixture Models

- Classify participants by the patterns of missing data
- Then, look at the effects / pattern of results within each group

Example:
- Analysis 1: Relation of personality & anxiety to stress for people with personality data reported
- Analysis 2: Relation of anxiety to stress for people without personality data reported
Encyclopedia Brown confronted local troublemaker “Bugs” Hauser about the missing data. Bugs says he distinctly remembers storing the missing sheet of data between pages 151 and 152 of his lab notebook. Bugs says that the sheet must have just fallen out when Bugs's gang, the Cotton-Top Tamarins, were cleaning their clubhouse.

How did Encyclopedia know Bugs was lying?

Pages 151 and 152 are the front and back of the same sheet.