Data Cleaning Workshop:

How to Prepare your Data Prior to Analysis

Abby L. Braitman

Old Dominion University

November 1, 2019

These slides are available on my website

- https://fs.wp.odu.edu/abraitma/workshops/

Outline for Today

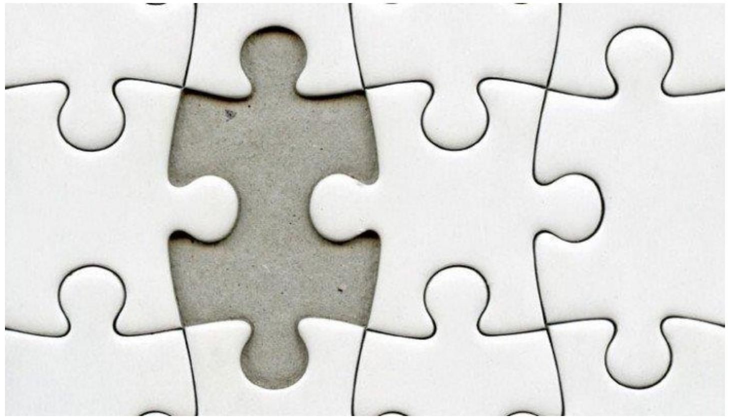
- Missing Data
  - Identifying, assessing type, imputation options
- Composite Scores
  - Total scores, recoding, dummy coding
- Outliers
  - Identifying and addressing univariate and multivariate outliers
- Normality
  - Assessing and addressing (e.g., transformations, analysis specifications)
- Bivariate Linearity
  - Reading scatterplots and what to do about them
- Documentation
  - The importance of codebooks and data logs

Outline for Today

- **Missing Data**
  - **Identifying, assessing type, imputation options**
- Composite Scores
  - Total scores, recoding, dummy coding
- Outliers
  - Identifying and addressing univariate and multivariate outliers
- Normality
  - Assessing and addressing (e.g., transformations, analysis specifications)
- Bivariate Linearity
  - Reading scatterplots and what to do about them
- Documentation
  - The importance of codebooks and data logs

Missing Data

- Types
  - MCAR, MAR, MNAR
- Types of Imputation
  - Multiple, EM, regression, mean, etc.
- How to impute



Why talk about missing data?

- Dissertation: college drinkers (SONA plus other classes)
    - Time 1 *n*= 353
    - 2 week follow-up *n* = 213 (60.3%)
    - 4 week follow-up *n* = 115 (32.6%)
- Fellowship: college drinkers (SONA)
    - Time 1 *n* = 537
    - 2 weeks (*n* = 338; 62.9%), 4 weeks (*n* = 284; 52.9%), and 6 weeks (*n* = 259; 48.2%) post-intervention
    - 3 months (*n* = 213; 39.7%), 6 months (*n* = 173; 32.2%), and 9 months (*n* = 140; 26.1%) post-intervention
- K01 Study 1a: college drinkers (SONA [75.8%] plus general student body)
    - Time 1 *n* = 546
    - 1 month (n = 364; 66.7%), 3 months (n = 312; 57.1%)
- K01 Study 1b: college drinkers (general student body only) – numbers are tentative
    - Time 1 n = 249
    - 1 month (n = 217; 87.1%), 3 month (n = 90 out of 110 possible; 81.8%)

Why talk about missing data?

Measure (and Items)	# of Missing Values	Result of Missingness
Baseline		
PBS Items (21 items)	51	Imputed
Alcohol Knowledge (15 items)*	18	Imputed
Number of Drinking Days (1 item)	0	
Days Intoxicated (1 item)	1	Imputed
Drinks on Highest Drinking Day (1 item)	0	
BAC on Highest Drinking Day (1 item)	0	
GPA (1 item)	198	Not imputed (missing > 50%)
Age (1 item)	0	
Follow-ups (both 2-week and 4-week)		
PBS Items (21 items)	154	Imputed (except 6 cases >20%)
Alcohol Knowledge (15 items)*	73	Imputed (except 1 case >20%)
Number of Drinking Days (1 item)	0	
Days Intoxicated (1 item)	2	Imputed
Drinks on Highest Drinking Day (1 item)	0	
BAC on Highest Drinking Day (1 item)	0	

Missing Data

- More Info:
  - Rubin, Donald B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.
    - Free online
  - Allison, Paul D. (2001). *Missing Data*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-136. Thousand Oaks, CA: Sage.
    - Cheap on Amazon
  - Enders, Craig K. (2010). *Applied Missing Data Analysis*. New York: Guilford Press.
    - Available in ODU library as a book and as an *e*-book
  - http://www.appliedmissingdata.com/ (Enders' website)
    - Slides and articles
    - Step-by-step guide for multiple methods in multiple software packages
  - Enders, C.K., Baraldi, A.N., & Cham, H. (2014). Estimating interaction effects with incomplete predictor variables. *Psychological Methods*, *19*, 39-55.
  - Enders, C.K. (2011). Analyzing longitudinal data with missing values. *Rehabilitation Psychology*, *56*, 267-288.
  - Enders, C.K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, *16*, 1-16.
  - Muthén, B., Asparouhov, T., Hunter, A. & Leuchter, A. (2011). Growth modeling with non-ignorable dropout: Alternative analyses of the STAR*D antidepressant trial. *Psychological Methods*, *16*, 17-33.

Missing Data Types

- Conceived by Donald B. Rubin (1976)
- Missing Completely At Random (**MCAR**)
  - “Probability of missing data on Y is unrelated to the value of Y itself or to the values of other variables in the data set”
  - Missing values are completely independent from observed or unobserved data
  - Truly random skips (nothing to do with answers)
    - E.g., someone skipped salary not because their salary was particularly high or low, and not because of their race, ethnicity, gender, depressive symptoms, anxiety, etc.
  - VIOLATED if: someone skipped this question because their salary was very low, or if everyone who skipped this question was very young
  - RARE (we think)
    - People often skip for a reason

Missing Data Types

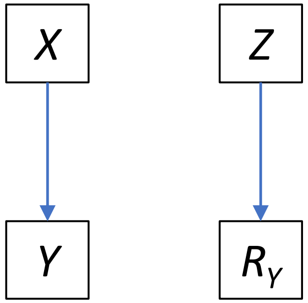
- Missing At Random (**MAR**)
  - “Probability of missing data on Y is unrelated to the value of Y, after controlling for other variables in the analysis”
  - Missing values are completely dependent on observed values
  - You can estimate what their answer WOULD BE based on other information in the data
    - E.g., People who skipped a question about binge drinking are very high on drinking quantity and drinking frequency
    - E.g., someone skips a single item on a multi-item scale
  - VIOLATED if: people skipped a question because they are high on binge drinking, but this is not strongly related to other variables in the sample (e.g., no other drinking variables, OR not a strong correlation)
  - MUCH MORE COMMON
    - The reason people skip an item is captured somewhere else in your data

Missing Data Types

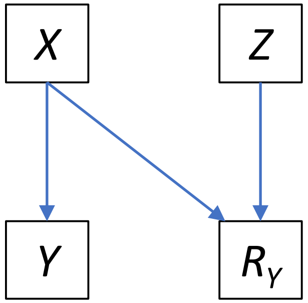
- Analyses usually assume MCAR or MAR
  - “Ignorable” missing data
- Missing Not At Random (**MNAR**)
  - Missing values are dependent on unobserved values
  - E.g., Skipped trauma questionnaires because high on trauma (not indicated with other variables)
  - “Nonignorable” missing data
    - VERY BAD
  - Cannot perform most analyses (results would be biased)
    - Excluding everyone high in drinking, or high in trauma, or low in income, etc.
    - Everyone who is getting worse in a clinical trial drops out
  - The missing data mechanism must be modeled in your analysis
    - E.g., Heckman’s (1976) two-stage estimator for regression models with selection bias on the dependent variable
    - Requires a lot of knowledge about why data are missing, sophisticated calculations, lack of software, difficulties in interpretability
    - UPDATE: Software is more accommodating, helpful articles

Missing Data Types

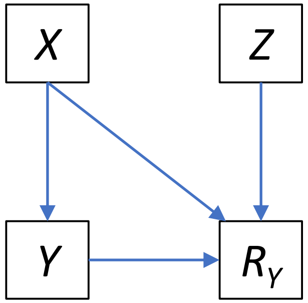
- If Z = set of UNOBSERVED variables uncorrelated with X and Y
- R_Y = missing data indicator for Y



MCAR



MAR



MNAR

Missing Data Types

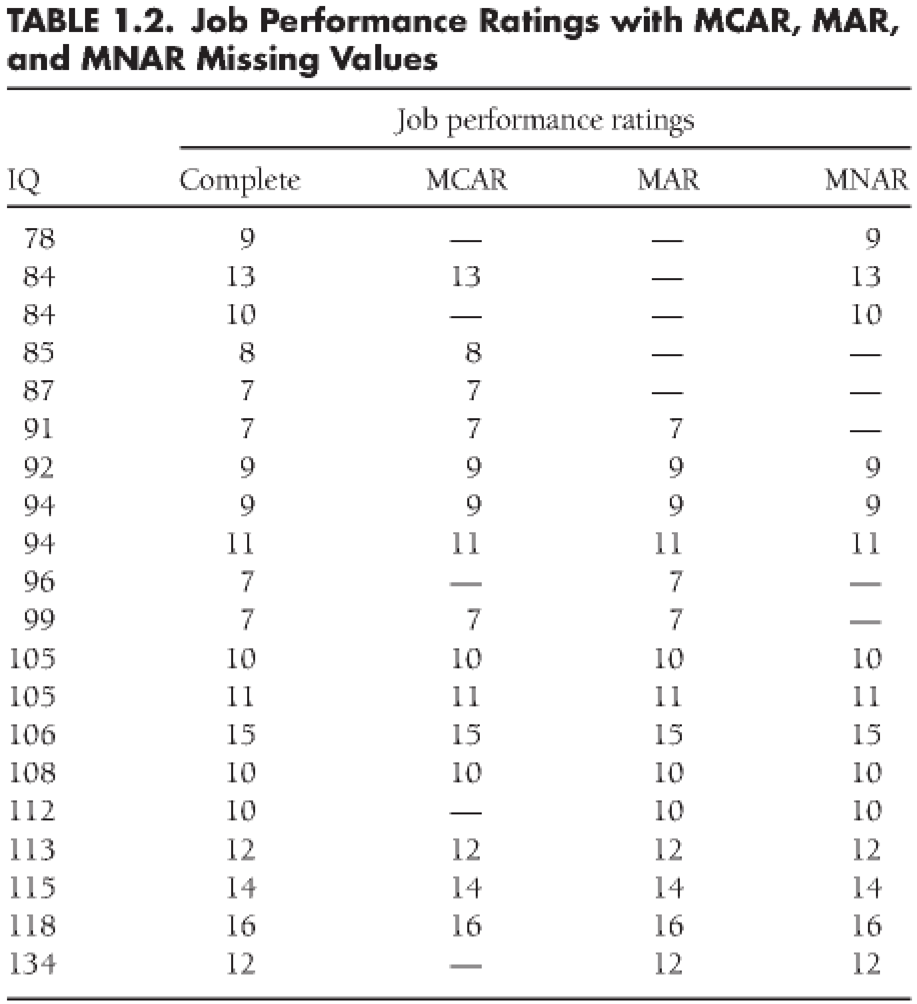


TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

	Job performance ratings			
IQ	Complete	MCAR	MAR	MNAR
78	9	—	—	9
84	13	13	—	13
84	10	—	—	10
85	8	8	—	—
87	7	7	—	—
91	7	7	7	—
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	—	7	—
99	7	7	7	—
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	—	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	—	12	12

Enders (2010, p. 7)

- *Y* = Job performance rating
- *X* = IQ

Missing Data Types

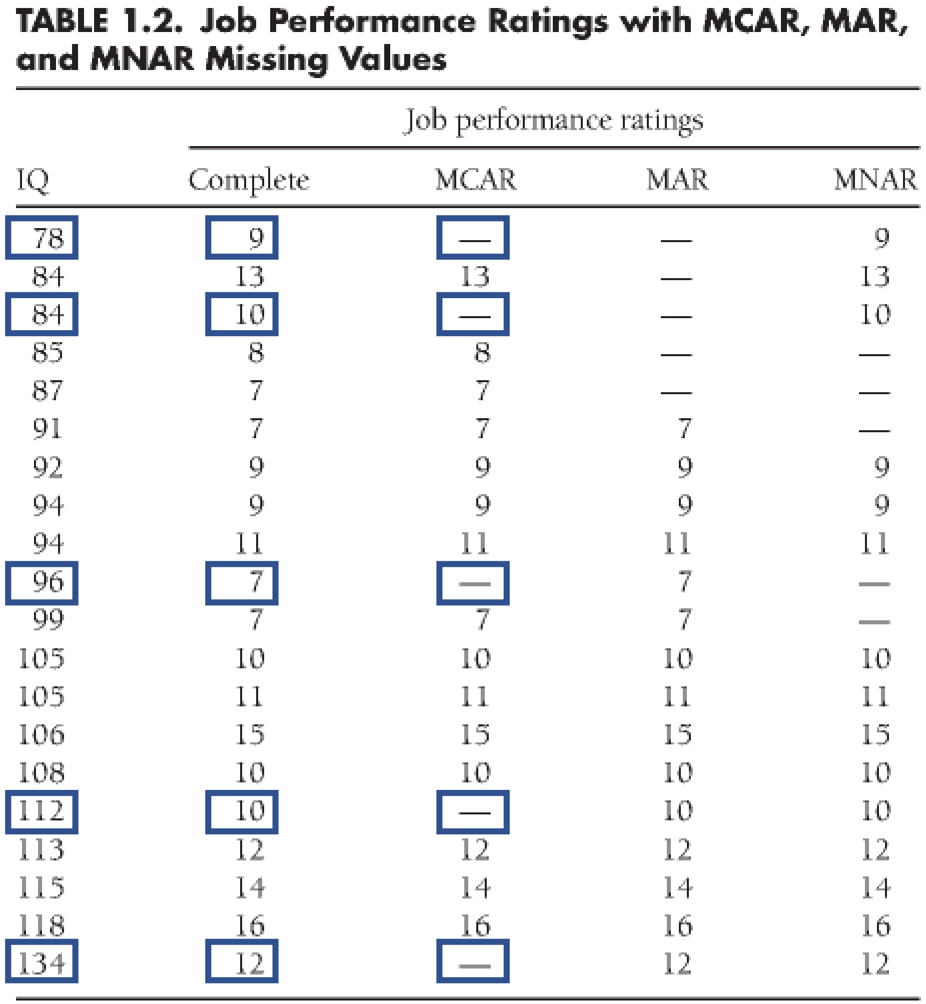


TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

	Job performance ratings			
IQ	Complete	MCAR	MAR	MNAR
78	9	—	—	9
84	13	13	—	13
84	10	—	—	10
85	8	8	—	—
87	7	7	—	—
91	7	7	7	—
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	—	7	—
99	7	7	7	—
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	—	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	—	12	12

Enders (2010, p. 7)

- *Y* = Job performance rating
- *X* = IQ
- Values missing for **MCAR** are truly random

Missing Data Types

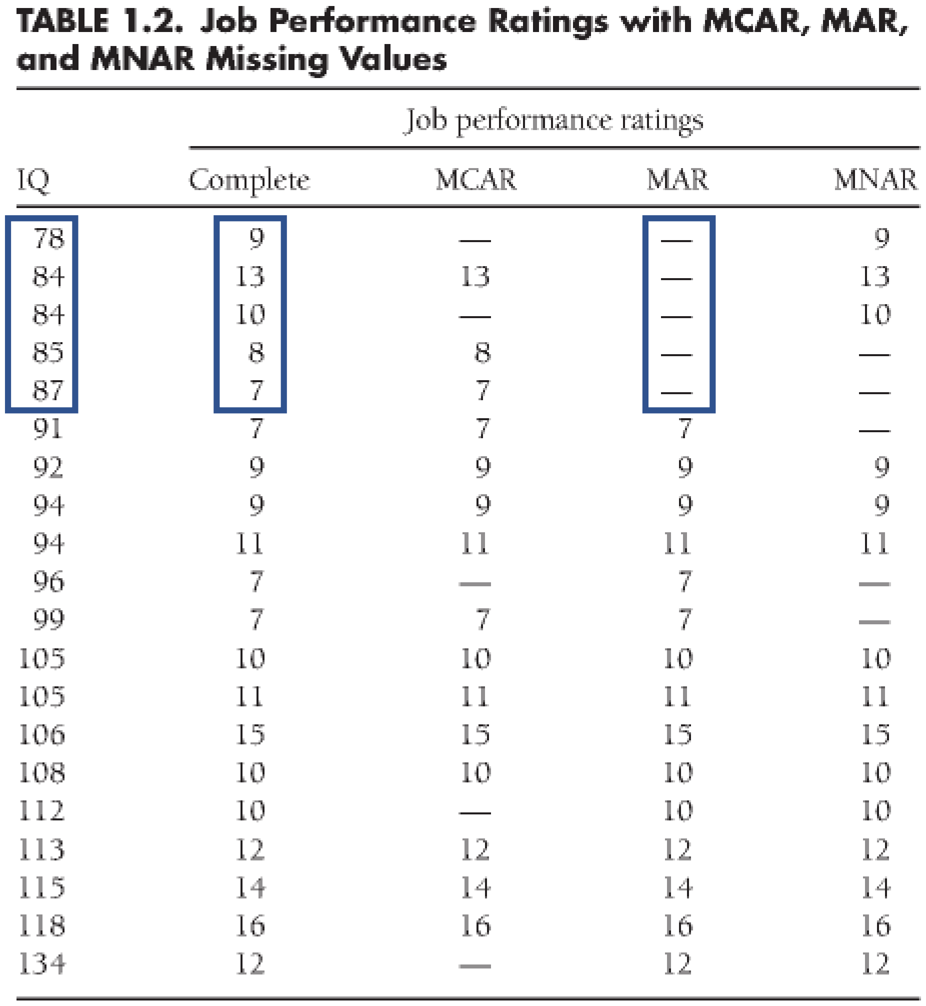


TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

	Job performance ratings			
IQ	Complete	MCAR	MAR	MNAR
78	9	—	—	9
84	13	13	—	13
84	10	—	—	10
85	8	8	—	—
87	7	7	—	—
91	7	7	7	—
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	—	7	—
99	7	7	7	—
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	—	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	—	12	12

Enders (2010, p. 7)

- *Y* = Job performance rating
- *X* = IQ
- Values missing for **MCAR** are truly random
- Values missing for **MAR** are related to *X*
  - Lowest IQ
  - X is correlated with Y (generally lower job performance ratings, but not the lowest)
  - Together with other information in the data, can potentially predict the missing values

Missing Data Types

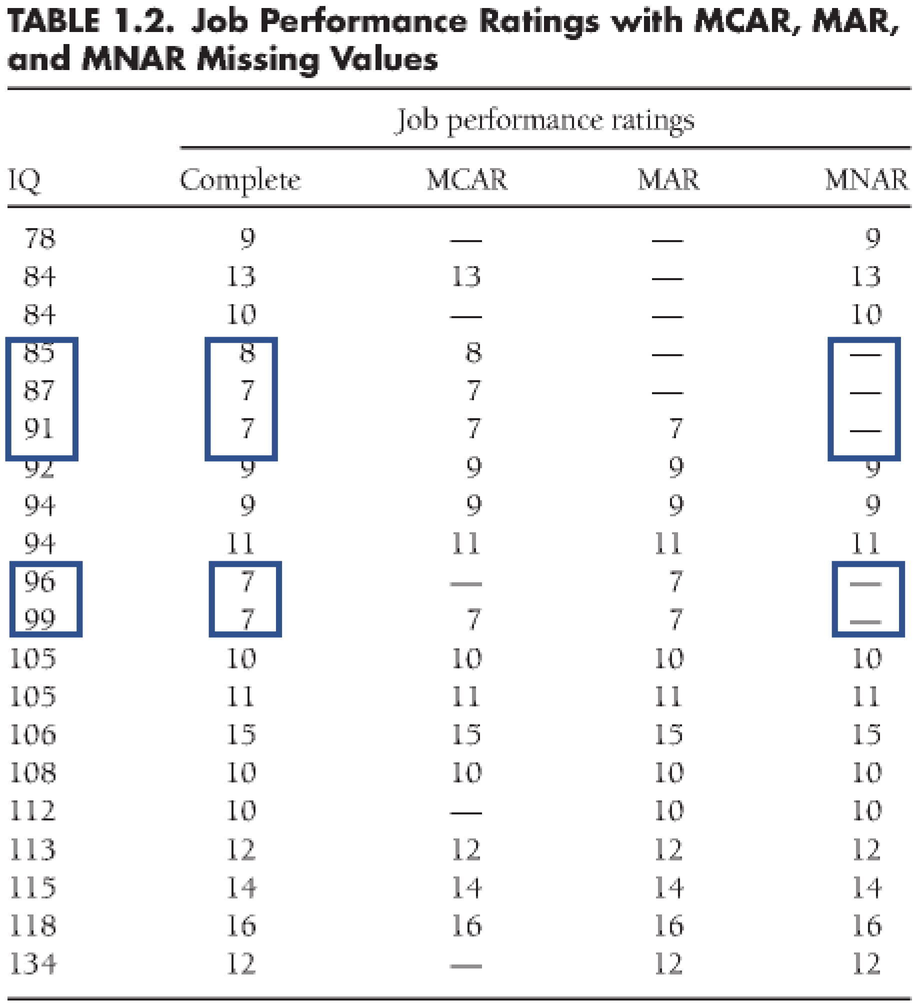


TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

IQ	Job performance ratings			
	Complete	MCAR	MAR	MNAR
78	9	—	—	9
84	13	13	—	13
84	10	—	—	10
85	8	8	—	—
87	7	7	—	—
91	7	7	7	—
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	—	7	—
99	7	7	7	—
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	—	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	—	12	12

Enders (2010, p. 7)

- *Y* = Job performance rating
- *X* = IQ
- Values missing for **MCAR** are truly random
- Values missing for **MAR** are related to *X*
  - Lowest IQ
- Values missing for **MNAR** are related to *Y*, even after controlling for *X*
  - Lowest job performance, even though IQ may be higher

Missing Data Types

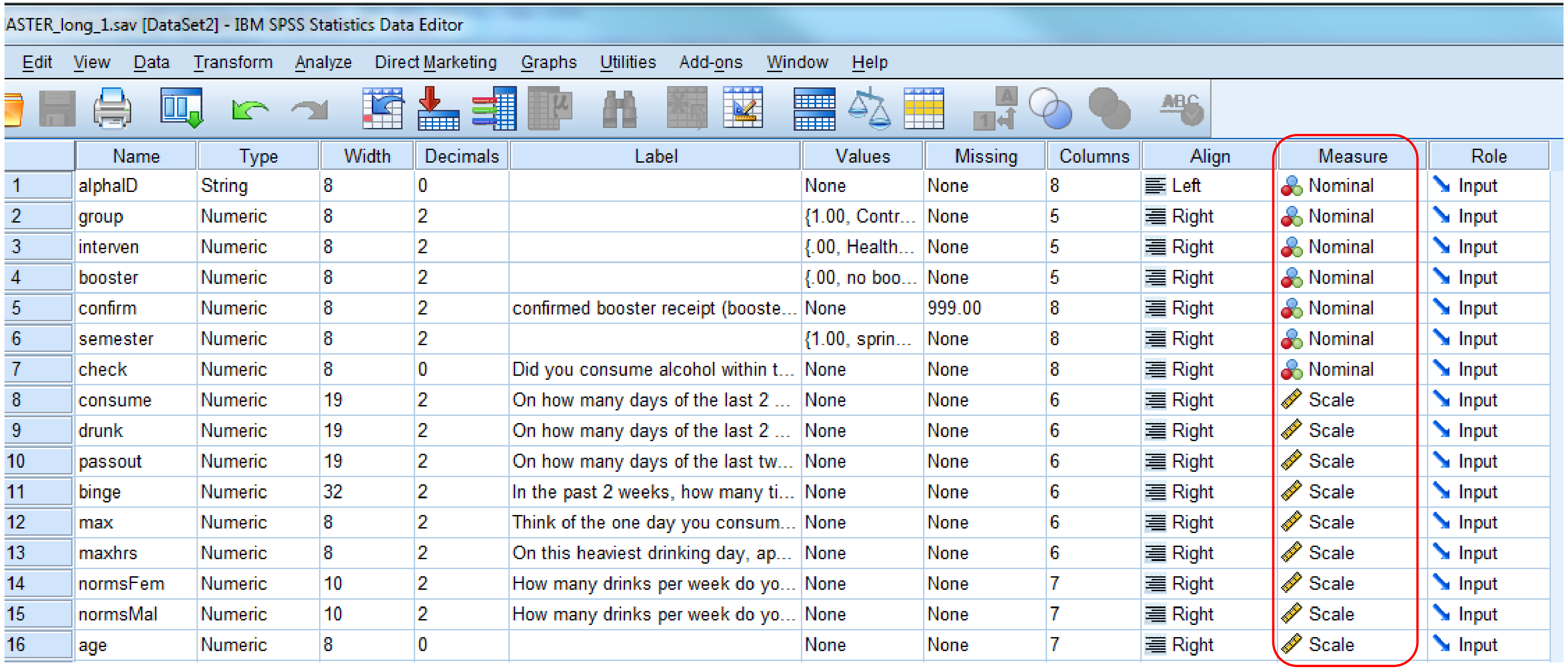
- How do I know what type of data I have?
  - Difficult to test. The information needed is missing!
  - Can test if missingness is related to variables in your sample
    - Little’s MCAR test in SPSS, create your own *t* tests, etc.
    - If related, supports MAR (that missing values depend on observed values)
    - If unrelated, cannot determine MCAR versus MNAR
    - Best techniques for addressing missing data do not distinguish between MCAR and MAR, so test is relatively pointless
      - Cannot tell you IF you can proceed
      - CAN tell you what variables need to be included in your model
  - I recommend identifying missingness for outcome variables, and identifying potential predictors associated/correlated with them
    - Can describe your sample for readers/reviewers
    - Can include these variables in your imputation process
    - Can include these variables in your model

