End-to-end Analysis:
a case study in using R

EPIC R 2015
Original Problem Statement

• Hypothesis: prenatal exposure to phthalates causes obesity in children
About Phthalates

• Phthalates are a broad class of chemicals, some naturally present in foods, some used as plasticizers
  – may be endocrine disruptors
• Hard to measure external dose – usually measure urinary metabolites
Initial Analysis

- Characterized 9 phthalate metabolites in urine samples from pregnant women
  - Compared with body size of kids at age 7
An aside on molecular epidemiology and metabolites...

• Suppose Z is a chemical which is metabolized as: Z -> Z1 -> Z2 -> Z3
  – Z3 is excreted in urine
  – Z2 causes cancer
  – Metabolic efficiency varies between people

• If person A has more urinary Z3 than person B does, who is at higher risk for cancer owing to Z exposure?
... but anyway

• It turns out that the levels of these 9 metabolites are log-normally distributed and are correlated

• So, analytic plan: use PCA to identify components of variance, and use components to predict obesity status
Reviewer feedback

- How do we know if your results generalize outside your sample if you use component scores?
So… our goal:

• See how similar the components identified by a PCA for NHANES women age 15-45 are to the components identified by the PCA in the CCCEH cohort

• Rundelian Jiu-jitsu: If a PCA identifies similar components in different people, then the components are more likely a reflection of between-individual variation in common exposures than an artifact of the sample
Specific plan

1. Download relevant NHANES data
2. Load NHANES data into R
3. Filter to women aged 15-45
4. Log transform phthalate metabolite variables
5. Run a PCA and compute component scores
6. Compare to component scores in CCCEH

• A Bonus question: what do we do about the complex survey sampling used by NHANES?
Get the data

• NHANES is publically available on the CDC's site
• We want demographics and phthalates for 2005-2006
Get the data into R

• The data come as a sas7bdat file
• Q: How do we get that into R?

• A: Need to use the read.sas7bdat method in the sas7bdat package
Merging data

• Google "R merge data frame"

• The merge function looks good

• Q1: How do we specify the id

• Q2: We don't need to here, but how would you do a 'vertical merge'
Filtering to women 15-45

• Q: Remember how to do this?

• A: Filter data frames by indexing and assigning the result
Find the variables

• Get the documentation file on the NHANES site

Can you send me the full PCA output and the correlation matrix? Make sure the metabolites are entered into the matrix in this order....

MEHP
MEHHP
MECPP
MEOHP
MCPP
MIBP
MBP
MBP
MBZP
MEP
Log transform variables

• Using indexing by name, we can log-transform them all at once

• Note: Check out the paste function
Loops

• There are several ways to loop in R, but the most common is using for:

```r
for (i in somevector) {
    #do something with i
}
```

• Sometimes it's easier to use the positional index, in which case the code takes this form:

```r
for (i in 1:length(somevector)) {
    #do something with somevector[i]
}
```
Loops vs. alternatives

• Computer scientists love vectorized operations, higher-order functions (e.g. apply), etc.
  – They are elegant (I think) and allow for writing compact code

• But many computer programmers find looping to be the natural way to think about repeated operations

• It's usually a wash in terms of performance in R (and elsewhere)
Doing the PCA

• Okay, so we've finally gotten the data we need – log transformed urinary phthalate levels for women aged 15-45 in 2005-2006 NHANES.

• So how do we do the PCA?
Which PCA function?

• Let's pick from among the PCA packages:
  • http://gastonsanchez.com/blog/how-to/2012/06/17/PCA-in-R.html

• (Which I found by searching for 'PCA in R')
So, some options

- stats::prcomp
- stats::princomp
- FactoMineR::PCA
- ade4::dudi.pca
- amap::acp
- Also… psych::principal, probably others
Aside: PCA vs. EFA

• PCA: identify components that are composites of observations

• EFA: identify factors that underlie variables – variables are composites of identified factors

• In practice, very similar, but (in my experience) PCA more appropriate for dimensionality reduction, EFA more appropriate for exploration and characterizing data
Q-mode PCA vs. R-mode PCA

• Q-mode analysis is looking for patterns of similarity in the subjects over variables
• R-mode analysis is looking for similarity in the variables over subjects.

• prcomp does Q-mode, princomp does R-mode
• Q: Which one do we want?
Weighted PCA

• In the past, I've used the 'survey' package to deal with complex survey weights in R

• Survey includes svyprcomp but not svyprincomp
  – Argh!
Next, FactoMineR

• FactoMineR's PCA comes recommended
• And googling suggests FactoMineR's PCA allows for row weights…
• Can use princomp with a covariance matrix
  – So if we supply a weighted covariance matrix, that suggests we should be able to get the right answer?

• …or will we?
Complex survey weighting

• Complex surveys often use clustered, stratified sampling

• Not going to get into the details, but essentially, you need to specify the clusters and strata correctly to get the right standard errors

• So, does just specifying the row weights work for us?
PCAs and complex surveys

- PCAs in complex survey data is an area of active research:
  - PCAs should account for the survey design, but they generally don't in practice
  - Technically, this is non-trivial
  - Biases do not appear to be huge for our purposes here
  - Read more: https://www.amstat.org/sections/SRMS/Proceedings/y2008/Files/302340.pdf
So, how does NHANES compare to CCCEH?
### PCA: CCCEH vs. NHANES

<table>
<thead>
<tr>
<th></th>
<th>Maternal DEHP Component(^a)</th>
<th>Maternal Non-DEHP Component(^a)</th>
<th>DEHP Component(^a)</th>
<th>Non-DEHP Component(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEHP</td>
<td>0.88</td>
<td>0.23</td>
<td>0.87</td>
<td>0.18</td>
</tr>
<tr>
<td>MEHHP</td>
<td>0.94</td>
<td>0.31</td>
<td>0.94</td>
<td>0.29</td>
</tr>
<tr>
<td>MECPP</td>
<td>0.92</td>
<td>0.29</td>
<td>0.93</td>
<td>0.27</td>
</tr>
<tr>
<td>MEOHP</td>
<td>0.92</td>
<td>0.35</td>
<td>0.93</td>
<td>0.31</td>
</tr>
<tr>
<td>MCPP</td>
<td>0.39</td>
<td>0.68</td>
<td>0.40</td>
<td>0.75</td>
</tr>
<tr>
<td>MIBP</td>
<td>0.26</td>
<td>0.79</td>
<td>0.31</td>
<td>0.80</td>
</tr>
<tr>
<td>MBP</td>
<td>0.19</td>
<td>0.89</td>
<td>0.33</td>
<td>0.84</td>
</tr>
<tr>
<td>MBZP</td>
<td>0.21</td>
<td>0.77</td>
<td>0.28</td>
<td>0.79</td>
</tr>
<tr>
<td>MEP</td>
<td>0.23</td>
<td>0.60</td>
<td>0.01</td>
<td>0.63</td>
</tr>
</tbody>
</table>

CCCEH

NHANES
Conclusion

• Pretty much the same
  – Paper was conditionally accepted

• Confession: I sanity-checked against SAS, because I felt like I was in uncharted waters with the complex survey stuff
  – SAS does not allow specification of survey design, just weights (like FactoMineR)