Identifying Missing Data Patterns

- From a workshop with Enders (no citation):
  - It's less about the significance test itself. Don't focus on *p* values.
  - It's about the effect size (cohen's *d*, or *r* or r^2). You don't want this to depend on sample size
  - Therefore, Little’s MCAR test (given in SPSS) is less helpful
    - Presents overall significance for all data, but does not show where differences/associations are or how large they are
- Little’s MCAR test is not the only thing provided in the output
- The patterns can be useful

Identifying Missing Data Patterns: Option 1

- Using “Missing Value Analysis” in SPSS
- Make sure “Measure” is correct in Variable View

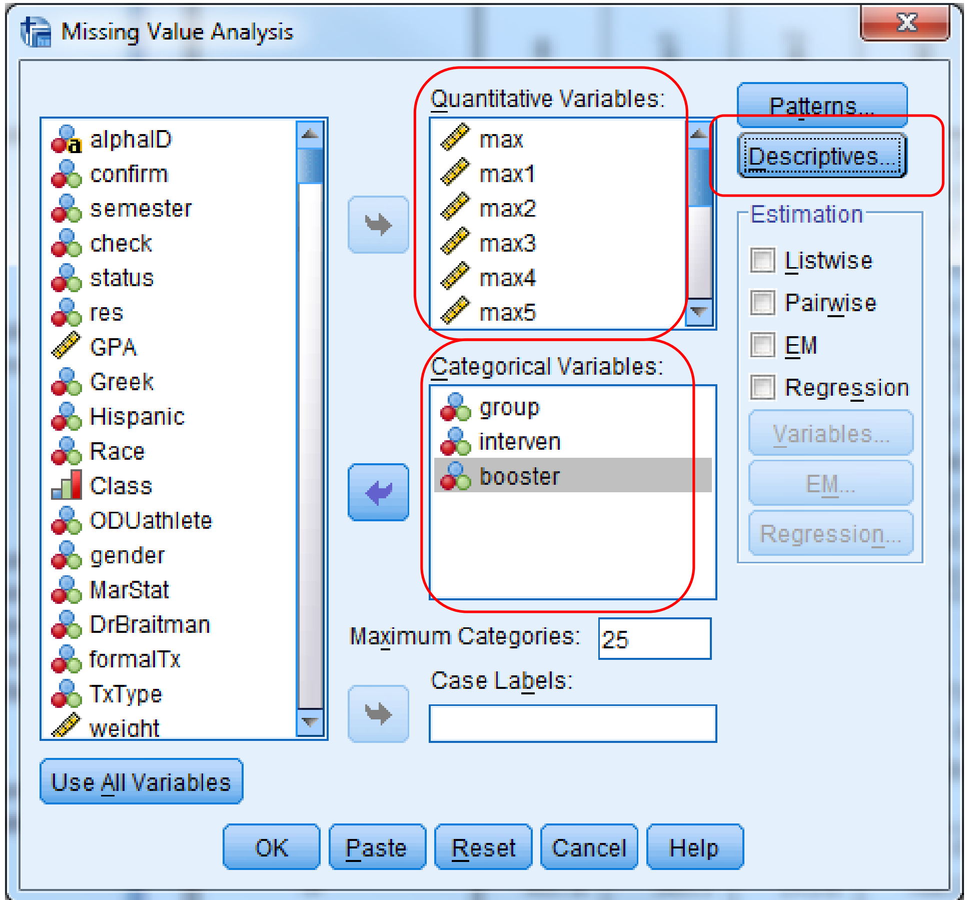


ASTER_long_1.sav [DataSet2] - IBM SPSS Statistics Data Editor

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	alphaID	String	8	0		None	None	8	Left	Nominal	Input
2	group	Numeric	8	2		{1.00, Contr...	None	5	Right	Nominal	Input
3	interven	Numeric	8	2		{.00, Health...	None	5	Right	Nominal	Input
4	booster	Numeric	8	2		{.00, no boo...	None	5	Right	Nominal	Input
5	confirm	Numeric	8	2	confirmed booster receipt (booste...	None	999.00	8	Right	Nominal	Input
6	semester	Numeric	8	2		{1.00, sprin...	None	8	Right	Nominal	Input
7	check	Numeric	8	0	Did you consume alcohol within t...	None	None	8	Right	Nominal	Input
8	consume	Numeric	19	2	On how many days of the last 2 ...	None	None	6	Right	Scale	Input
9	drunk	Numeric	19	2	On how many days of the last 2 ...	None	None	6	Right	Scale	Input
10	passout	Numeric	19	2	On how many days of the last tw...	None	None	6	Right	Scale	Input
11	binge	Numeric	32	2	In the past 2 weeks, how many ti...	None	None	6	Right	Scale	Input
12	max	Numeric	8	2	Think of the one day you consum...	None	None	6	Right	Scale	Input
13	maxhrs	Numeric	8	2	On this heaviest drinking day, ap...	None	None	6	Right	Scale	Input
14	normsFem	Numeric	10	2	How many drinks per week do yo...	None	None	7	Right	Scale	Input
15	normsMal	Numeric	10	2	How many drinks per week do yo...	None	None	7	Right	Scale	Input
16	age	Numeric	8	0		None	None	7	Right	Scale	Input

Identifying Missing Data Patterns

- Analyze > Missing Value Analysis
- Move “scale” variables into “quantitative” box, and “nominal” variables into “categorical” box.



Missing Value Analysis

alphaID
confirm
semester
check
status
res
GPA
Greek
Hispanic
Race
Class
ODUathlete
gender
MarStat
DrBraitman
formalTx
TxType
weight

Quantitative Variables:
max
max1
max2
max3
max4
max5

Categorical Variables:
group
interven
booster

Patterns...

Descriptives...

Estimation
□ Listwise
□ Pairwise
□ EM
□ Regression

Variables...
EM...
Regression...

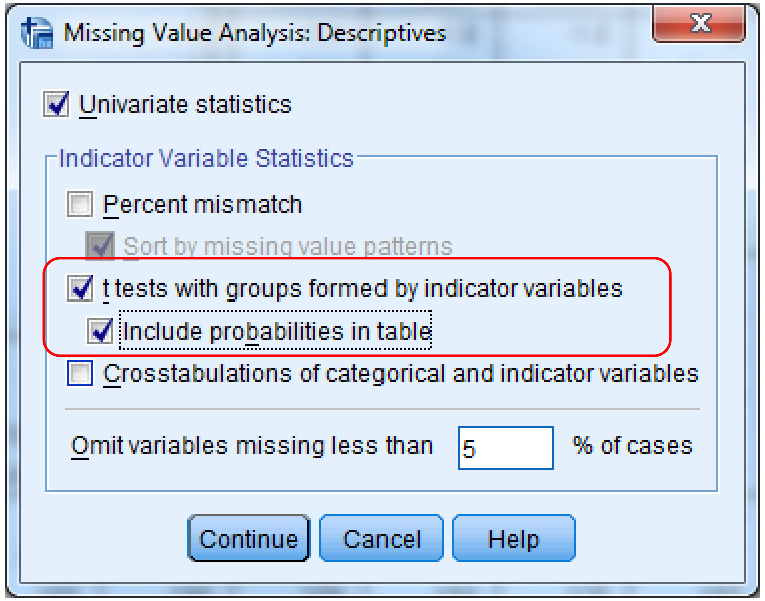
Maximum Categories: 25

Case Labels:

Use All Variables

OK Paste Reset Cancel Help

- Select “Descriptives” button
- Request “t tests with groups formed by indicator variables”
- Include probabilities



Missing Value Analysis: Descriptives

☑ Univariate statistics

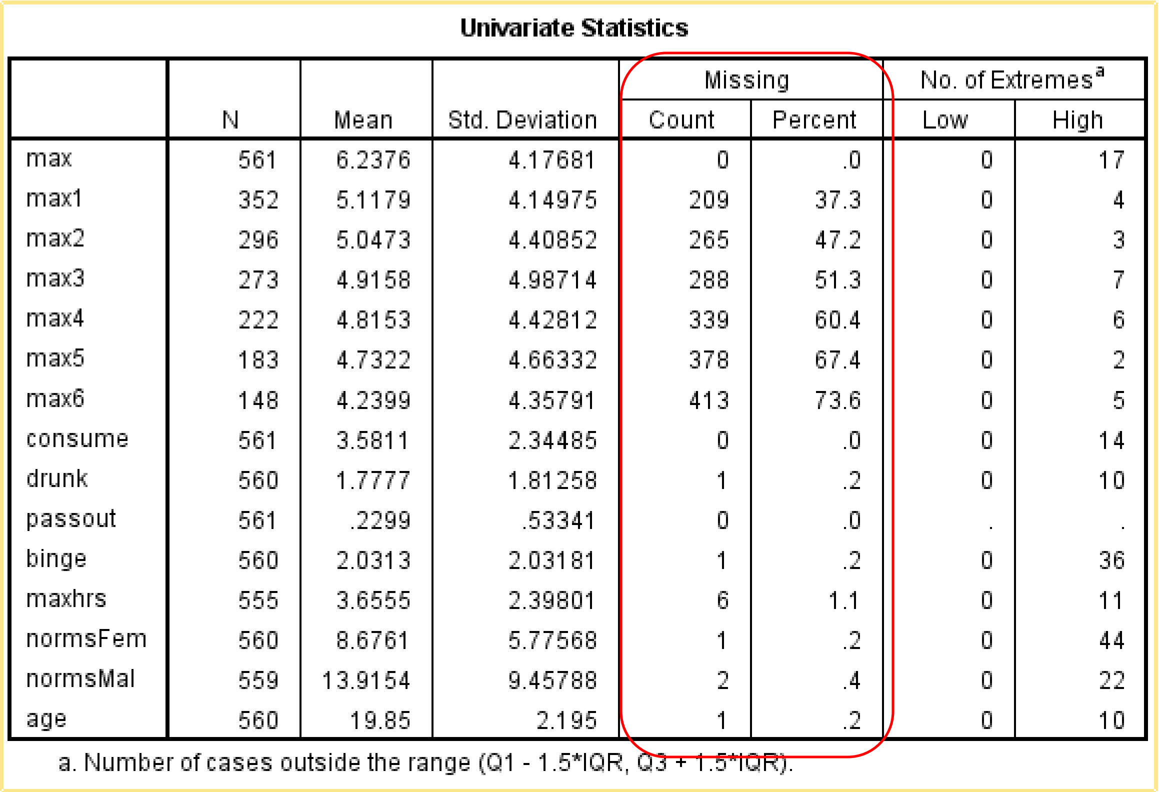
Indicator Variable Statistics
□ Percent mismatch
☑ Sort by missing value patterns
☑ t tests with groups formed by indicator variables
☑ Include probabilities in table
□ Crosstabulations of categorical and indicator variables

Omit variables missing less than 5 % of cases

Continue Cancel Help

Identifying Missing Data Patterns

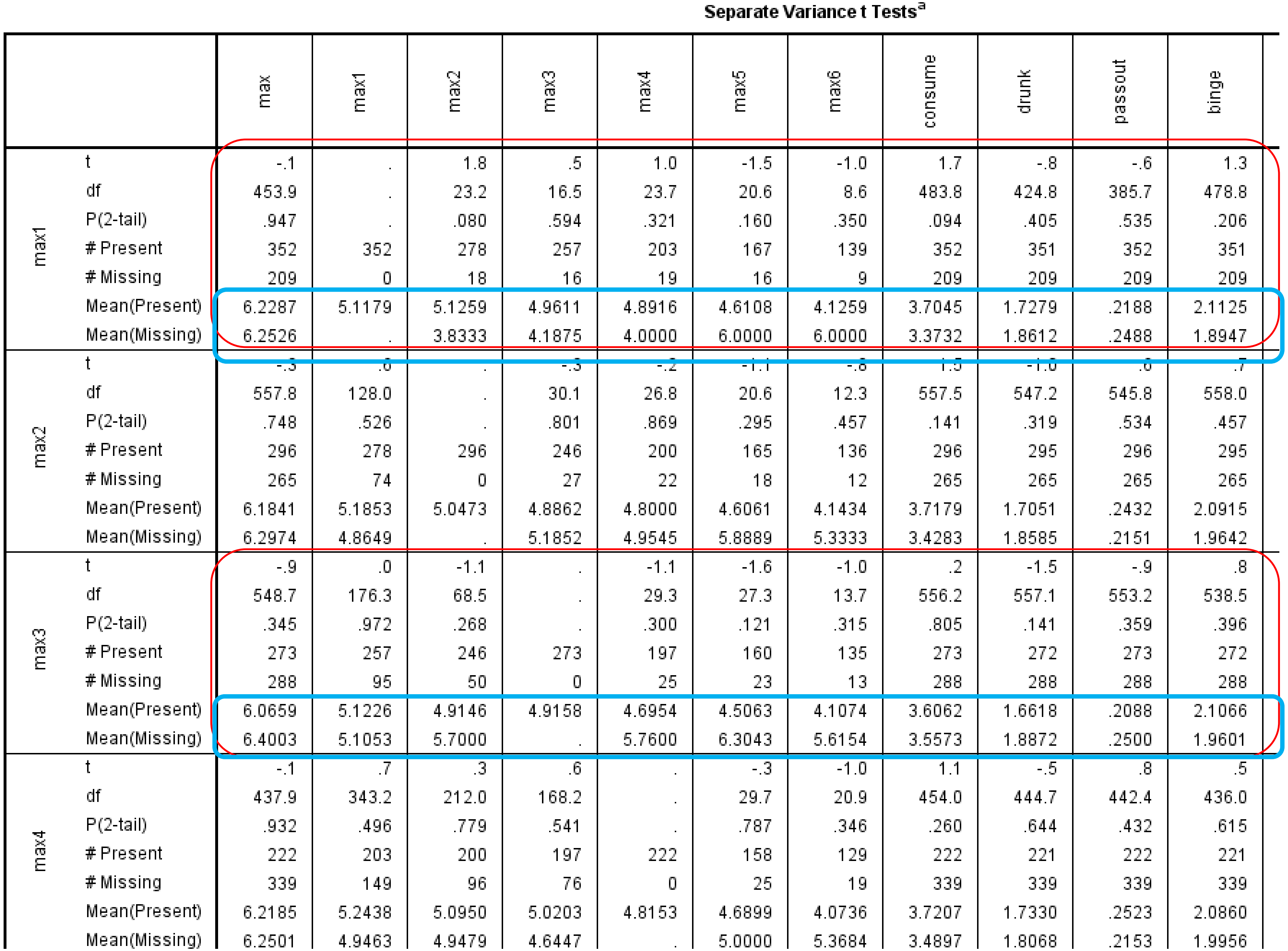
- Output will tell you how many cases are missing (*f* and %)
- Will tell you if missingness is associated with other variables



Univariate Statistics

	N	Mean	Std. Deviation	Missing		No. of Extremes[a]	
				Count	Percent	Low	High
max	561	6.2376	4.17681	0	.0	0	17
max1	352	5.1179	4.14975	209	37.3	0	4
max2	296	5.0473	4.40852	265	47.2	0	3
max3	273	4.9158	4.98714	288	51.3	0	7
max4	222	4.8153	4.42812	339	60.4	0	6
max5	183	4.7322	4.66332	378	67.4	0	2
max6	148	4.2399	4.35791	413	73.6	0	5
consume	561	3.5811	2.34485	0	.0	0	14
drunk	560	1.7777	1.81258	1	.2	0	10
passout	561	.2299	.53341	0	.0	.	.
binge	560	2.0313	2.03181	1	.2	0	36
maxhrs	555	3.6555	2.39801	6	1.1	0	11
normsFem	560	8.6761	5.77568	1	.2	0	44
normsMal	559	13.9154	9.45788	2	.4	0	22
age	560	19.85	2.195	1	.2	0	10

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

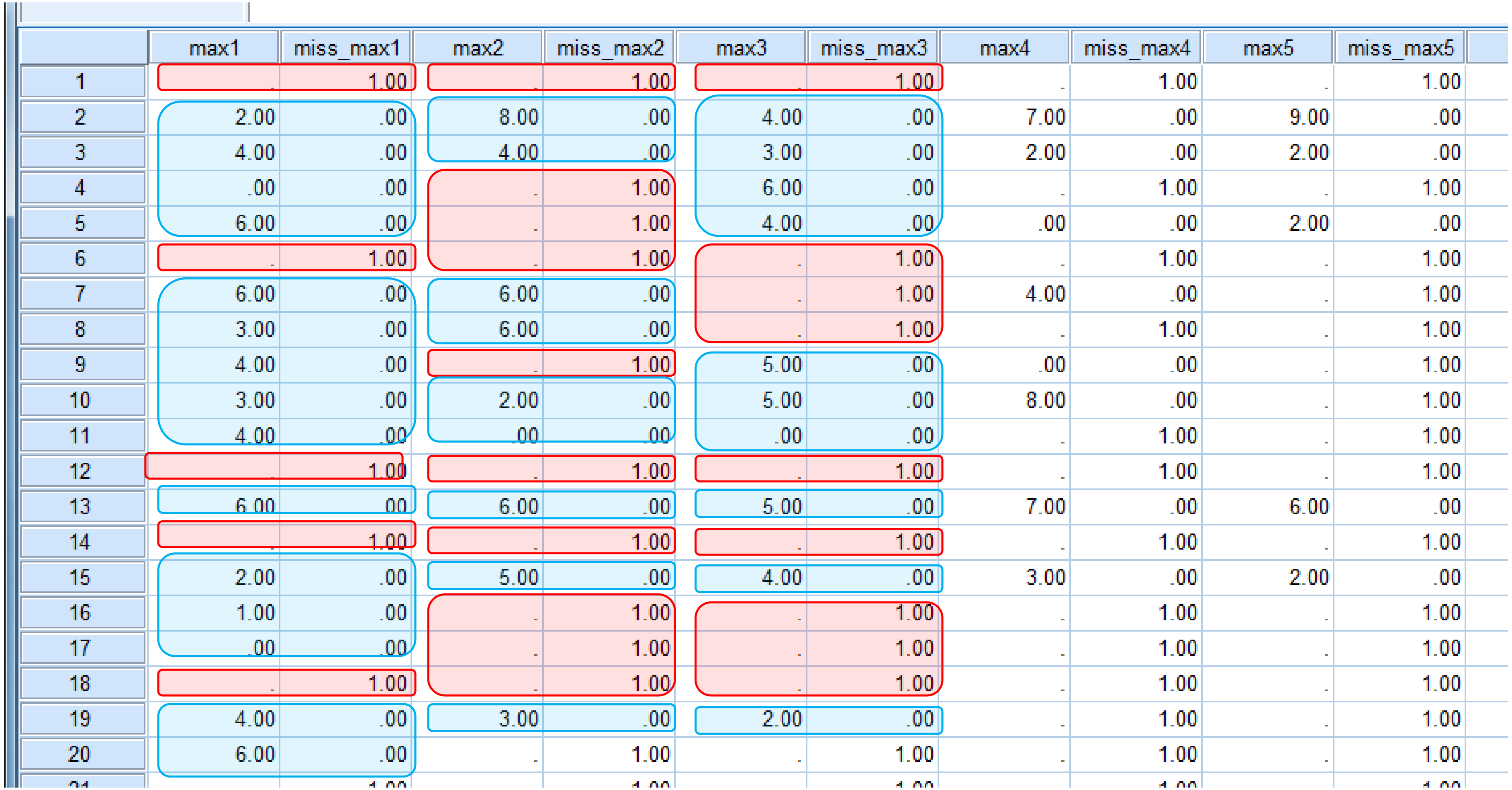


Separate Variance t Tests[a]

		max	max1	max2	max3	max4	max5	max6	consume	drunk	passout	binge
max1	t	-.1	.	1.8	.5	1.0	-1.5	-1.0	1.7	-.8	-.6	1.3
	df	453.9	.	23.2	16.5	23.7	20.6	8.6	483.8	424.8	385.7	478.8
	P(2-tail)	.947	.	.080	.594	.321	.160	.350	.094	.405	.535	.206
	# Present	352	352	278	257	203	167	139	352	351	352	351
	# Missing	209	0	18	16	19	16	9	209	209	209	209
	Mean(Present)	6.2287	5.1179	5.1259	4.9611	4.8916	4.6108	4.1259	3.7045	1.7279	.2188	2.1125
	Mean(Missing)	6.2526	.	3.8333	4.1875	4.0000	6.0000	6.0000	3.3732	1.8612	.2488	1.8947
max2	t	-.3	.6	.	-.3	-.2	-1.1	-.8	1.5	-1.0	.6	.7
	df	557.8	128.0	.	30.1	26.8	20.6	12.3	557.5	547.2	545.8	558.0
	P(2-tail)	.748	.526	.	.801	.869	.295	.457	.141	.319	.534	.457
	# Present	296	278	296	246	200	165	136	296	295	296	295
	# Missing	265	74	0	27	22	18	12	265	265	265	265
	Mean(Present)	6.1841	5.1853	5.0473	4.8862	4.8000	4.6061	4.1434	3.7179	1.7051	.2432	2.0915
	Mean(Missing)	6.2974	4.8649	.	5.1852	4.9545	5.8889	5.3333	3.4283	1.8585	.2151	1.9642
max3	t	-.9	.0	-1.1	.	-1.1	-1.6	-1.0	.2	-1.5	-.9	.8
	df	548.7	176.3	68.5	.	29.3	27.3	13.7	556.2	557.1	553.2	538.5
	P(2-tail)	.345	.972	.268	.	.300	.121	.315	.805	.141	.359	.396
	# Present	273	257	246	273	197	160	135	273	272	273	272
	# Missing	288	95	50	0	25	23	13	288	288	288	288
	Mean(Present)	6.0659	5.1226	4.9146	4.9158	4.6954	4.5063	4.1074	3.6062	1.6618	.2088	2.1066
	Mean(Missing)	6.4003	5.1053	5.7000	.	5.7600	6.3043	5.6154	3.5573	1.8872	.2500	1.9601
max4	t	-.1	.7	.3	.6	.	-.3	-1.0	1.1	-.5	.8	.5
	df	437.9	343.2	212.0	168.2	.	29.7	20.9	454.0	444.7	442.4	436.0
	P(2-tail)	.932	.496	.779	.541	.	.787	.346	.260	.644	.432	.615
	# Present	222	203	200	197	222	158	129	222	221	222	221
	# Missing	339	149	96	76	0	25	19	339	339	339	339
	Mean(Present)	6.2185	5.2438	5.0950	5.0203	4.8153	4.6899	4.0736	3.7207	1.7330	.2523	2.0860
	Mean(Missing)	6.2501	4.9463	4.9479	4.6447	.	5.0000	5.3684	3.4897	1.8068	.2153	1.9956

Identifying Missing Data Patterns: Option 2

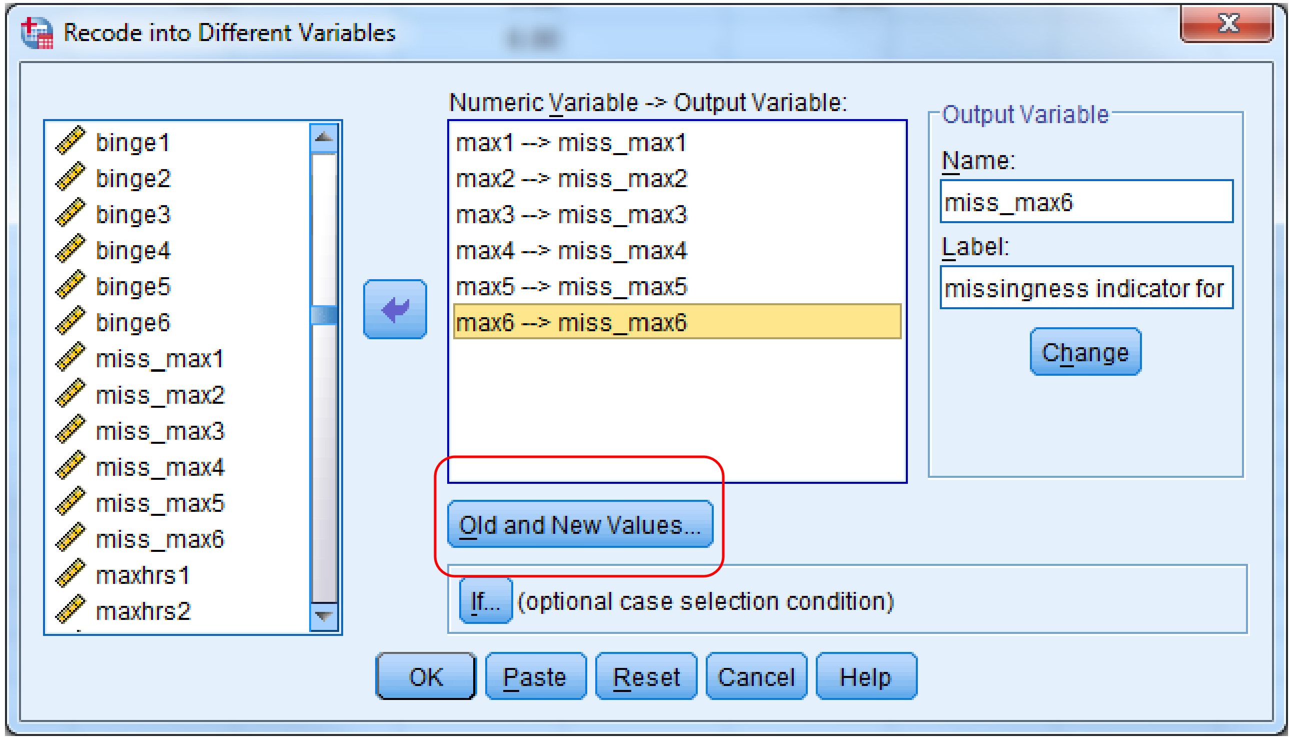
- Create the missingness variables yourself
- Run your own *t* tests or correlations



	max1	miss_max1	max2	miss_max2	max3	miss_max3	max4	miss_max4	max5	miss_max5
1	.	1.00	.	1.00	.	1.00	.	1.00	.	1.00
2	2.00	.00	8.00	.00	4.00	.00	7.00	.00	9.00	.00
3	4.00	.00	4.00	.00	3.00	.00	2.00	.00	2.00	.00
4	.00	.00	.	1.00	6.00	.00	.	1.00	.	1.00
5	6.00	.00	.	1.00	4.00	.00	.00	.00	2.00	.00
6	.	1.00	.	1.00	.	1.00	.	1.00	.	1.00
7	6.00	.00	6.00	.00	.	1.00	4.00	.00	.	1.00
8	3.00	.00	6.00	.00	.	1.00	.	1.00	.	1.00
9	4.00	.00	.	1.00	5.00	.00	.00	.00	.	1.00
10	3.00	.00	2.00	.00	5.00	.00	8.00	.00	.	1.00
11	4.00	.00	.00	.00	.00	.00	.	1.00	.	1.00
12	.	1.00	.	1.00	.	1.00	.	1.00	.	1.00
13	6.00	.00	6.00	.00	5.00	.00	7.00	.00	6.00	.00
14	.	1.00	.	1.00	.	1.00	.	1.00	.	1.00
15	2.00	.00	5.00	.00	4.00	.00	3.00	.00	2.00	.00
16	1.00	.00	.	1.00	.	1.00	.	1.00	.	1.00
17	.00	.00	.	1.00	.	1.00	.	1.00	.	1.00
18	.	1.00	.	1.00	.	1.00	.	1.00	.	1.00
19	4.00	.00	3.00	.00	2.00	.00	.	1.00	.	1.00
20	6.00	.00	.	1.00	.	1.00	.	1.00	.	1.00
21		1.00		1.00		1.00		1.00		1.00

Identifying Missing Data Patterns

- Transform > Recode into Different Variables



Numeric Variable -> Output Variable:

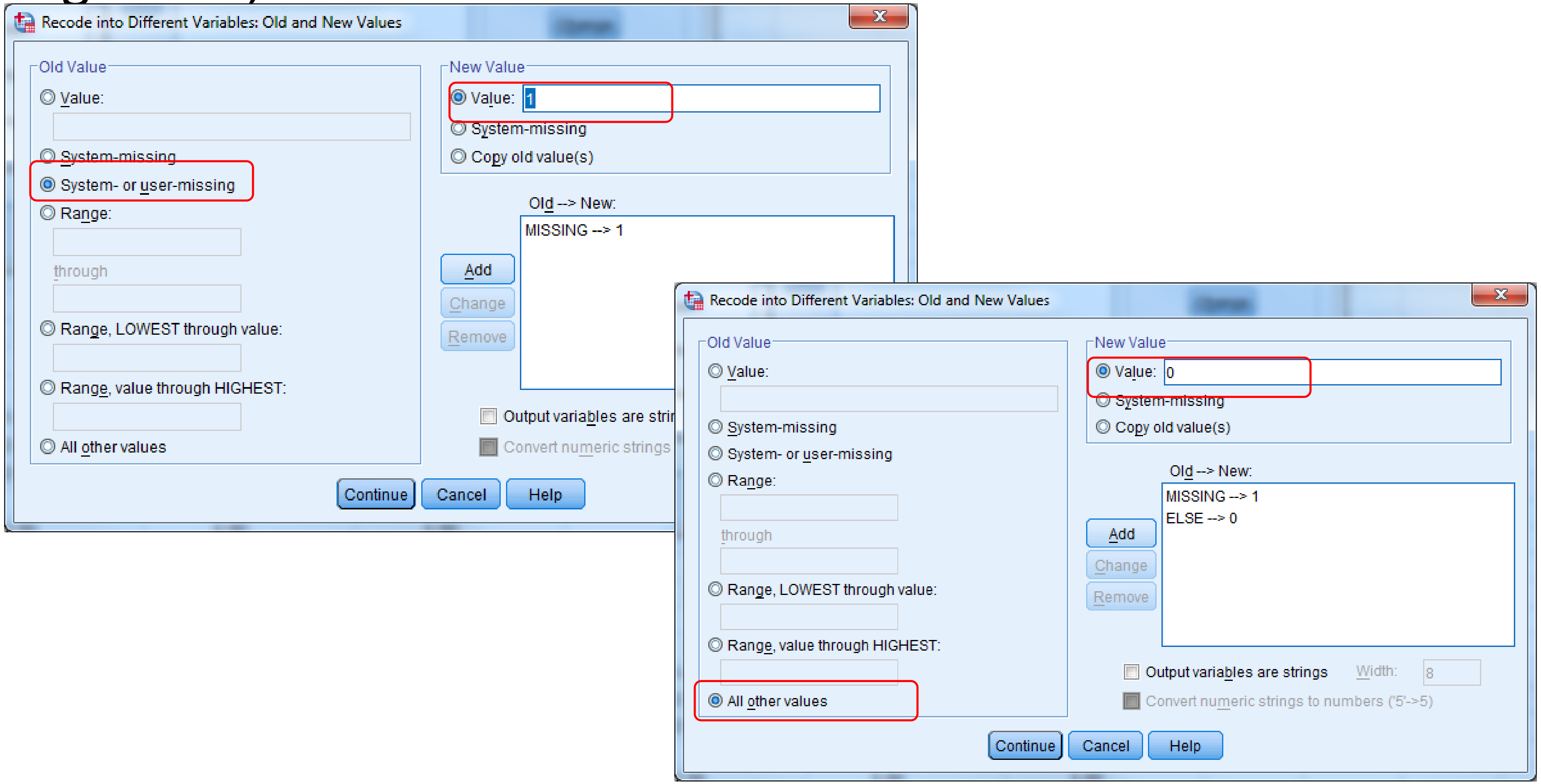
max1 --> miss_max1
max2 --> miss_max2
max3 --> miss_max3
max4 --> miss_max4
max5 --> miss_max5
max6 --> miss_max6

Name: miss_max6

Label: missingness indicator for

Identifying Missing Data Patterns

- Missing into 1, all else into 0



Recode into Different Variables: Old and New Values

Old Value: ● System- or user-missing

New Value: ● Value: 1

Old --> New:

MISSING --> 1

Recode into Different Variables: Old and New Values

Old Value: ● All other values

New Value: ● Value: 0

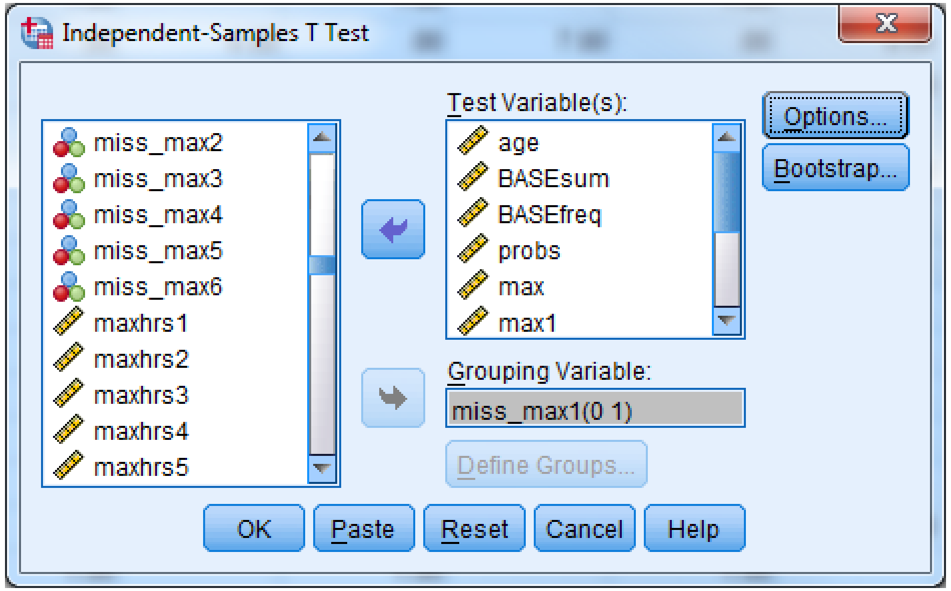
Old --> New:

MISSING --> 1
ELSE --> 0

Old Value
○ Value:
○ System-missing
○ System- or user-missing
○ Range:
through
○ Range, LOWEST through value:
○ Range, value through HIGHEST:
○ All other values
New Value
○ Value:
○ System-missing
○ Copy old value(s)
Add
Change
Remove
□ Output variables are strings
Width: 8
■ Convert numeric strings to numbers ('5'->5)
Continue
Cancel
Help

Identifying Missing Data Patterns

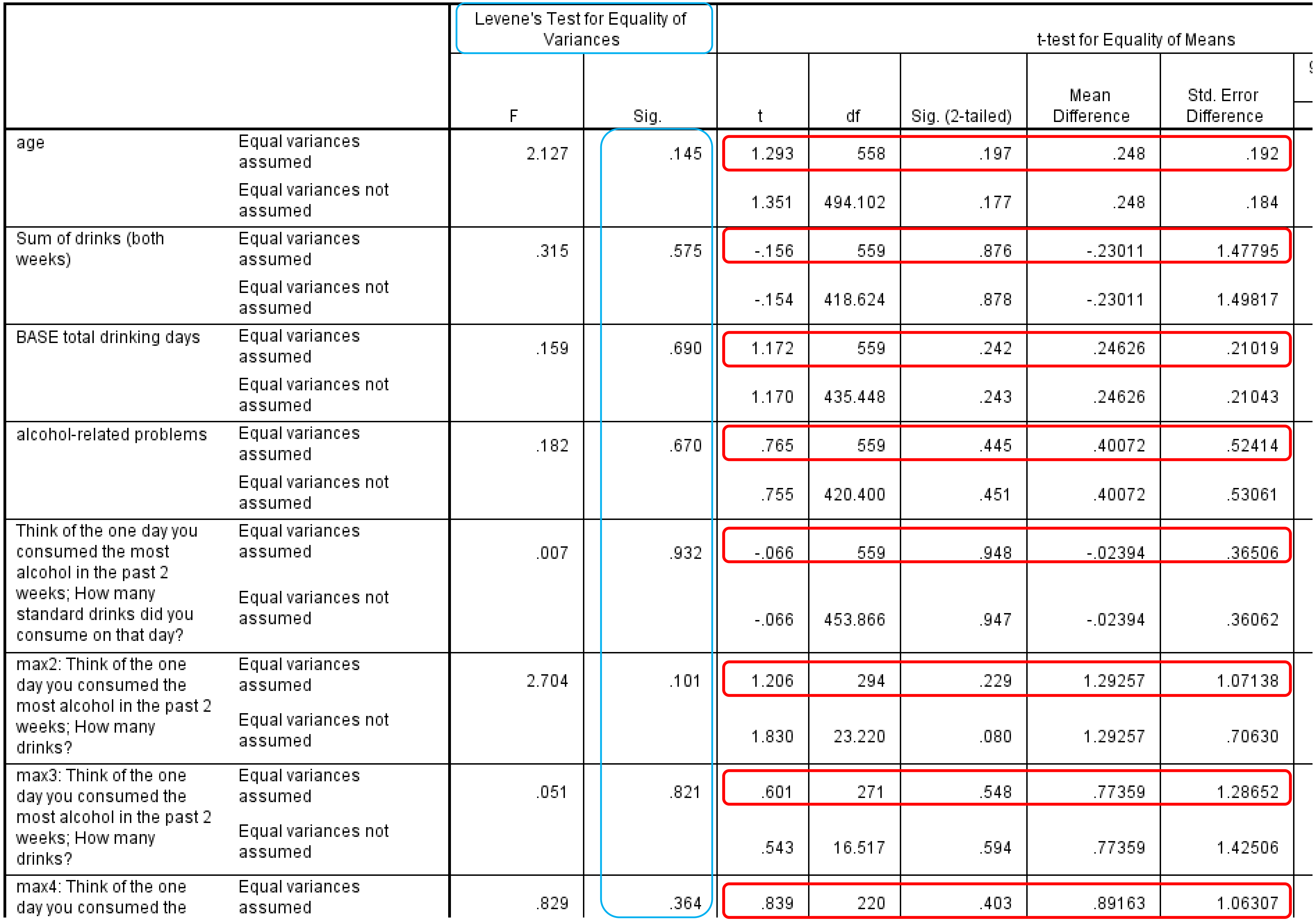
- Conduct *t* tests (or correlations, or chi-squares) to detect associations between missingness for that variable and the values of other variables in the dataset



Independent-Samples T Test

Test Variable(s):
age
BASEsum
BASEfreq
probs
max
max1

Grouping Variable:
miss_max1(0 1)



		Levene's Test for Equality of Variances		t-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
age	Equal variances assumed	2.127	.145	1.293	558	.197	.248	.192
	Equal variances not assumed			1.351	494.102	.177	.248	.184
Sum of drinks (both weeks)	Equal variances assumed	.315	.575	-.156	559	.876	-.23011	1.47795
	Equal variances not assumed			-.154	418.624	.878	-.23011	1.49817
BASE total drinking days	Equal variances assumed	.159	.690	1.172	559	.242	.24626	.21019
	Equal variances not assumed			1.170	435.448	.243	.24626	.21043
alcohol-related problems	Equal variances assumed	.182	.670	.765	559	.445	.40072	.52414
	Equal variances not assumed			.755	420.400	.451	.40072	.53061
Think of the one day you consumed the most alcohol in the past 2 weeks; How many standard drinks did you consume on that day?	Equal variances assumed	.007	.932	-.066	559	.948	-.02394	.36506
	Equal variances not assumed			-.066	453.866	.947	-.02394	.36062
max2: Think of the one day you consumed the most alcohol in the past 2 weeks; How many drinks?	Equal variances assumed	2.704	.101	1.206	294	.229	1.29257	1.07138
	Equal variances not assumed			1.830	23.220	.080	1.29257	.70630
max3: Think of the one day you consumed the most alcohol in the past 2 weeks; How many drinks?	Equal variances assumed	.051	.821	.601	271	.548	.77359	1.28652
	Equal variances not assumed			.543	16.517	.594	.77359	1.42506
max4: Think of the one day you consumed the	Equal variances assumed	.829	.364	.839	220	.403	.89163	1.06307

Missing Data Patterns

- Missingness for “maximum drinks” at follow-up 1 was unrelated to age, quantity of alcohol drinks consumed, frequency of drinking, alcohol-related problems experienced, maximum drinks at baseline, maximum drinks at follow-up 2, maximum drinks at follow-up 3, etc.
- IF missingness for “maximum drinks” HAD been significantly associated with a variable, you’d want to include that variable in your model (for ML estimation), or in the variables used for imputation

Missing Data Patterns

- Code missingness (R_t)
- For cross-sectional data, would do this variable by variable
- For longitudinal data, important to code each time point
- Compare constructs for those with complete vs missing data
    - Complete vs any missing?
    - $t1$ present vs missing? $t2$ present vs missing?
    - Number of missing time points (0, 1, 2, 3, 4)?
    - ***All of the above***

Types of Missing Data

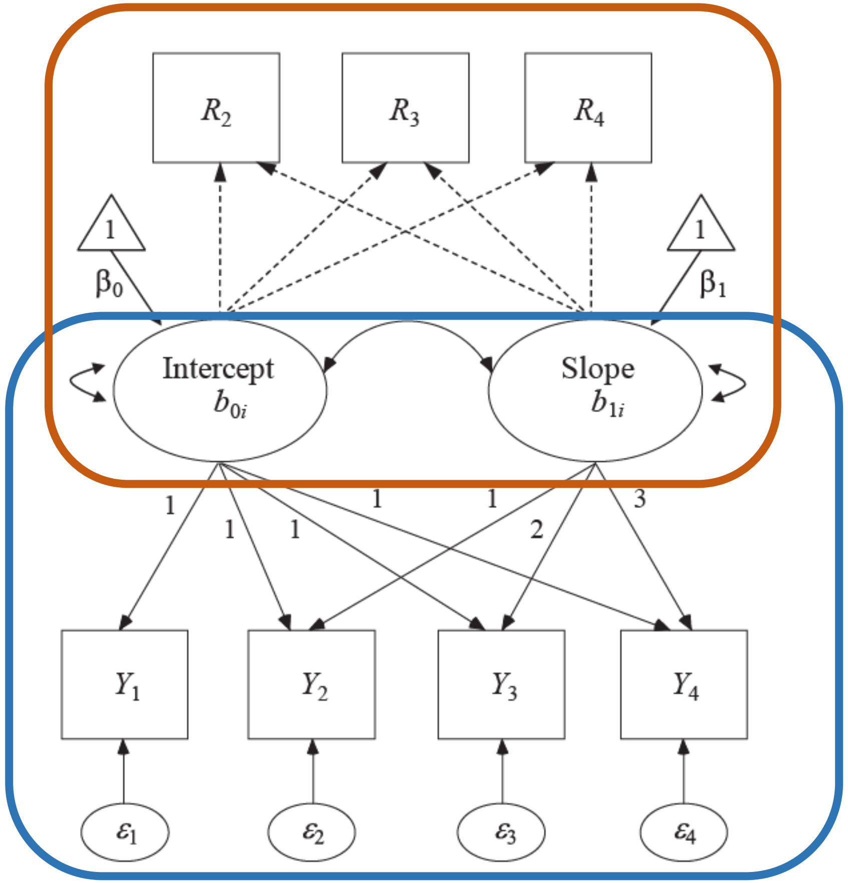
- MCAR
  - Compatible with all analyses; no limits on what can be done with data
- MAR
  - The datapoint is missing because of a value recorded in the dataset
    - Skipped income question is strongly correlated with education question
      - Good to ask multiple items regarding potentially sensitive questions (or at least, constructs relating to it)
    - Skipped that timepoint because I am a heavy drinker and not good at following through with surveys
      - Missing time 3 alcohol use BECAUSE OF **TIME 1** ALCOHOL USE
  - Easiest to detect
  - Compatible with *most* fun analyses
    - **IF** you do some prep work such as imputation or carefully choosing estimation method
    - Minimizes bias and increases power
    - Data are missing in a systematic way that we can control for
  - Rarely limits what can be done with the data

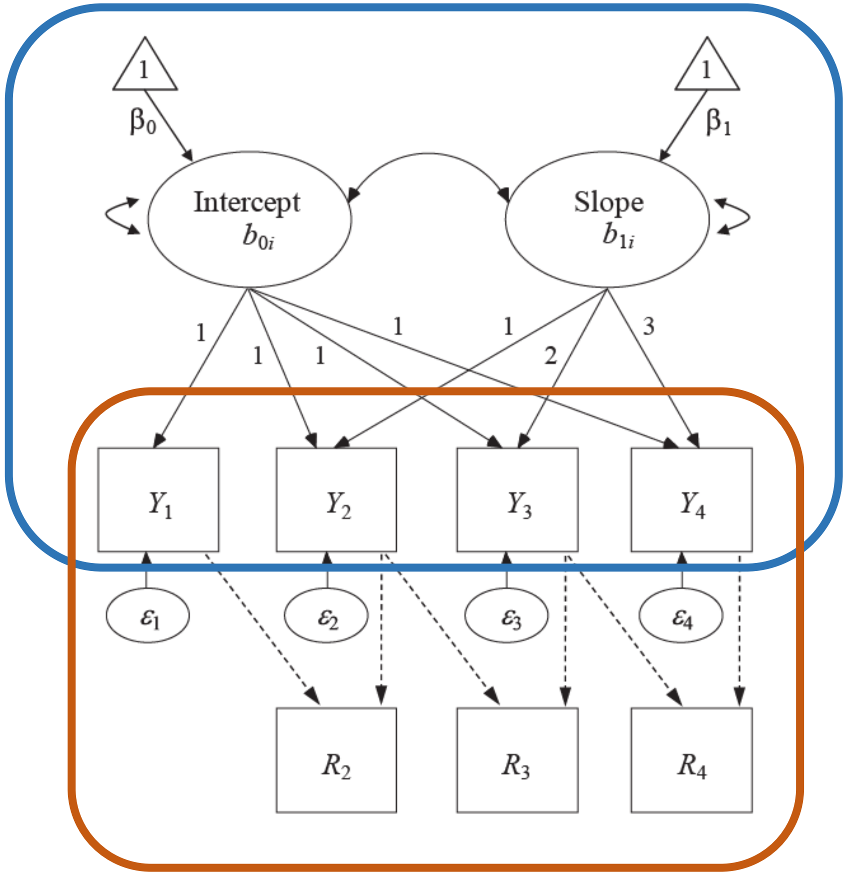
Types of Missing Data

- MNAR requires a change in analysis approach (Enders, 2011)
  - Allow for association between outcome variable and the propensity for missing data
  - Longitudinal: section model and pattern mixture model
    - Not without assumptions, particularly about normality
  - Enders says execution is easy in Mplus
  - Note that he did not invent these models, but does a great job explaining them

Selection Models (Enders, 2011)

- Combine substantive proposed model with prediction of missingness
- *R_t* is missing data indicator for time *t*

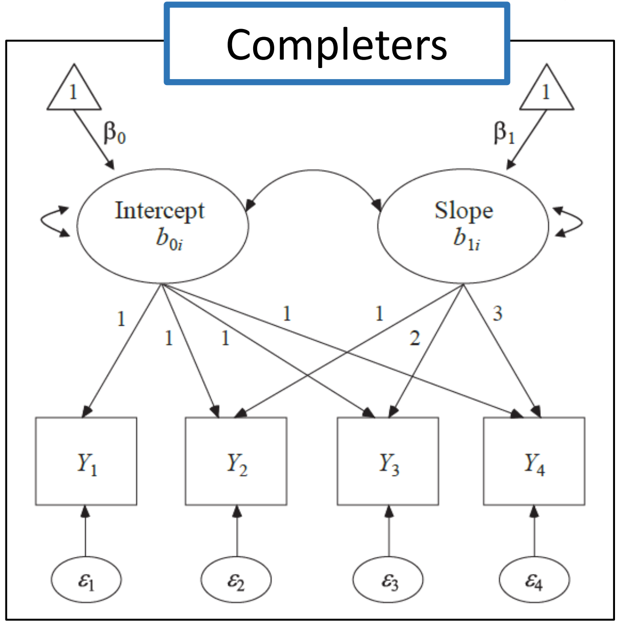




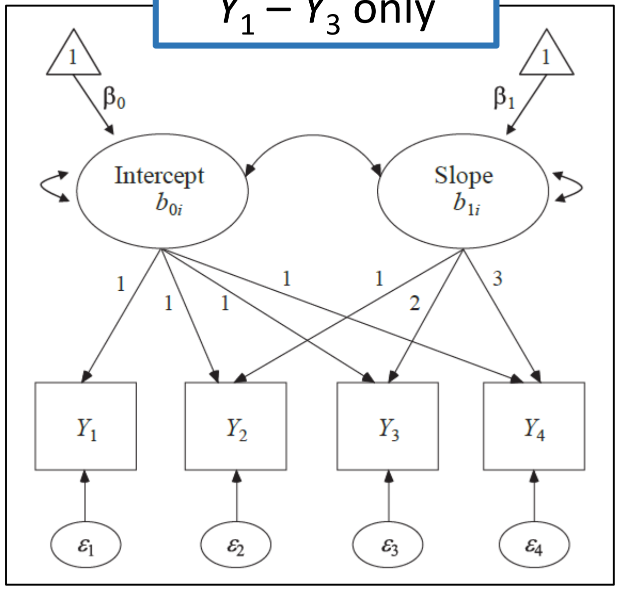
Pattern Mixture Models (Enders, 2011)

- Identify subgroup that share the same missing data pattern
  - Complete cases vs drop out after t3, drop out after t2, drop out after baseline
    - Assuming they stay out, 4 groups for 4 timepoints
- Estimate model for each subgroup
  - 4 intercepts, 4 slopes

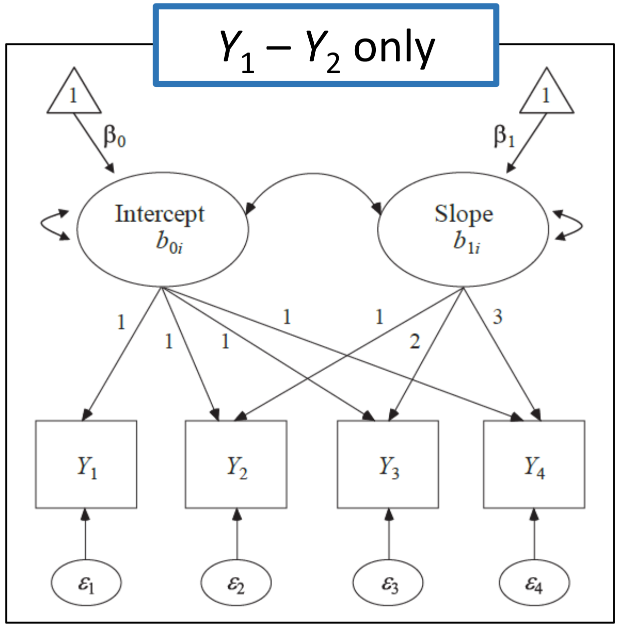
Completers



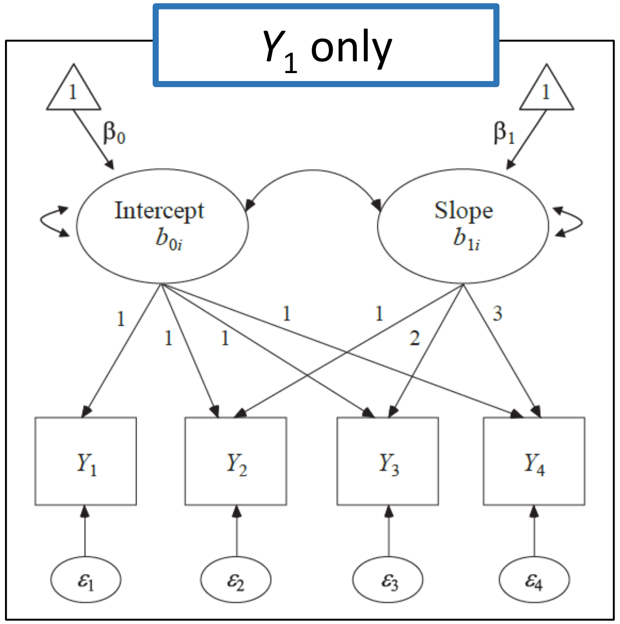
$Y_1 – Y_3$ only



$Y_1 – Y_2$ only



Y_1 only



Pattern Mixture Models (Enders, 2011)

- With 4 models, want overall pattern (average across distribution of missingness)
  - Mplus can do this
- Looks a lot like multigroup, but do not use multiple approach
  - Lose benefits of pattern-specific conditioning
  - Computes SEs a different way
- Model identification is tricky (especially for higher dropout groups)
  - Enders has some suggestions
  - Again, incorporating missingness as a predictor/outcome can help
- This approach gets messy with more patterns of missingness (e.g., lots of timepoints, dropout is not permanent)
- Need large samples to support the multiple models

Sensitivity Analyses

- In addition to your main hypothesized model, fit multiple MNAR models
- Each alternative model should free an assumption of the original hypothesized model
    - Run these models using MI
- If conclusions/inferences are fairly stable across models, you can be fairly certain about your conclusions
    - Does not tell you if your data are MNAR, but tells you the degree to which your model would be influenced IF they did not meet MAR assumptions
- Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.
    - "Accessible to substantive readers while providing a level of detail that will satisfy quantitative specialists."
- National Research Council (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
    - Free

What to do about Missing Data

- Delete incomplete cases?
- Complete Case Analysis (aka Listwise Deletion)
  - Delete everyone from your sample who has missing data
  - Final sample includes only individuals with all data
- Available Case Analysis (aka Pairwise Deletion)
  - Exclude people from relevant analyses who have missing data
  - E.g., If missing on depressive symptoms, then missing from regression that examines influence of meditation on depressive symptoms
    - Present for regression that examines influence of meditation on anxiety

Deleting/Omitting Incomplete Cases

- Reduces sample size
  - Reduces power (or ability to detect effects)
- Introduces bias (unless MCAR)
  - May be eliminating people who drink the most, or have the lowest salary, or experience the worse trauma, or are not responding to the medication, etc.
  - May lead to falsely non-significant results
    - e.g., without high-trauma individuals in sample, cannot detect association between alcohol and trauma
  - May lead to falsely significant effects
    - e.g., if people who don’t get better drop out, those who remain lead to false belief the drug leads to lower depression scores
- Bad choice unless data are MCAR ☹
- What SPSS does by default if you do nothing to address missingness!

Imputation

- Impute/replace missing values (.) with best estimates
- A lot of methods exist that were once popular, but not great
  - Easy to do (easier; not computationally intensive)
- Mean imputation
  - Person mean (average of their responses for other items on that scale)
  - Item mean (average of everyone else in the sample)
  - Does not take all known information into account
  - Falsely deflates random error (makes smaller SDs and smaller SEs)
- Regression Imputation (or Conditional Mean Imputation)
  - Predicted value of Y given values of X
  - Falsely deflates random error (makes smaller SDs and smaller SEs)
- Hot Deck Imputation
  - Find someone similar in the sample, and use their value
- Last Observation Carried Forward (longitudinal only)
  - Replace missing follow-up values with the last value the participant reported
  - Assumes no change, which is typically wrong

Outdated Imputation

- The aforementioned methods were the best available methods for decades
- Can be seen in older (and some newer) published articles
- Can be executed using older versions of SPSS, or via hand calculations
- All tend to falsely reduce your SD/SE
- Also tend to falsely increase test statistics (*t*, *F*, *r*, etc.)
- In some cases, are worse (more biased) than deletion
- Need a technique that accounts for the *uncertainty* about the value of the missing parameter

Better Imputation

- Better methods exist that take into account the uncertainty of imputation
- Been around for decades, but becoming popular because software is starting to incorporate them more easily
- All require MAR at least (could be MCAR)
- Multiple Imputation
  - Generates multiple datasets that represent multiple imputed values possible
- Maximum Likelihood (ML) Estimation
  - Not actually imputing data, but using all available data
- Expectation Maximization (EM) Imputation
  - Iterative 2-step process

Historical Methods are Flawed

- Simulation data: Hallgren & Witkiewitz, 2013
- Real data: Witkiewitz et al., 2014
- Took real datasets and poked holes in them:
  - At random (MCAR)
  - Based on observable data X (MAR)
  - Or based on unobserved data Y (MNAR)
  - Like our class exercise
- Tried listwise and pairwise deletion (i.e., excluding cases from relevant analyses if they weren’t complete)
- Tried lots of single imputation methods
- Tried ML estimation and MI

Historical Methods are Flawed

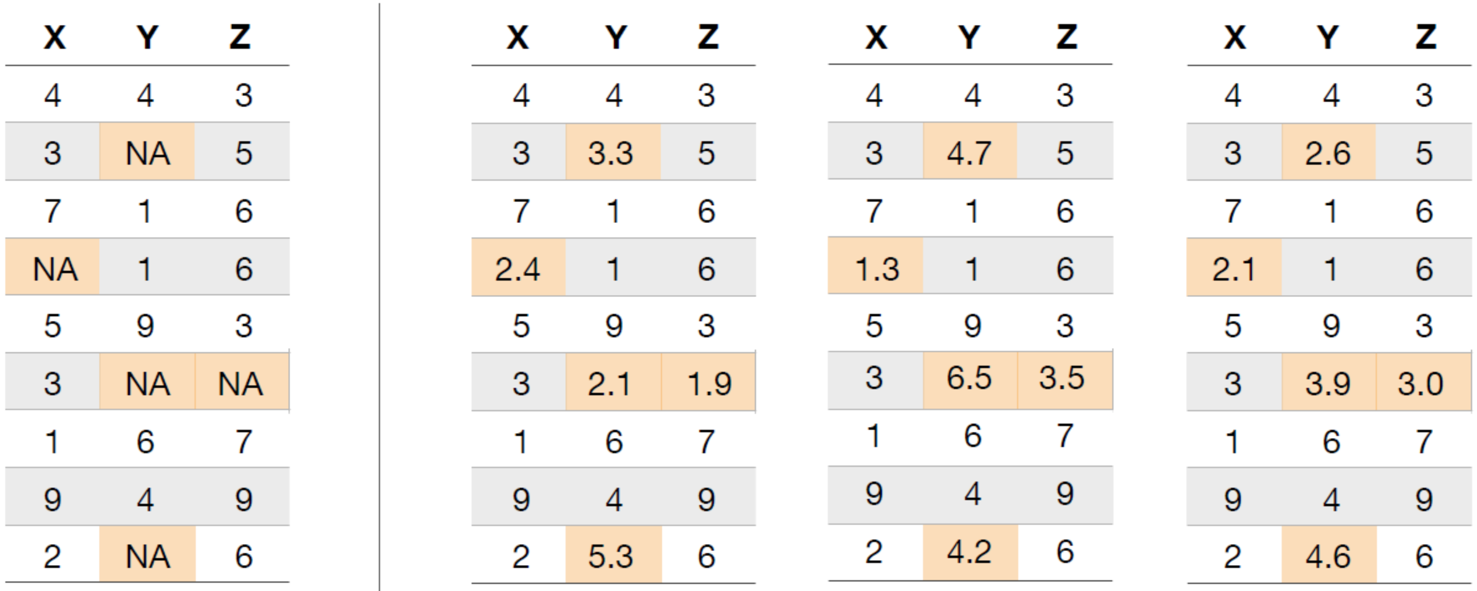
- “Consistent with 30 years of prior research on missing data approaches, the findings from both studies clearly showed full information maximum likelihood and multiple imputation to generate the least biased estimates of intervention effects in alcohol clinical trials”
  - If you have MAR data and do nothing (i.e., deletion), results will be biased
- “Although they are the preferred approaches, full information maximum likelihood estimation and multiple imputation can introduce bias if data are missing not at random. **Sensitivity analyses** to examine the effects of different missing data models can be useful in evaluating the impact of missing data on the analysis of primary outcomes”
  - Enders, 2010 for example

Multiple Imputation (MI)

- Generates multiple datasets with different possible values for each missing datapoint
  - For each dataset, the observed datapoints stay the same
  - The missing values are different in each version
  - Imputed values represent a combination of the linear regression of that variable on other variables of interest, plus random error (a random draw from the residual normal distribution for that variable)
- Main analysis is conducted on all datasets, and results are combined
- Computationally difficult (getting easier)
  - Easy to incorporate in SPSS for a regression, mean, etc.
  - Easily done now in Mplus, or done in “blimp” and imported into other software
  - Layering MI on top of complex analyses can become overly complicated
- Gives slightly different results every time you do it
  - It is possible to specify the random “seed” so that you get the same results every time, but this removes one of the benefits of MI (introducing random error)

Multiple Imputation (MI)

- Generates multiple datasets with different possible values for each missing datapoint
  - For each dataset, the observed datapoints stay the same
  - The missing values are different in each version
  - Imputed values represent a combination of the linear regression of that variable on other variables of interest, plus random error (a random draw from the residual normal distribution for that variable)



X	Y	Z
4	4	3
3	NA	5
7	1	6
NA	1	6
5	9	3
3	NA	NA
1	6	7
9	4	9
2	NA	6

X	Y	Z
4	4	3
3	3.3	5
7	1	6
2.4	1	6
5	9	3
3	2.1	1.9
1	6	7
9	4	9
2	5.3	6

X	Y	Z
4	4	3
3	4.7	5
7	1	6
1.3	1	6
5	9	3
3	6.5	3.5
1	6	7
9	4	9
2	4.2	6

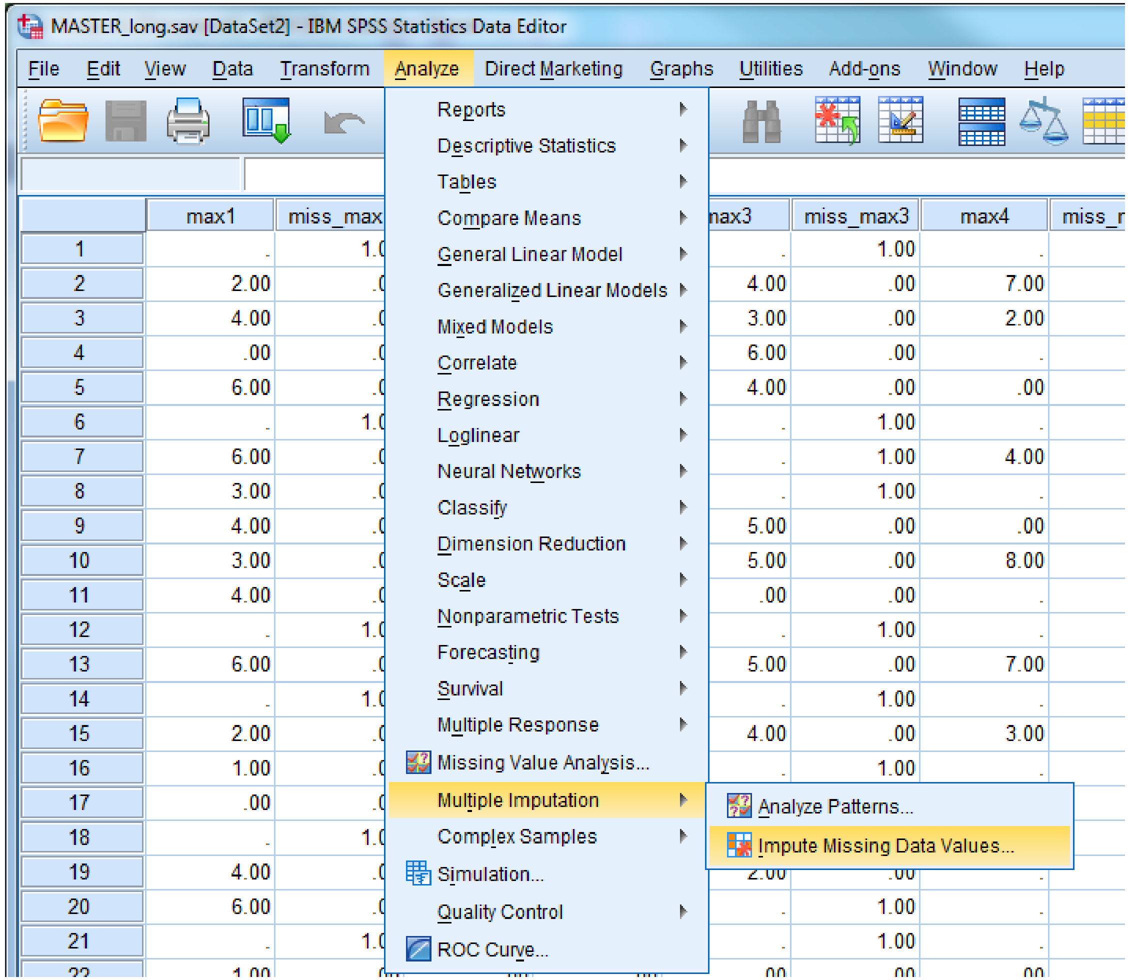
X	Y	Z
4	4	3
3	2.6	5
7	1	6
2.1	1	6
5	9	3
3	3.9	3.0
1	6	7
9	4	9
2	4.6	6

How to do MI?

- Can do it in the software that will also be used for the analysis
  - SPSS
    - Default is 5 iterations, but want at least 20
  - Mplus
  - SAS
- Can do it in MI software, then export to another program for analysis
  - Blimp!
  - Mplus, SAS, etc.

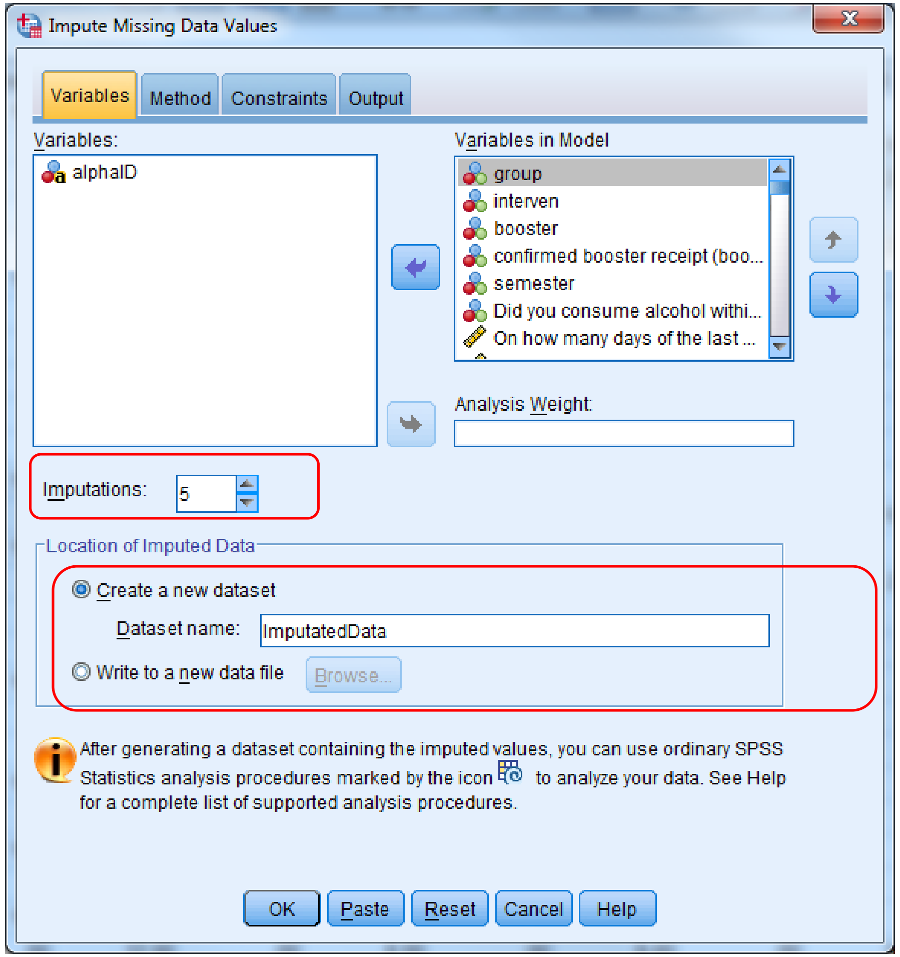
MI in SPSS

- Analyze > Multiple Imputation > Impute Missing Data Values



	max1	miss_max	nax3	miss_max3	max4	miss_r
1	.	1.(	.	1.00	.	
2	2.00	.(	4.00	.00	7.00	
3	4.00	.(	3.00	.00	2.00	
4	.00	.(	6.00	.00	.	
5	6.00	.(	4.00	.00	.00	
6	.	1.(	.	1.00	.	
7	6.00	.(	.	1.00	4.00	
8	3.00	.(	.	1.00	.	
9	4.00	.(	5.00	.00	.00	
10	3.00	.(	5.00	.00	8.00	
11	4.00	.(	.00	.00	.	
12	.	1.(	.	1.00	.	
13	6.00	.(	5.00	.00	7.00	
14	.	1.(	.	1.00	.	
15	2.00	.(	4.00	.00	3.00	
16	1.00	.(	.	1.00	.	
17	.00	.(				
18	.	1.(				
19	4.00	.(	2.00	.00	.	
20	6.00	.(	.	1.00	.	
21	.	1.(	.	1.00	.	
22	1.00		.00	.00	.00	

MI in SPSS



Variables:

- alphaID

Variables in Model

- group
- interven
- booster
- confirmed booster receipt (boo...
- semester
- Did you consume alcohol withi...
- On how many days of the last ...

Analysis Weight:

Imputations: 5

Location of Imputed Data

● Create a new dataset

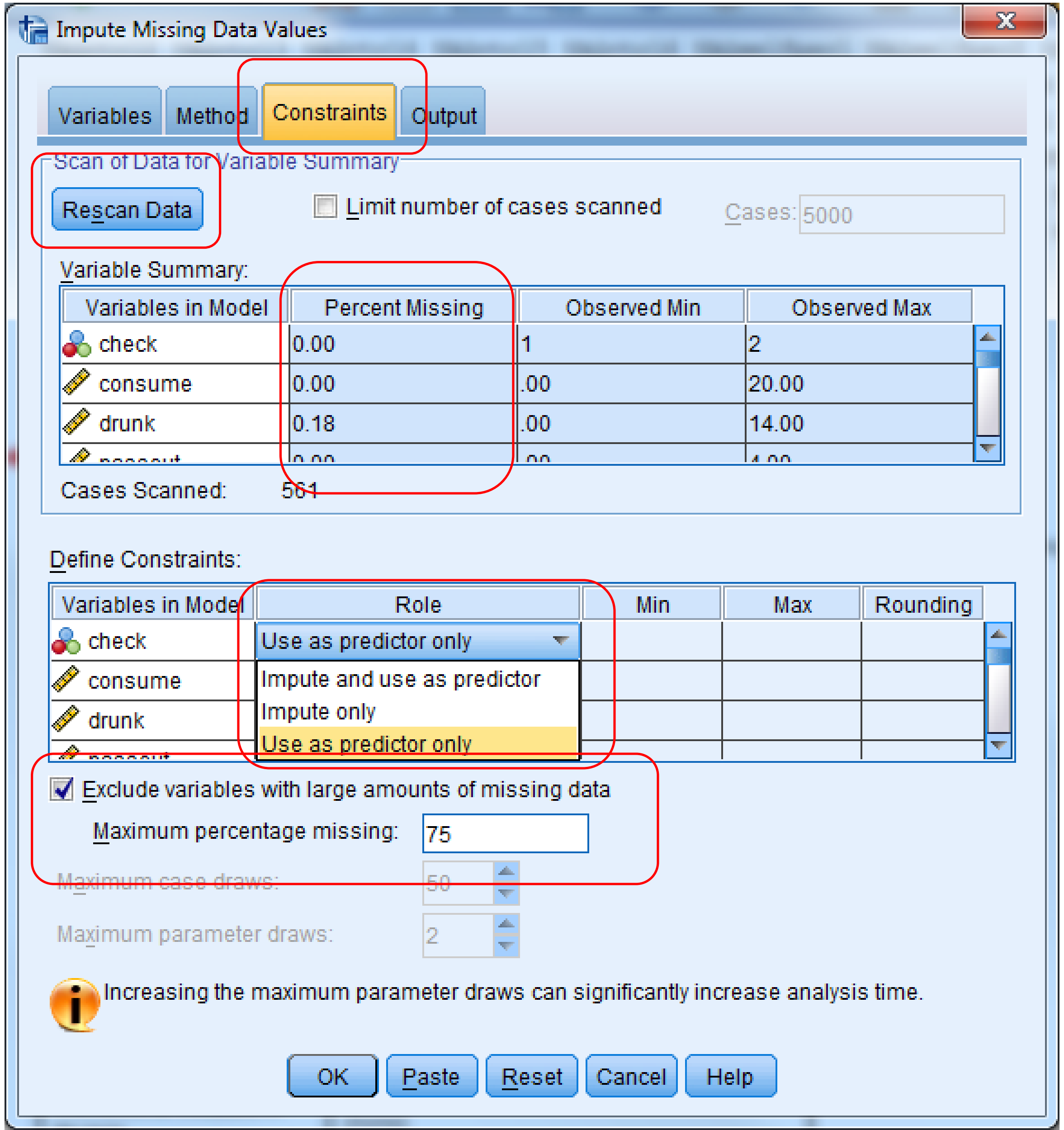
Dataset name: ImputatedData

○ Write to a new data file

After generating a dataset containing the imputed values, you can use ordinary SPSS Statistics analysis procedures marked by the icon to analyze your data. See Help for a complete list of supported analysis procedures.

1. Select all numeric variables
   a. Except ID
2. Choose number of datasets/imputations to generate
   a. **Default is 5. Never use 5! Use at least 20 (Enders).**
3. Give a name to your new dataset
   a. ImputedData

MI in SPSS



Scan of Data for Variable Summary

☐ Limit number of cases scanned   Cases: 5000

Variable Summary:

Variables in Model	Percent Missing	Observed Min	Observed Max
check	0.00	1	2
consume	0.00	.00	20.00
drunk	0.18	.00	14.00

Cases Scanned: 561

Define Constraints:

Variables in Model	Role	Min	Max	Rounding
check	Use as predictor only			
consume				
drunk				

☑ Exclude variables with large amounts of missing data

Maximum percentage missing: 75

Maximum case draws: 50

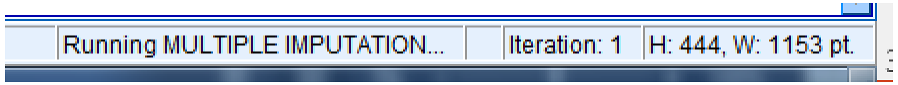
Maximum parameter draws: 2

Increasing the maximum parameter draws can significantly increase analysis time.

- Can fail if too many variables and/or too much missing data for some variables
- Might pare the dataset down to 500 variables or fewer
  - Can do the process again if you identify a different 500 variables for other analyses
- Can limit what is imputed versus what is a predictor
- Can exclude variables with too much missing data
  - If not main outcome

MI in SPSS

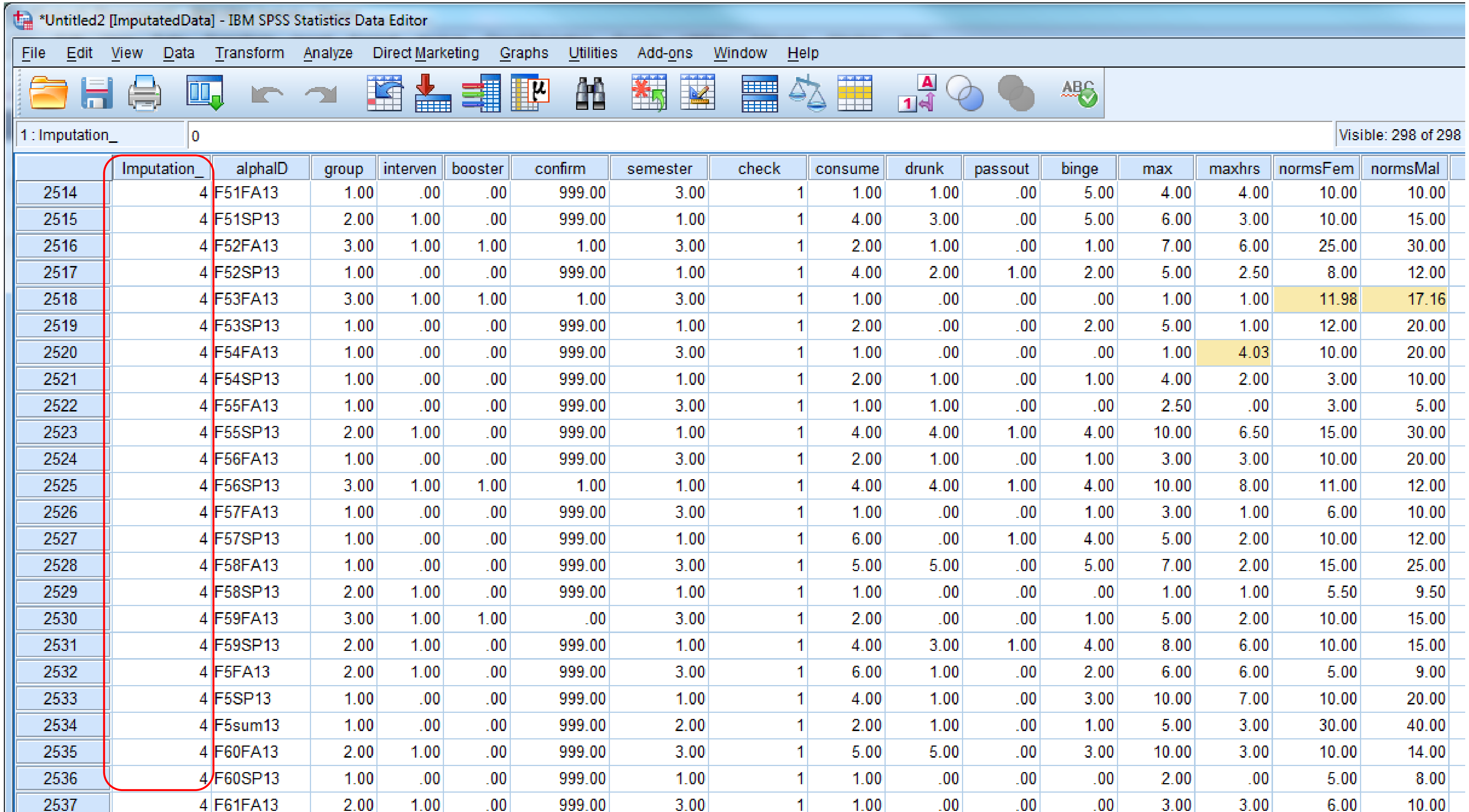
- Might take a while



Running MULTIPLE IMPUTATION... Iteration: 1 H: 444, W: 1153 pt.

- New imputed data file(s) will open in a new window
- Can scroll down to see multiple datasets
- Imputed values are highlighted

MI in SPSS



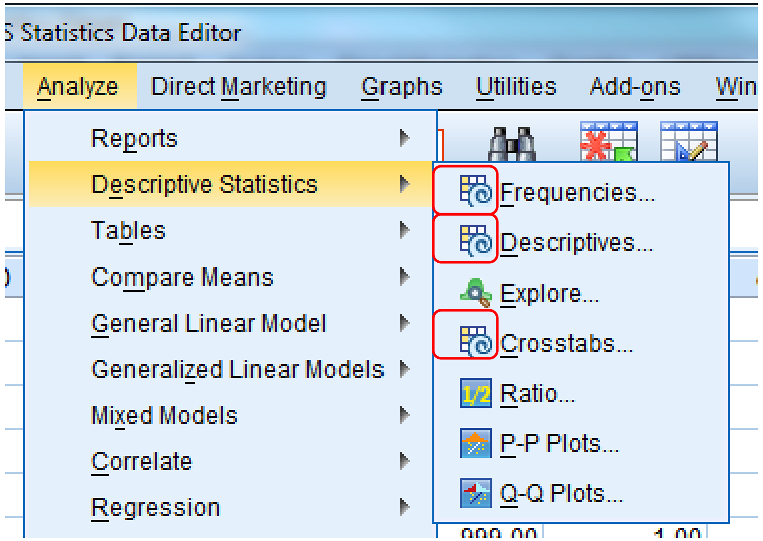
1 : Imputation_ 0

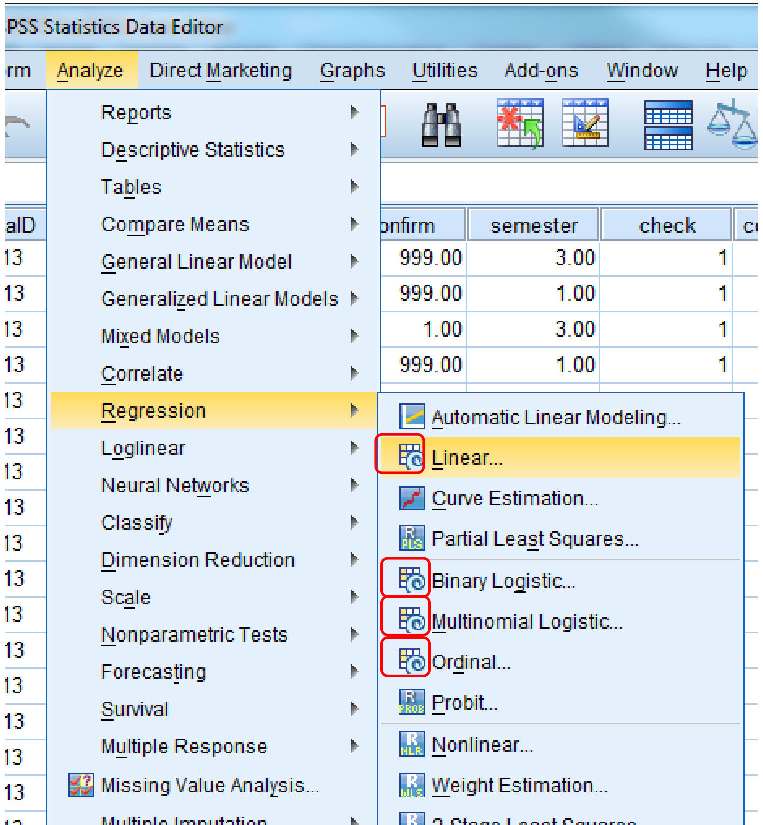
Visible: 298 of 298

	Imputation_	alphaID	group	interven	booster	confirm	semester	check	consume	drunk	passout	binge	max	maxhrs	normsFem	normsMal
2514	4	F51FA13	1.00	.00	.00	999.00	3.00	1	1.00	1.00	.00	5.00	4.00	4.00	10.00	10.00
2515	4	F51SP13	2.00	1.00	.00	999.00	1.00	1	4.00	3.00	.00	5.00	6.00	3.00	10.00	15.00
2516	4	F52FA13	3.00	1.00	1.00	1.00	3.00	1	2.00	1.00	.00	1.00	7.00	6.00	25.00	30.00
2517	4	F52SP13	1.00	.00	.00	999.00	1.00	1	4.00	2.00	1.00	2.00	5.00	2.50	8.00	12.00
2518	4	F53FA13	3.00	1.00	1.00	1.00	3.00	1	1.00	.00	.00	.00	1.00	1.00	11.98	17.16
2519	4	F53SP13	1.00	.00	.00	999.00	1.00	1	2.00	.00	.00	2.00	5.00	1.00	12.00	20.00
2520	4	F54FA13	1.00	.00	.00	999.00	3.00	1	1.00	.00	.00	.00	1.00	4.03	10.00	20.00
2521	4	F54SP13	1.00	.00	.00	999.00	1.00	1	2.00	1.00	.00	1.00	4.00	2.00	3.00	10.00
2522	4	F55FA13	1.00	.00	.00	999.00	3.00	1	1.00	1.00	.00	.00	2.50	.00	3.00	5.00
2523	4	F55SP13	2.00	1.00	.00	999.00	1.00	1	4.00	4.00	1.00	4.00	10.00	6.50	15.00	30.00
2524	4	F56FA13	1.00	.00	.00	999.00	3.00	1	2.00	1.00	.00	1.00	3.00	3.00	10.00	20.00
2525	4	F56SP13	3.00	1.00	1.00	1.00	1.00	1	4.00	4.00	1.00	4.00	10.00	8.00	11.00	12.00
2526	4	F57FA13	1.00	.00	.00	999.00	3.00	1	1.00	.00	.00	1.00	3.00	1.00	6.00	10.00
2527	4	F57SP13	1.00	.00	.00	999.00	1.00	1	6.00	.00	1.00	4.00	5.00	2.00	10.00	12.00
2528	4	F58FA13	1.00	.00	.00	999.00	3.00	1	5.00	5.00	.00	5.00	7.00	2.00	15.00	25.00
2529	4	F58SP13	2.00	1.00	.00	999.00	1.00	1	1.00	.00	.00	.00	1.00	1.00	5.50	9.50
2530	4	F59FA13	3.00	1.00	1.00	.00	3.00	1	2.00	.00	.00	1.00	5.00	2.00	10.00	15.00
2531	4	F59SP13	2.00	1.00	.00	999.00	1.00	1	4.00	3.00	1.00	4.00	8.00	6.00	10.00	15.00
2532	4	F5FA13	2.00	1.00	.00	999.00	3.00	1	6.00	1.00	.00	2.00	6.00	6.00	5.00	9.00
2533	4	F5SP13	1.00	.00	.00	999.00	1.00	1	4.00	1.00	.00	3.00	10.00	7.00	10.00	20.00
2534	4	F5sum13	1.00	.00	.00	999.00	2.00	1	2.00	1.00	.00	1.00	5.00	3.00	30.00	40.00
2535	4	F60FA13	2.00	1.00	.00	999.00	3.00	1	5.00	5.00	.00	3.00	10.00	3.00	10.00	14.00
2536	4	F60SP13	1.00	.00	.00	999.00	1.00	1	1.00	.00	.00	.00	2.00	.00	5.00	8.00
2537	4	F61FA13	2.00	1.00	.00	999.00	3.00	1	1.00	.00	.00	.00	3.00	3.00	6.00	10.00

MI in SPSS

- Run your analyses from these new data
- Swirls will appear over analyses that can incorporate the multiply imputed data
- From simple (means, correlations) to complex (chi-squares, regressions)



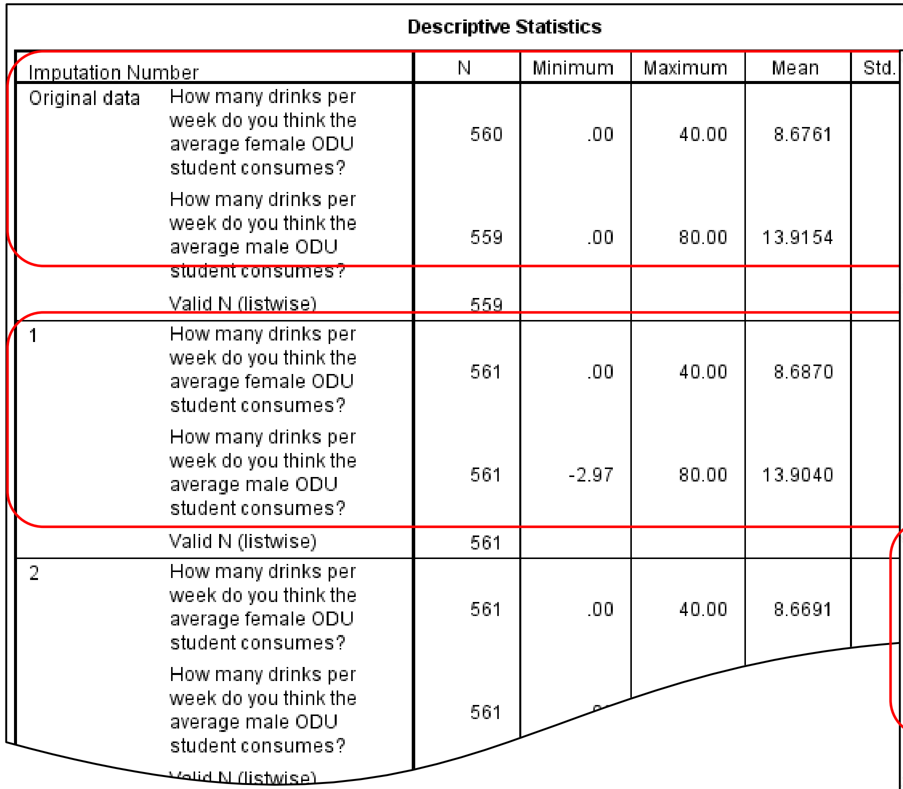


MI in SPSS

- Will display results for: 1) original dataset, 2) each imputed dataset, and 3) **pooled results** across imputations

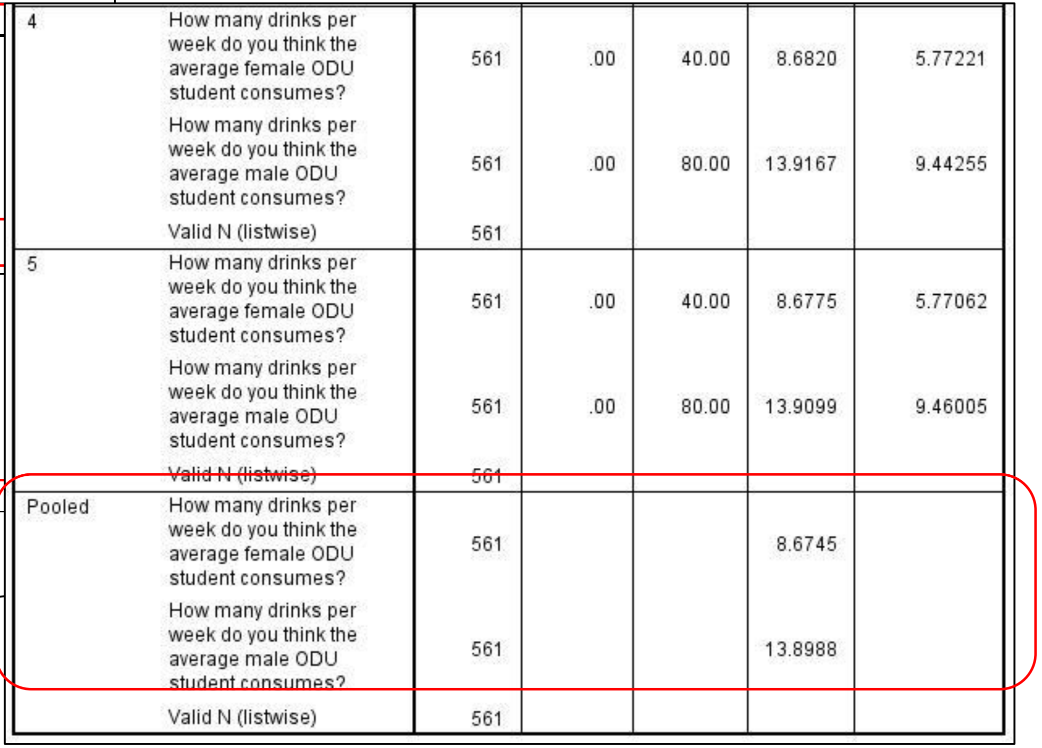
Notice not everything is presented in pooled results!

No min, max, or *SD!* That can be problematic



Descriptive Statistics

Imputation Number		N	Minimum	Maximum	Mean	Std.
Original data	How many drinks per week do you think the average female ODU student consumes?	560	.00	40.00	8.6761	
	How many drinks per week do you think the average male ODU student consumes?	559	.00	80.00	13.9154	
	Valid N (listwise)	559				
1	How many drinks per week do you think the average female ODU student consumes?	561	.00	40.00	8.6870	
	How many drinks per week do you think the average male ODU student consumes?	561	-2.97	80.00	13.9040	
	Valid N (listwise)	561				
2	How many drinks per week do you think the average female ODU student consumes?	561	.00	40.00	8.6691	
	How many drinks per week do you think the average male ODU student consumes?	561				
	Valid N (listwise)					



4	How many drinks per week do you think the average female ODU student consumes?	561	.00	40.00	8.6820	5.77221
	How many drinks per week do you think the average male ODU student consumes?	561	.00	80.00	13.9167	9.44255
	Valid N (listwise)	561				
5	How many drinks per week do you think the average female ODU student consumes?	561	.00	40.00	8.6775	5.77062
	How many drinks per week do you think the average male ODU student consumes?	561	.00	80.00	13.9099	9.46005
	Valid N (listwise)	561				
Pooled	How many drinks per week do you think the average female ODU student consumes?	561			8.6745	
	How many drinks per week do you think the average male ODU student consumes?	561			13.8988	
	Valid N (listwise)	561				

MI in Mplus

- Just generating data (from Mplus User’s Guide, version 7)

```
TITLE: this is an example of multiple imputation
for a set of variables with missing values
DATA: FILE = ex11.5.dat;
VARIABLE: NAMES = x1 x2 y1-y4 v1-v50 z1-z5;
USEVARIABLES = x1 x2 y1-y4 z1-z5;
AUXILIARY = v1-v10;
MISSING = ALL (999);
DATA IMPUTATION:
IMPUTE = y1-y4 x1 (c) x2;
NDATASETS = 20;
SAVE = missimp*.dat;
ANALYSIS: TYPE = BASIC;
OUTPUT: TECH8;
```

- Variables actually to be used in imputation

- Not involved, but saved in data

- Variables w/ missing values being imputed

- Number of datasets to be generated – USE AT LEAST 20!

- Name of file to be created

MI in Mplus

- MI followed by latent growth model (not saving data)

```
DATA: FILE = ex11.5.dat;
VARIABLE: NAMES = x1 x2 y1-y4 v1-v50 z1-z5;
USEVARIABLES = x1 x2 y1-y4 z1-z5;
MISSING = ALL (999);
DATA IMPUTATION:
IMPUTE = y1-y4 x1 (c) x2;
NDATASETS = 20;
ANALYSIS: ESTIMATOR = ML;
MODEL: i s | y1@0 y2@1 y3@2 y4@3;
       i s ON x1 x2;
OUTPUT: TECH1 TECH8;OUTPUT: TECH8;
```

- Variables actually to be used in imputation

- Variables w/ missing values being imputed

- Number of datasets to be generated

- Using ML estimation for the LGM

- Analysis being conducted on imputed data

- “AUXILIARY” and “SAVE” commands not necessary because not saving data

blimp

- Free software by Craig Enders
- Can import from any stats package
- Can export to any stats package
- Can handle multilevel data (e.g., daily or EMA or couples, etc.)

MI in blimp: Diagnostic phase

- DATA: ~/desktop/examples/smoking.dat;
- VARNAMES: id txgroup txdum1 txdum2 male age years
  cigs heavycig efficacy stress;
- MISSING: -99;
- MODEL: ~ years cigs efficacy
- SEED: 90291;
- BURN: 3000;
- THIN: 1;
- NIMPS: 2;
- OUTFILE: ~/desktop/examples/imp*.csv;
- OPTIONS: separate psr;
- CHAINS 2 processors 2;

We're letting it go for a long time (burn) because this is diagnostic, and it generates diagnostics every 50 iterations by default. That's also why we're using a low thin because we're not actually saving any of the data, so it doesn't matter. The imputed values in the generated data don't mater because we're not saving/using it. This is just for diagnostics.

The "psr" tells it to give us the diagnostics.

The "separate" saves each imputed dataset as a separate file (for Mplus, and HLM).

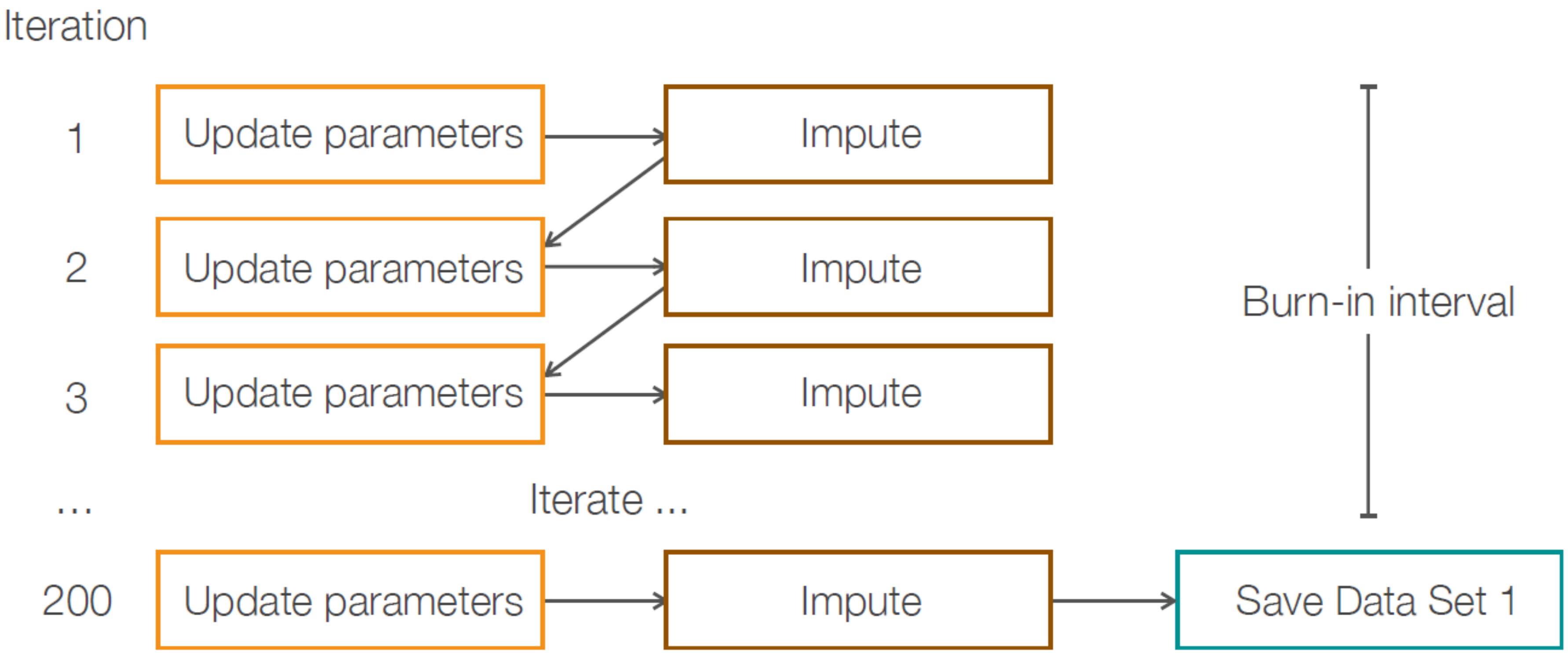
Most other software (R, SAS, etc.) want it "stacked" where each dataset is stacked on top of the other. The default is stacked unless you say otherwise.

Burn In and Thinning

- Notice you need to specify “burn in” and “thinning” in blimp (and can probably change defaults for these in other programs)
- Imputed values are based on parameter estimates (what would Y be if X is 3?
- Parameter estimates are themselves influenced by the imputed values, so change is slow
    - Imputed Y is based on B = 1.000, so then B is adjusted to 1.0001 based on new Y, but then new imputed Y is based on B = 1.0001, so then B is adjusted to 1.0002 based on this new imputed Y, etc.

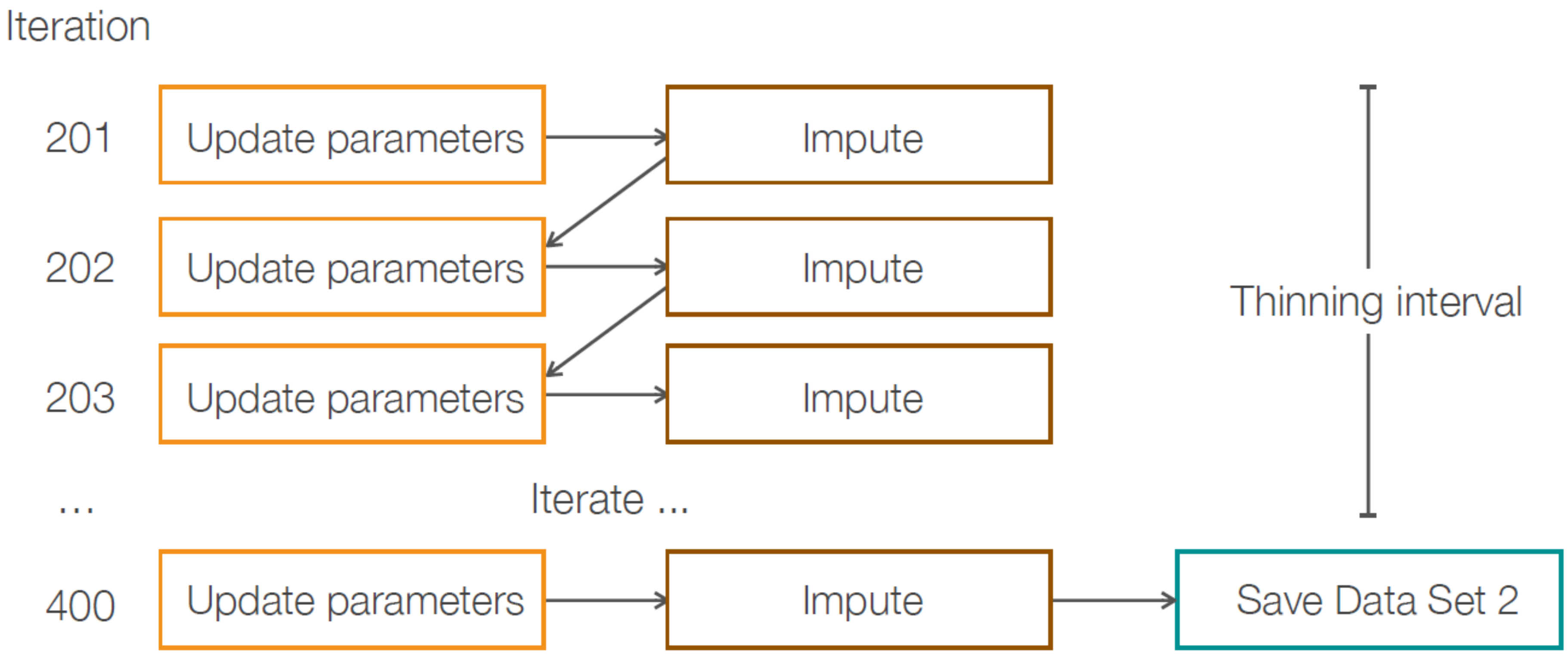
Burn In

- Period before first dataset is saved, as parameters get better so imputations get better



Thinning

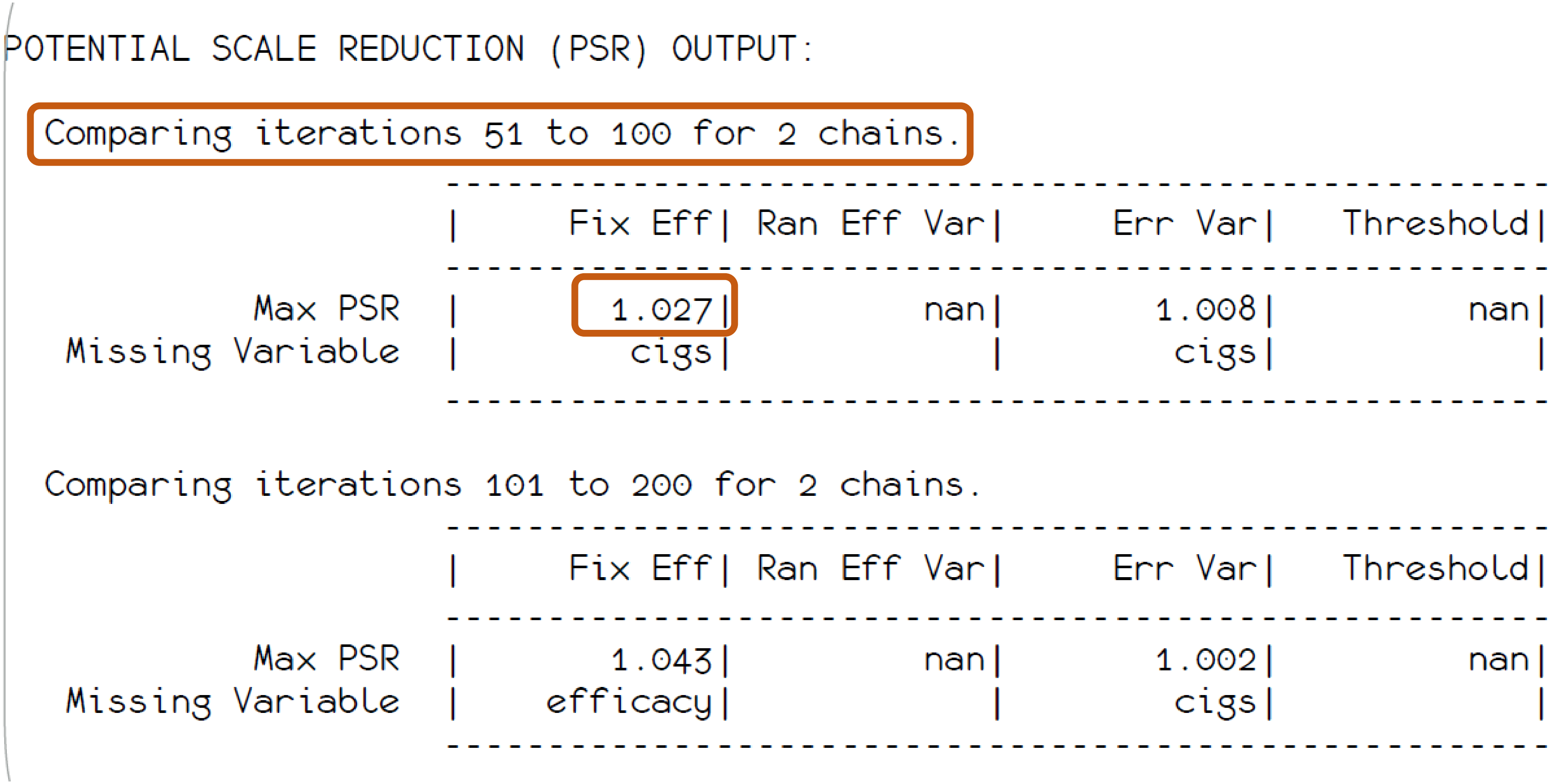
- Interval between saved datasets (to allow for greater adjustments to parameters and imputations)



Burn in and Thinning in blimp

- Potential scale reduction (PSR) captures the degree of similarity between imputations generated from two separate MCMC runs
- PSR < 1.05 to 1.10 is often considered acceptable
- Use the PSR to specify burn-in and thinning
- Run the blimp model once (don’t need many imputations) to get PSR; use PSR to specify burn-in and thinning
    - Diagnostic phase
- After specifying burn-in and thinning, run real model with at least 20 imputations

Diagnostics (PSR)



```
POTENTIAL SCALE REDUCTION (PSR) OUTPUT:

 Comparing iterations 51 to 100 for 2 chains.
                   ---------------------------------------------------
                   |     Fix Eff| Ran Eff Var|     Err Var|   Threshold|
                   ---------------------------------------------------
          Max PSR  |       1.027|         nan|       1.008|         nan|
 Missing Variable  |        cigs|            |        cigs|            |
                   ---------------------------------------------------

 Comparing iterations 101 to 200 for 2 chains.
                   ---------------------------------------------------
                   |     Fix Eff| Ran Eff Var|     Err Var|   Threshold|
                   ---------------------------------------------------
          Max PSR  |       1.043|         nan|       1.002|         nan|
 Missing Variable  |    efficacy|            |        cigs|            |
                   ---------------------------------------------------
```

MI in blimp: Actual imputation (mplus, HLM)

- DATA: ~/desktop/examples/smoking.dat;
- VARNAMES: id txgroup txdum1 txdum2 male age years
  cigs heavycig efficacy stress;
- MISSING: -99;
- MODEL: ~ years cigs efficacy
- SEED: 90291;
- BURN: 100;
- THIN: 100;
- NIMPS: 20;
- OUTFILE: ~/desktop/examples/imp*.csv;
- OPTIONS: separate;
- CHAINS 2 processors 2;

Specifying burn to be 100; no reason why thin should be different from burn.
NIMPS is 20 because saving 20 datasets (real run this time)

Imp* means they will be named imp1, imp2, imp3, … imp20.csv

The "separate" saves each imputed dataset as a separate file for Mplus, and HLM (as opposed to “stacked”). Wouldn’t need asterisk for file name if stacked because only one file.

“psr” is gone.

MI in blimp: Actual imputation (R, SAS, SPSS)

- DATA: ~/desktop/examples/smoking.dat;
- VARNAMES: id txgroup txdum1 txdum2 male age years
  cigs heavycig efficacy stress;
- MISSING: -99;
- MODEL: ~ years cigs efficacy
- SEED: 90291;
- BURN: 100;
- THIN: 100;
- NIMPS: 20;
- OUTFILE: ~/desktop/examples/imp.csv;
- OPTIONS: stacked;
- CHAINS 2 processors 2;

Now just “imp” (not “imp*”) because only one file.

“stacked” instead of “separate”
Stacked is default so could have left off OPTIONS.

Using blimp data in Mplus

- DATA:
- file = implist.csv;
- type = imputation;
- VARIABLE:
- Names = id txgroup txdum1 txdum2 male age years cigs heavycig efficacy stress;
- Usevariables = years cigs efficacy;
- MODEL:
- Efficacy on years (b1) cigs (b2);
- MODEL TEST:
- b1 = 0; b2 = 0;
- OUTPUT standardized (stdyx);

Only difference from any other mplus model is specifying the correct file and indicating it is an imputation file.

Nothing to do with imputation, but I wanted to point out that one can test things besides if a parameter is different from 0. If, for example, you want to test if it is different from another established parameter from the literature, you could simply list:
MODEL TEST:
b1 = 1.1; b2 = 0.8

MI models

- Regardless of software chosen, need to tell it what variables to consider when generating imputed values
- If your final hypothetical model is complex, some of these complexities need to be taken into account when imputing
- If you think the effectiveness of your intervention varies by sex, or that developmental trajectories vary by age at baseline, then these interactions need to be part of *imputation* model
  - Condition*sex
  - Time*age
  - I believe this can be done in SAS, Mplus, and blimp, but not SPSS
- “Researchers often have a pretty good intuition about why data are missing. If the poorest kids who move in and out of districts are the ones most likely to miss follow-ups, can use free-and-reduced-lunch variable to condition for missingness.” – Enders (no citation)

MI in blimp: Imputing Interactions

- DATA: ~/desktop/examples/pain.dat;
- VARNAMES: female diagnose sleep pain posaff negaff stress;
- MISSING: -99;
- MODEL: ~ female diagnose sleep pain negaff stress pain*female;
- SEED: 90291;
- OUTCOME: stress:
- BURN: 500;
- THIN: 500;
- NIMPS: 20;
- OUTFILE: ~/desktop/examples/imp*.csv;
- OPTIONS: separate;
- CHAINS 2 processors 2;

Included product/interaction (and outcome) in model.

Would do this for both diagnostic phase and real model generating the data.

Imputed values

- Imputed values may not match other values in the dataset
  - E.g., Response values are 1-5 in whole numbers, and imputed value is 2.364
- This is fine and actually preferred in most situations
  - Continuous, ordinal, dichotomous scales
- This may seem intuitive for continuous scales, but give pause for ordinal or dichotomous scales
- Wu, Jia, and Enders (2015) compared a variety of imputation strategies for dichotomous and ordinal data
  - Examined reliability coefficients, mean scale scores, and regressions coefficients
  - Compared normal data models with or without rounding, latent variable models, and categorical data models
  - The only strategies that did not perform well (i.e., introduced substantial bias) were methods that forced whole numbers (i.e., normal data models WITH rounding, or logistic/categorical models)

Imputing categorical values (blimp)

```
DATA: ~/desktop/examples/probsolve.dat;
VARIABLES: school condition esolpercent student abilitygrp female
   stanmath frlunch efficacy probsolve1 probsolve7;
ORDINAL: efficacy;
NOMINAL: abilitygrp female frlunch;
MISSING: -99;
MODEL: ~ abilitygrp female stanmath frlunch efficacy probsolve1 probsolve7;
NIMPS: 20;
BURN: 1000;
THIN: 1000;
SEED: 90291;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate;
CHAINS: 2 processors 2;
```

Using imputed categorical values (mplus)

```
DATA: file = implist.csv;
Type = imputation;
VARIABLE: names = school condition esolpercent student abilitygrp female
  stanmath frlunch efficacy probsolve1 probsolve7;
Usevars = female efficacy probsolve1 probsolve7 abilgrp2 abilgrp3;
DEFINE:
abilgrp2 = 0;
abilgrp3 = 0;
If (abilgrp eq 2) then abilgrp2 = 1;
If (abilgrp eq 3) then abilgrp3 = 1;
MODEL:
probsolve7 on probsolve1 efficacy female abilgrp2 abilgrp3;
OUTPUT: standardized;
```

Imputing **multilevel** data (blimp)

```
DATA: ~/desktop/examples/probsolve.dat;
VARIABLES: school condition esolpercent student abilitygrp female
   stanmath frlunch efficacy probsolve1 probsolve7;
ORDINAL: condition female frlunch efficacy;
OUTCOME: probsolve7;
MISSING: -99;
MODEL: school ~ condition esolpercent female stanmath
   frlunch probsolve1 efficacy:probsolve7;
NIMPS: 20;
BURN: 1000;
THIN: 1000;
SEED: 90291;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate;
CHAINS: 2 processors 2;
```

Tilda! The ID variable for clusters comes before the tilda (in this case, school goes from 1 to 99 for the 99 clusters).

It can do 3 levels. Would do for both diagnostic and actual phases.

"efficacy:probsolve7" says there is a random efficacy slope predicting probsolve7. Same logic as interaction; allows the random slope to exist.

Maximum Likelihood (ML) Estimation

- Identifies estimates that maximize the probability of observing what has been observed
  - If beta = 0.3213, what is the probability/likelihood of the current data?
  - If beta = 0.3212, what is the probability/likelihood of the current data?
  - Yields parameter estimates with the maximum (highest) likelihood (probability) given your data
- When data are missing
  - The likelihood is computed separately for those cases with complete data on ***some*** variables and those with complete data on ***all*** variables
  - These two likelihoods are then maximized together to find the estimates
- Uses all data available (complete and incomplete)
- Want to make sure you include (or control for) any predictors that influenced missingness for your outcomes when you specify your model
- Does not require imputation
- Much less complex than MI!

ML in Software

- SPSS
  - Default estimation is Ordinary Least Square (OLS)
  - Can switch to ML if you have the AMOS add-on
    - Must be using SEM approach or path analysis
- ML is default in Mplus for most analyses
  - Can add “ANALYSIS: ESTIMATOR= ML” to be sure
  - Will use all available data for *endogenous* variables
    - Variables being predicted by anything
  - Will still exclude cases with missing data on *exogenous* predictors
    - Variables not predicted by anything (e.g., gender, intervention status)
    - Easy to make a variable endogenous (estimate its variance or mean)
  - Could combine ML estimation with imputed predictors
- ML is easily available in SAS

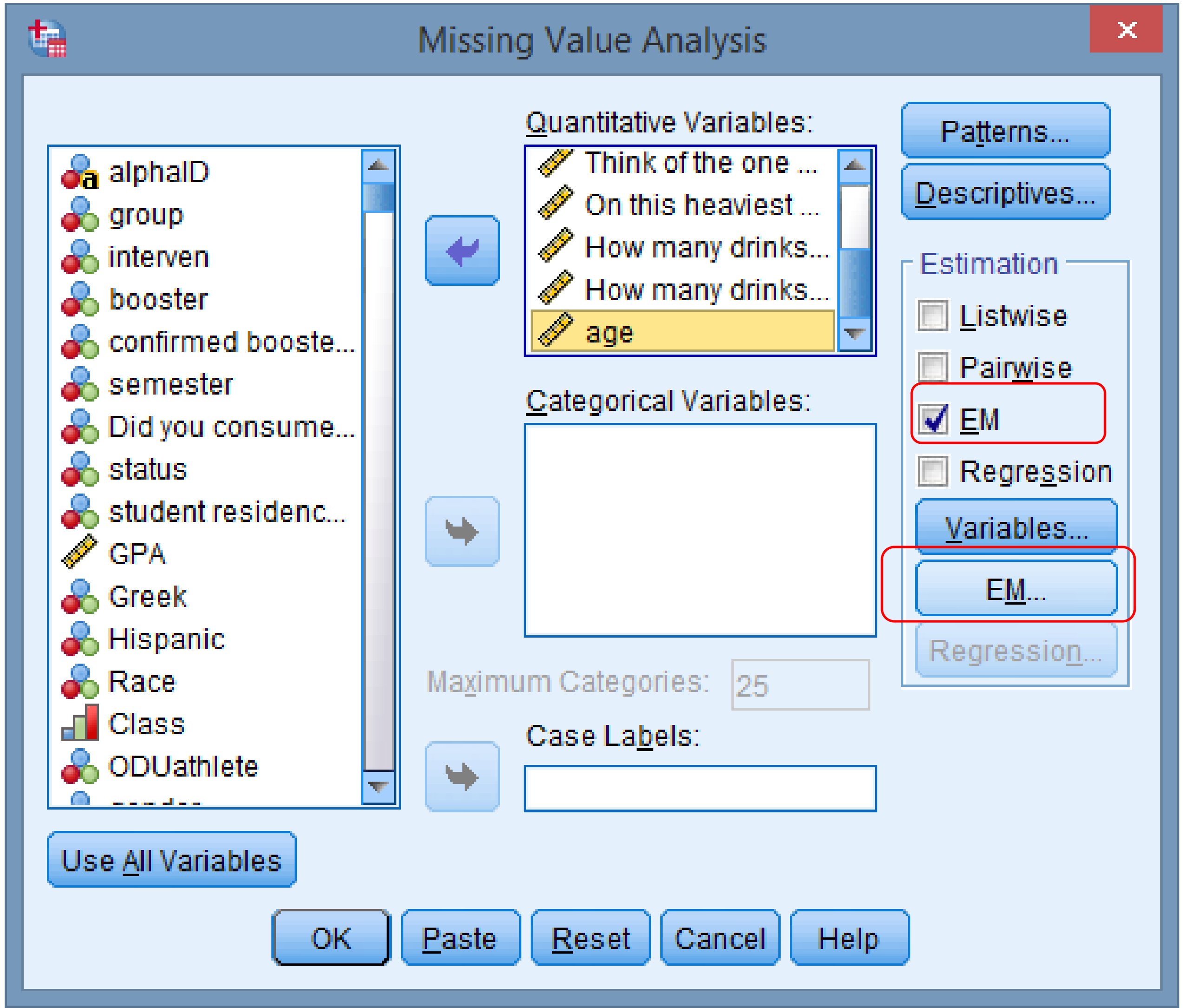
Expectation Maximization (EM) Imputation

- Consists of 2 steps
- *1) <u>Expectation</u>*: Choose values for unknown data
  - Generates means and covariance matrix based on complete data
  - Start with expected value based on regression imputation
  - What would X4 be, given what we know about X1, X2, and X3?
- *2) <u>Maximization</u>*: Calculate new means and covariance matrix using imputed data
  - Use the new means and covariance matrix to go back to step 1 and re-estimate missing values
- Iterative
  - Repeats until convergence (i.e., until the numbers stop changing or the changes are barely perceptible)
- Generates only ONE dataset with imputed values
  - Parameter estimates themselves are unbiased
  - Can still generate deflated SEs, so avoid as only solution, but layer on with ML

EM Imputation in SPSS

- Can be done easily in EQS, SAS, and SPSS
- In SPSS:
- Analyze > Missing Value Analysis
- Transfer all relevant numerical variables into “Quantitative Variables”
- Transfer all relevant categorical variables into “Categorical Variables”
- Select the EM option
- Press the EM button

EM Imputation in SPSS



Missing Value Analysis

Quantitative Variables:
- Think of the one ...
- On this heaviest ...
- How many drinks...
- How many drinks...
- age

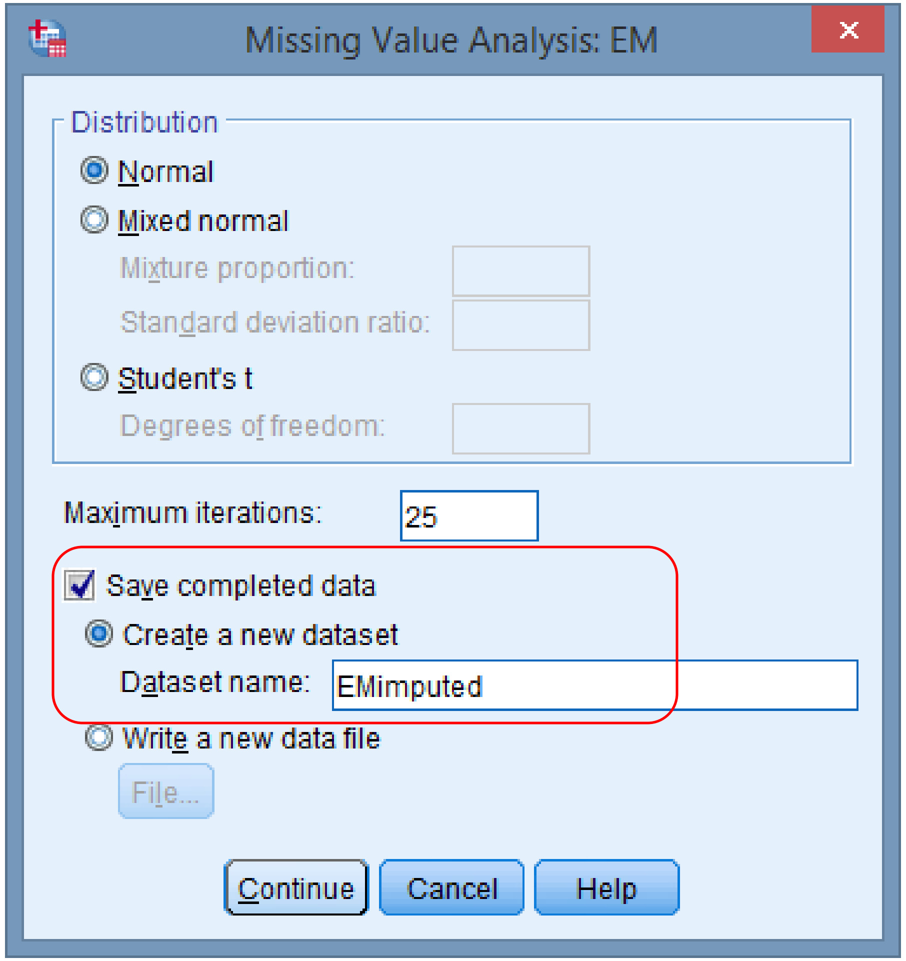
Categorical Variables:

Maximum Categories: 25

Case Labels:

Estimation
- □ Listwise
- □ Pairwise
- ☑ EM
- □ Regression

EM Imputation in SPSS



Missing Value Analysis: EM

Distribution

● Normal

○ Mixed normal

Mixture proportion:

Standard deviation ratio:

○ Student's t

Degrees of freedom:

Maximum iterations: 25

☑ Save completed data

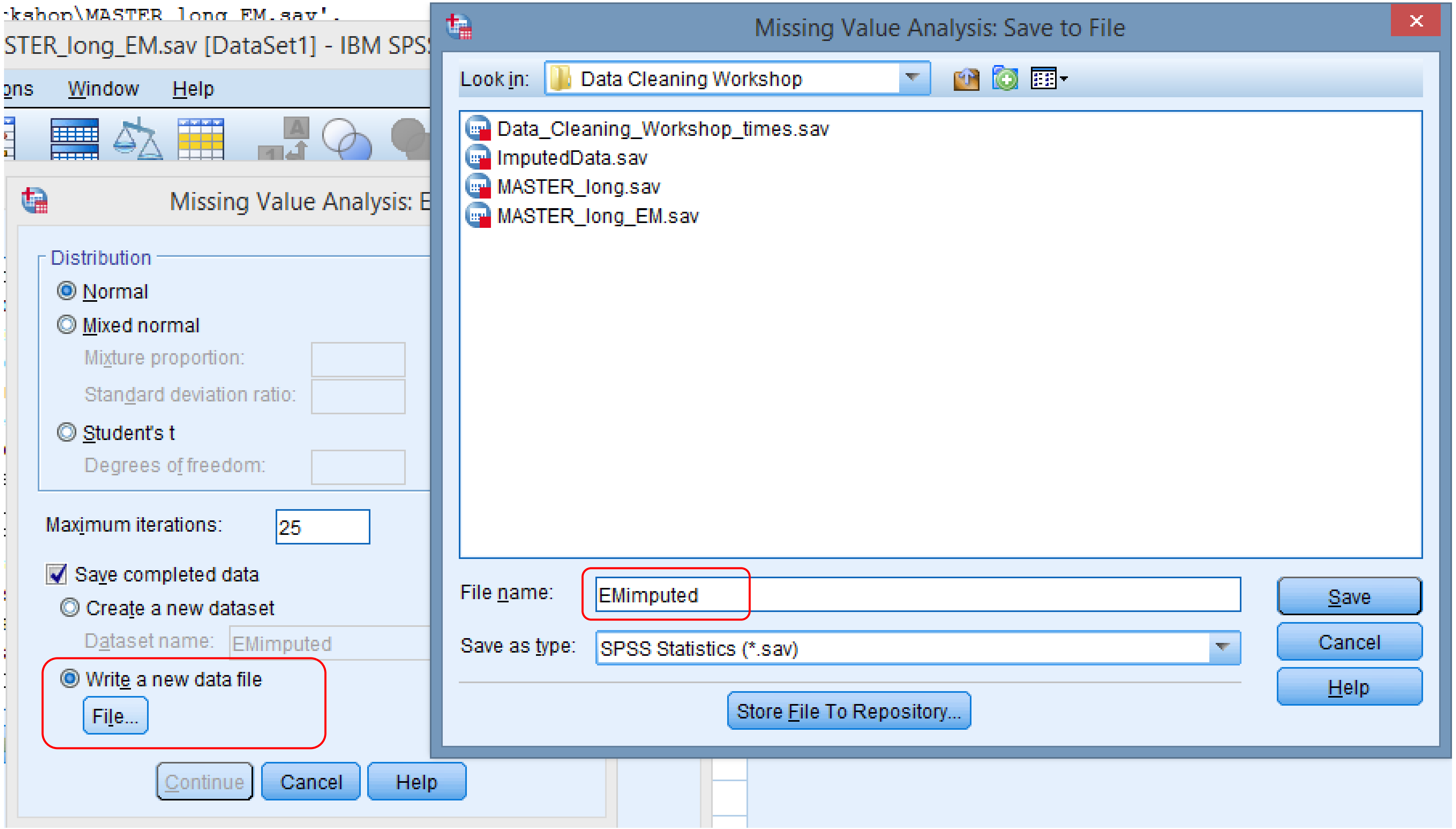
● Create a new dataset

Dataset name: EMimputed

○ Write a new data file

- Select “Save completed data”
- Choose “Create a new dataset” and name it ***OR*** “Write a new data file”
  - Press File and type a filename
- Open this new file
- Should include the observed data together with imputed data
- Conduct analyses on this file

EM Imputation in SPSS



kshop\MASTER_long_EM.sav'.

STER_long_EM.sav [DataSet1] - IBM SPS

Missing Value Analysis: E

Distribution

● Normal

○ Mixed normal

Mixture proportion:

Standard deviation ratio:

○ Student's t

Degrees of freedom:

Maximum iterations: 25

☑ Save completed data

○ Create a new dataset

Dataset name: EMimputed

● Write a new data file

Missing Value Analysis: Save to File

Look in: Data Cleaning Workshop

- Data_Cleaning_Workshop_times.sav
- ImputedData.sav
- MASTER_long.sav
- MASTER_long_EM.sav

File name: EMimputed

Save as type: SPSS Statistics (*.sav)

EM Imputation in SPSS

- The saved dataset with imputed values will only contain variables involved in the imputation
- If you excluded some variables, you'll want to merge the files
  - Use merge feature in SPSS
  - OR copy/paste variables
    - Be sure to sort both datasets by ID first so the cases match up

Choosing Imputation Type

- EM seems to be falling out of favor
- Multiple imputation and maximum likelihood estimation are the two favored approaches
  - Lots of papers comparing the two and ideal circumstances for each
- Multiple imputation is always a good choice
  - Often yields the least biased results
  - No one will ever question your choice (nothing is considered better)
  - Relatively easy to implement with simple analyses
    - Although the holes in the pooled output can be a real pain
    - Excel spreadsheets and hand calculations
  - Can be difficult to combine with complex analyses
    - CAN combine MI (20+ datasets) with bootstapping for mediation (5000+ resamples), but very time consuming
- Maximum likelihood estimation is also a very good choice
  - Often yields similar results to MI
  - There are some circumstances where MI could yield less biased results than ML
  - It would be unusual to be questioned in this choice (by reviewers, committee members, etc.)
  - Both are much, much better than alternative approaches

Choosing Imputation Type

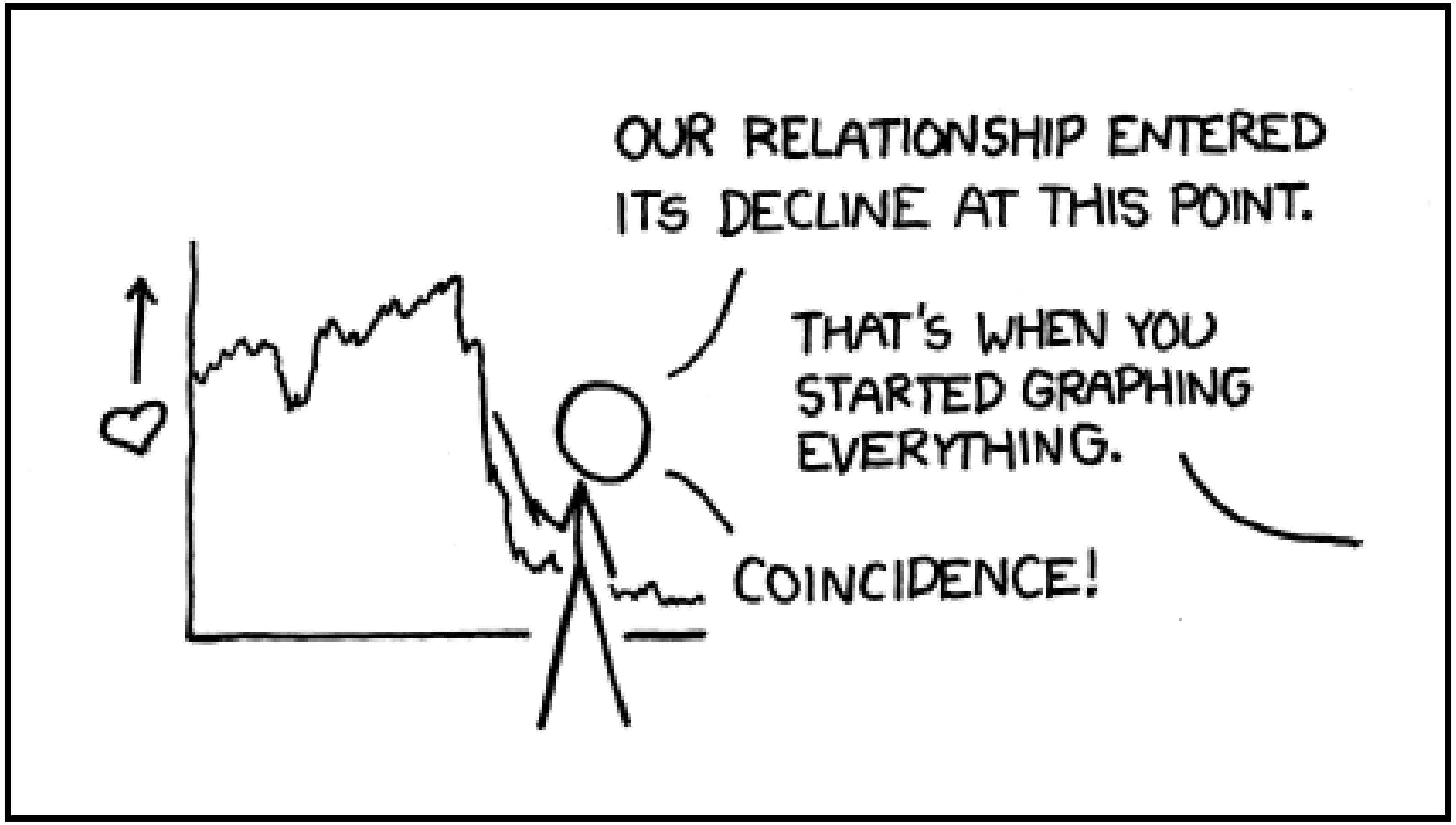
- Think about analysis type
- Think about software familiarity
- If easily conducted in SPSS and most familiar, might choose multiple imputation
- If SEM or HLM is required and/or comfortable with Mplus:
  - If analysis is straightforward, might choose MI
  - If analysis is complex, might choose ML estimation
    - If predictors have a lot of missing data, might layer on EM imputation for predictors
  - I personally tend to use ML estimation
- Either way, still examine associations between missingness and other variables
  - So they can inform the MI process, or be included in your ML model

When to impute?

- Eekhout et al. (2015) says at the item level prior to composites
  - Best practices for simpler models
    - i.e., a handful of variables
    - OR only one or two data analysis sessions
- Onerous if conducting analyses in multiple sessions
  - Would need to create composites using syntax in EVERY MODEL (Mplus, SAS, etc.) or EVERY SESSION (spss syntax).
  - SPSS does not recognize multiply imputed datasets after closing, so would need to re-impute each time
  - Blimp allows for export into Mplus, SAS, so could save code at start of each model. But means including all items rather than streamlining to composites only
  - Variables are repeated in longitudinal data
    - 5 constructs, assessed 3 times, 10 items each = 150 variables
    - 20 constructs, assessed 7 times, 10 items each = 1400 variables
- Eekhout et al. (2015) suggests including item scores as auxiliary variables
  - Should not affect model or focus on constructs, but improves missing data handling
  - Improves power

Words of Wisdom

- "…Although some missing data methods are clearly better than others, none of them can really be described as good. The only really good solution to the missing data problem is not to have any" (Allison, 2001, *p*. 2)
- Put lots of thought and effort into recruitment and retention
  - Longitudinal
    - Strong incentives for follow-ups, Bonuses for complete data
    - Lots of reminders for longitudinal data
  - Take advantage of advanced survey features
    - Point out when people skip questions
  - Include multiple items/scales to tap into the same construct
    - Allows you to model missingness if necessary (MAR instead of MNAR)



OUR RELATIONSHIP ENTERED ITS DECLINE AT THIS POINT.
THAT'S WHEN YOU STARTED GRAPHING EVERYTHING.
COINCIDENCE!

Outline for Today

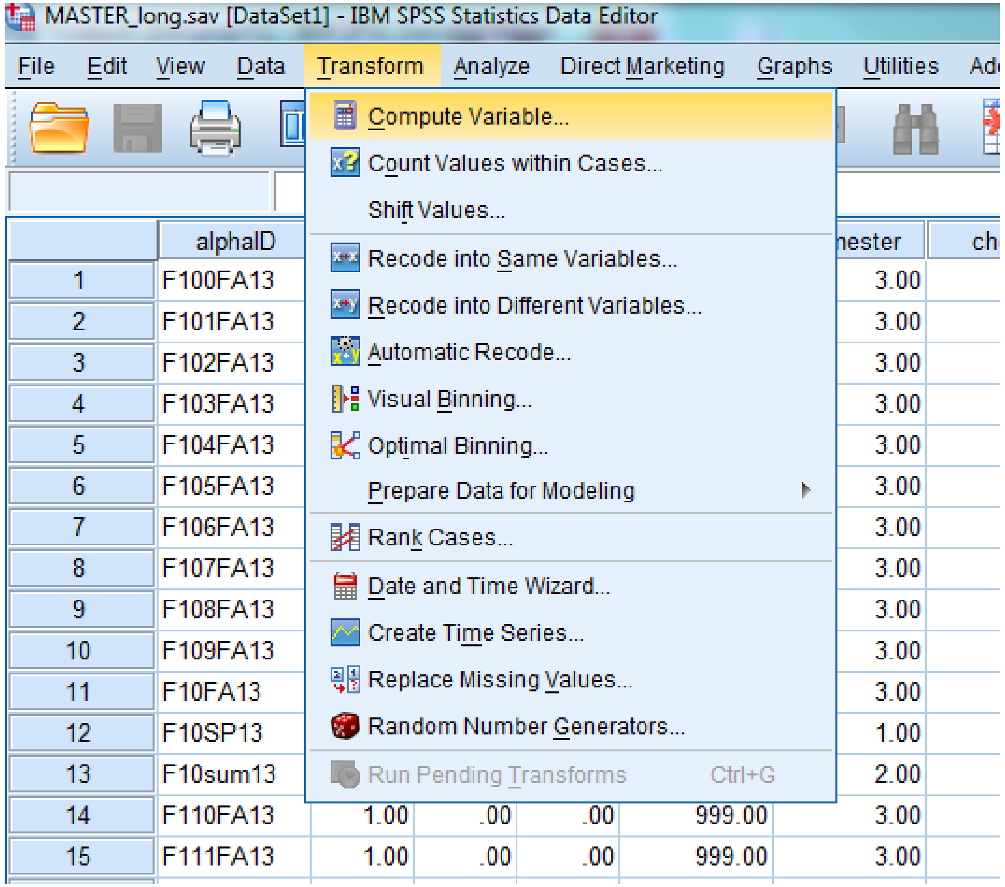
- Missing Data
    - Identifying, assessing type, imputation options
- **Composite Scores**
    - **Total scores, recoding, dummy coding**
- Outliers
    - Identifying and addressing univariate and multivariate outliers
- Normality
    - Assessing and addressing (e.g., transformations, analysis specifications)
- Bivariate Linearity
    - Reading scatterplots and what to do about them
- Documentation
    - The importance of codebooks and data logs

Composite Scores

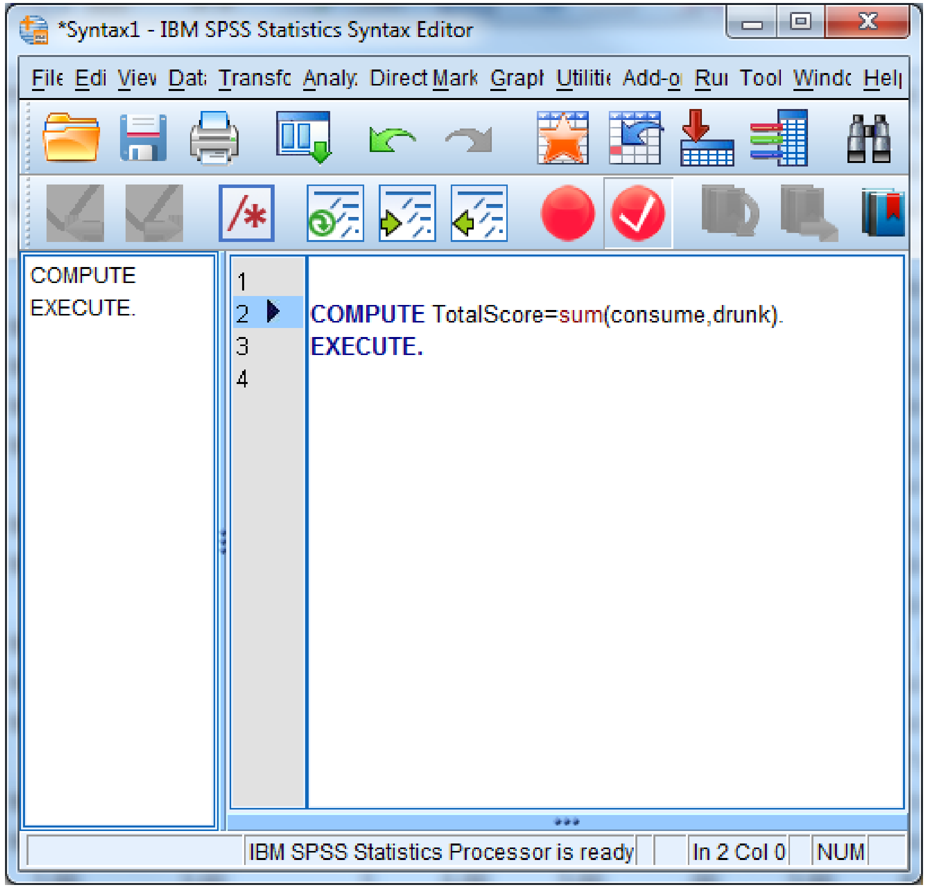
- Total Scores
  - E.g., Means, Sums
- Reverse scoring
  - Often select items within a scale
- Dummy Coding
  - Turning nominal variables into a series of dichotomous (0/1) variables
  - OR into one dichotomous variable
  - Race, gender, etc.
  - Only necessary for linear models (e.g., regression, SEM, HLM)
    - Not for ANOVA, chi-square

Total Scores

- Transform > Compute
- Syntax window


MASTER_long.sav [DataSet1] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Direct Marketing
Graphs
Utilities
Ad
Compute Variable...
Count Values within Cases...
Shift Values...
Recode into Same Variables...
Recode into Different Variables...
Automatic Recode...
Visual Binning...
Optimal Binning...
Prepare Data for Modeling
Rank Cases...
Date and Time Wizard...
Create Time Series...
Replace Missing Values...
Random Number Generators...
Run Pending Transforms Ctrl+G

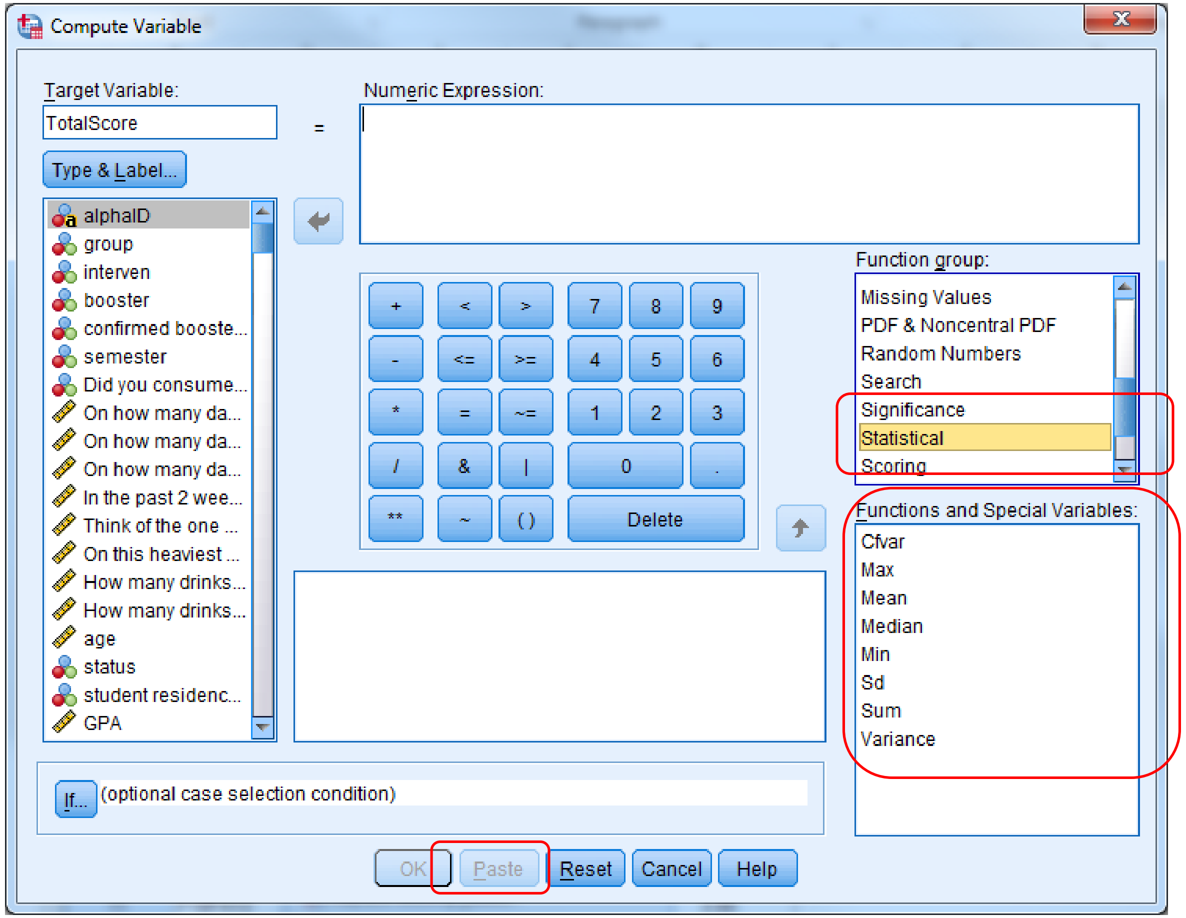
	alphaID					nester	ch
1	F100FA13					3.00	
2	F101FA13					3.00	
3	F102FA13					3.00	
4	F103FA13					3.00	
5	F104FA13					3.00	
6	F105FA13					3.00	
7	F106FA13					3.00	
8	F107FA13					3.00	
9	F108FA13					3.00	
10	F109FA13					3.00	
11	F10FA13					3.00	
12	F10SP13					1.00	
13	F10sum13					2.00	
14	F110FA13	1.00	.00	.00	999.00	3.00	
15	F111FA13	1.00	.00	.00	999.00	3.00	



```
COMPUTE TotalScore=sum(consume,drunk).
EXECUTE.
```

Total Score

- Typically sum or mean, but could be max, median, variance, etc.



Compute Variable

Target Variable: TotalScore

Numeric Expression:

Function group: Statistical

Functions and Special Variables:

(optional case selection condition)

Total Scores: Means

- *depress* = (*x*1 + *x*2 + *x*3 + *x*4 + *x*5) / 5
  - Gives average/mean when all 5 values are present
  - Gives [missing] if any values are missing
- *depress* = **MEAN**(*x*1,*x*2,*x*3,*x*4,*x*5)
  - Gives the total of all values present

- *EXAMPLE*
- *depress =* (3 + 2 + 3 + 5 + [.]) / 5 = [.]
- *depress* = MEAN(3,2,3,5,[.]) = 3.25

Total Scores: Sums

- *depress* = *x*1 + *x*2 + *x*3 + *x*4 + *x*5
  - Gives total/sum when all 5 values are present
  - Gives [missing] if any of the values are missing
- *depress* = **SUM**(*x*1,*x*2,*x*3,*x*4,*x*5)
  - Gives the total of all values present
- *depress* = **MEAN**(*x*1,*x*2,*x*3,*x*4,*x*5)***5**
  - Gives the mean of the present values, then multiplies by number of items
  - As if that value were present, and the mean of the other values

- When do you use SUM vs MEAN*[items]?
- I tend to use SUM when the missing item tends to be a zero (number of symptoms, behaviors, etc.)
  - # of drinks, alcohol-related problems, etc.
- I tend to use MEAN*[items] when missing item is likely not zero (level of agreement)
  - CESD, drinking motives, etc.

- *EXAMPLE*
- *depress* = 3 + 2 + 3 + 5 + [.] = [.]
- *depress* = SUM(3,2,3,5,[.]) = 13
- *depress* = MEAN(3,2,3,5,[.]) = 3.25*5 = 16.25

Total Scores

- If you’d rather exclude cases that have too much missing data, can add a digit to the command
- Digit represents the minimum number of values that must be present to execute the command
- *depress* = SUM.**3**(*x*1,*x*2,*x*3,*x*4,*x*5)
- *depress* = MEAN.**3**(*x*1,*x*2,*x*3,*x*4,*x*5)

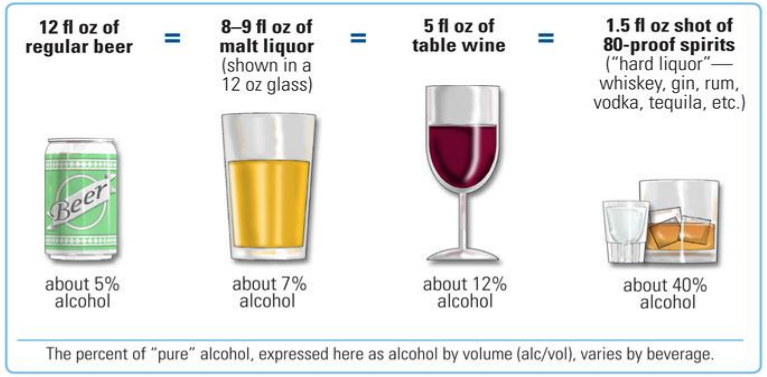
I tend to use a value that reflects at least half of the relevant items

- EXAMPLES
- *depress* = SUM.3 (3,2,3,5,[.]) = 13
- *depress* = SUM.3 (3,2,3,[.],[.]) = 8
- *depress* = SUM.3 (3,2,[.],[.],[.]) = [.]

Total Scores

Please think about your typical drinking over the **PAST 3 MONTHS**. On a typical day, how many drinks would you have, and over how many hours would you have them? That is, how many drinks would you typically have on each day in the 3 months? How long (in hours) would a typical drinking occasion last on that day? Use any applicable number, starting with 0, and please note that each space must be filled in.

NOTE: 1 drink = 1 Beer (12 oz.) = 1 Wine Cooler (12 oz.) = 1 Glass of Wine (5 oz.) = 1 Shot of Liquor (1-1.5 oz.) = 1 Mixed Drink (1-1.5 oz. of liquor)
Over the **PAST 3 MONTHS**, on a….


12 fl oz of regular beer
=
8–9 fl oz of malt liquor (shown in a 12 oz glass)
=
5 fl oz of table wine
=
1.5 fl oz shot of 80-proof spirits ("hard liquor"— whiskey, gin, rum, vodka, tequila, etc.)
Beer
about 5% alcohol
about 7% alcohol
about 12% alcohol
about 40% alcohol
The percent of "pure" alcohol, expressed here as alcohol by volume (alc/vol), varies by beverage.

	TYPICAL MONDAY	TYPICAL TUESDAY	TYPICAL WEDNESDAY	TYPICAL THURSDAY	TYPICAL FRIDAY	TYPICAL SATURDAY	TYPICAL SUNDAY
NUMBER OF DRINKS							
NUMBER OF HOURS							

Total Scores

	NO	YES
1. While drinking, I have said or done embarrassing things.		
2. The quality of my work or schoolwork has suffered because of my drinking.		
3. I have felt badly about myself because of my drinking.		
4. I have driven a car when I knew I had too much to drink to drive safely.		
5. I have had a hangover (headache, sick stomach) the morning after I had been drinking.		
6. I have passed out from drinking.		
7. I have taken foolish risks when I have been drinking.		
8. I have felt very sick to my stomach or thrown up after drinking.		
9. My drinking has created problems between myself and my boyfriend/girlfriend/spouse, parents, or other near relatives.		
10. I have spent too much time drinking.		
11. I have not gone to work or missed classes at school because of drinking, a hangover, or illness caused by drinking.		
12. I have felt like I needed a drink after I'd gotten up (that is, before breakfast).		
13. I have become very rude, obnoxious or insulting after drinking.		
14. I have woken up in an unexpected place after heavy drinking.		

Total Scores

The following is a list of reasons that some people give for drinking alcohol. Thinking of all the times you drink, how often would you say that you drink for each of the following reasons?

	Almost never/never	Some of the time	Half of the time	Most of the time	Almost always/always
1. To forget your worries	1	2	3	4	5
2. Because your friends pressure you to drink	1	2	3	4	5
3. Because it helps you enjoy a party	1	2	3	4	5
4. Because it helps you when you feel depressed or nervous	1	2	3	4	5
5. To be sociable	1	2	3	4	5
6. To cheer up when you are in a bad mood	1	2	3	4	5
7. Because you like the feeling	1	2	3	4	5
8. So that others won't kid you about not drinking	1	2	3	4	5
9. Because it's exciting	1	2	3	4	5
10. To get high	1	2	3	4	5
11. Because it makes social gatherings more fun	1	2	3	4	5
12. To fit in with a group you like	1	2	3	4	5
13. Because it gives you a pleasant feeling	1	2	3	4	5
14. Because it improves parties and celebrations	1	2	3	4	5
15. Because you feel more self-confident and sure of yourself	1	2	3	4	5

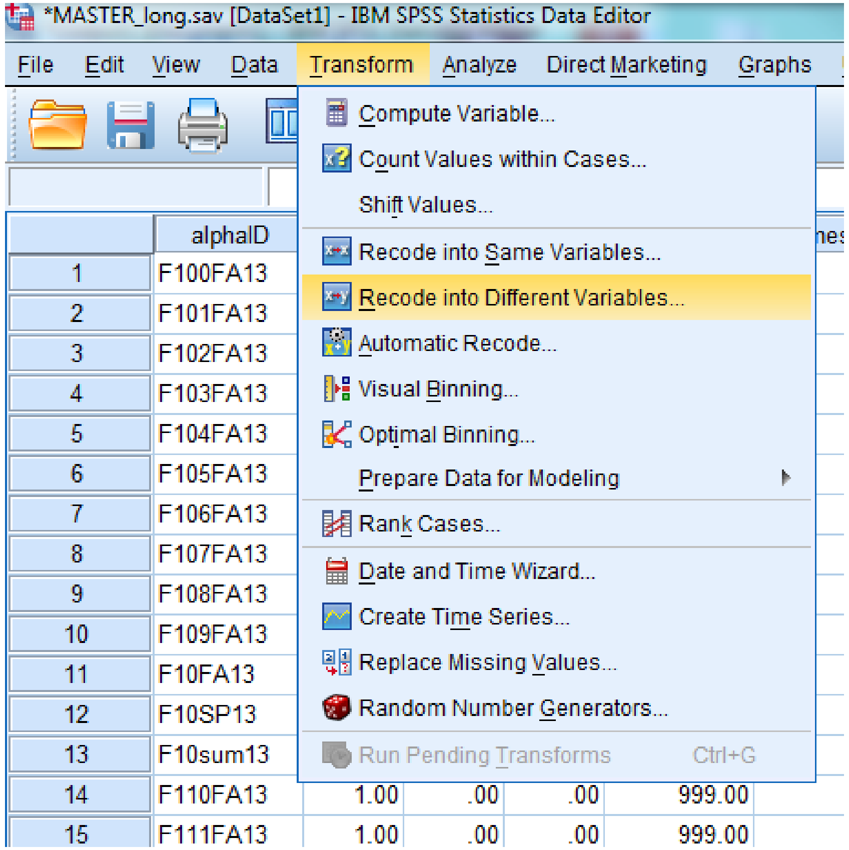
Reverse Scoring

1. I was bothered by things that usually don’t bother me.
2. I did not feel like eating; my appetite was poor.
3. I felt that I could not shake off the blues even with help from my family.
4. ***I felt that I was just as good as other people.****
5. I had trouble keeping my mind on what I was doing.
6. I felt depressed.
7. I felt that everything I did was an effort.
8. ***I felt hopeful about the future.****
9. I thought my life had been a failure.
10. I felt fearful.

Source: Radloff, L.S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*: 385-401.
Note. First 10 items only

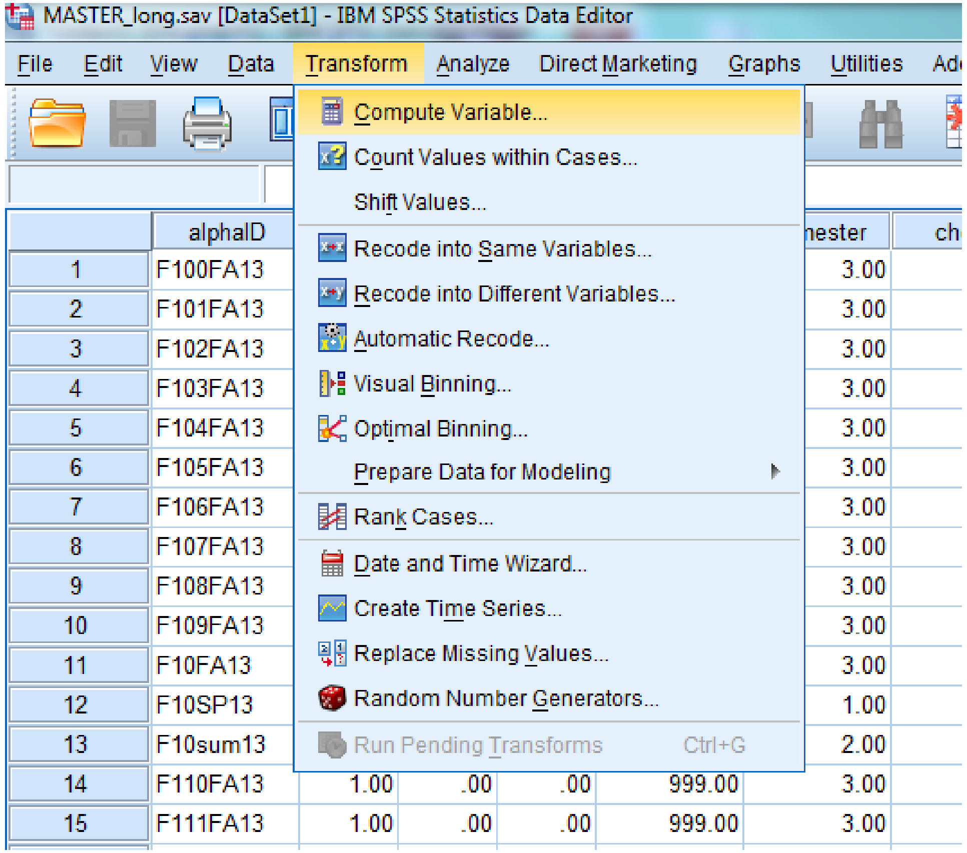
Reverse Scoring

- Recode versus Compute



*MASTER_long.sav [DataSet1] - IBM SPSS Statistics Data Editor

	alphaID				nes
1	F100FA13				
2	F101FA13				
3	F102FA13				
4	F103FA13				
5	F104FA13				
6	F105FA13				
7	F106FA13				
8	F107FA13				
9	F108FA13				
10	F109FA13				
11	F10FA13				
12	F10SP13				
13	F10sum13				
14	F110FA13	1.00	.00	.00	999.00
15	F111FA13	1.00	.00	.00	999.00

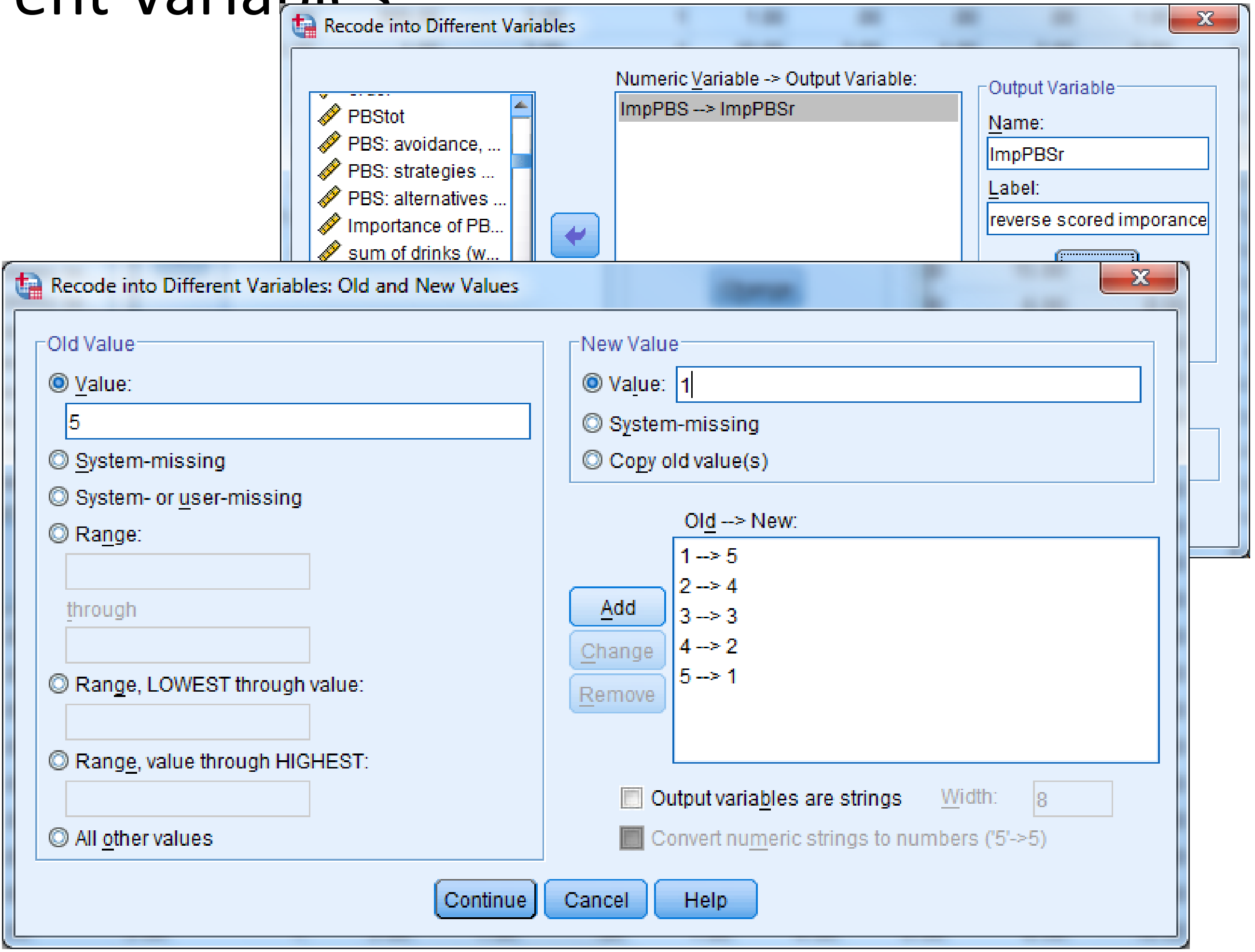


MASTER_long.sav [DataSet1] - IBM SPSS Statistics Data Editor

	alphaID					nester	ch
1	F100FA13					3.00	
2	F101FA13					3.00	
3	F102FA13					3.00	
4	F103FA13					3.00	
5	F104FA13					3.00	
6	F105FA13					3.00	
7	F106FA13					3.00	
8	F107FA13					3.00	
9	F108FA13					3.00	
10	F109FA13					3.00	
11	F10FA13					3.00	
12	F10SP13					1.00	
13	F10sum13					2.00	
14	F110FA13	1.00	.00	.00	999.00	3.00	
15	F111FA13	1.00	.00	.00	999.00	3.00	

Reverse Scoring

- RECODE into Different Variables
- 1 to 5
- 2 to 4
- 3 to 3
- 4 to 2
- 1 to 5



Recode into Different Variables

PBStot
PBS: avoidance, ...
PBS: strategies ...
PBS: alternatives ...
Importance of PB...
sum of drinks (w...

Numeric Variable -> Output Variable:
ImpPBS --> ImpPBSr

Output Variable
Name: ImpPBSr
Label: reverse scored imporance

Recode into Different Variables: Old and New Values

Old Value
● Value: 5
○ System-missing
○ System- or user-missing
○ Range:
through
○ Range, LOWEST through value:
○ Range, value through HIGHEST:
○ All other values

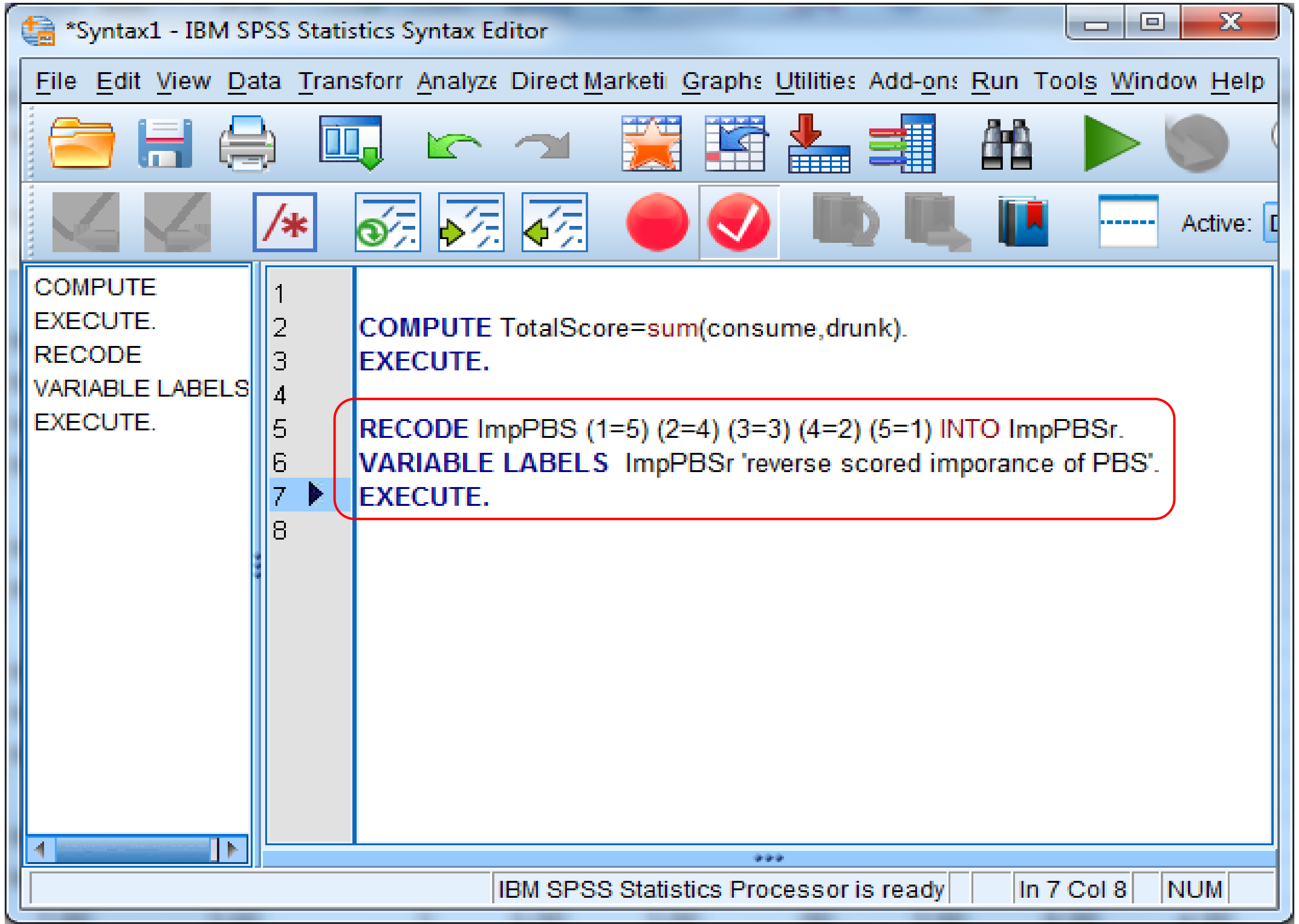
New Value
● Value: 1
○ System-missing
○ Copy old value(s)

Old --> New:
1 --> 5
2 --> 4
3 --> 3
4 --> 2
5 --> 1

□ Output variables are strings Width: 8
□ Convert numeric strings to numbers ('5'->5)

Reverse Scoring

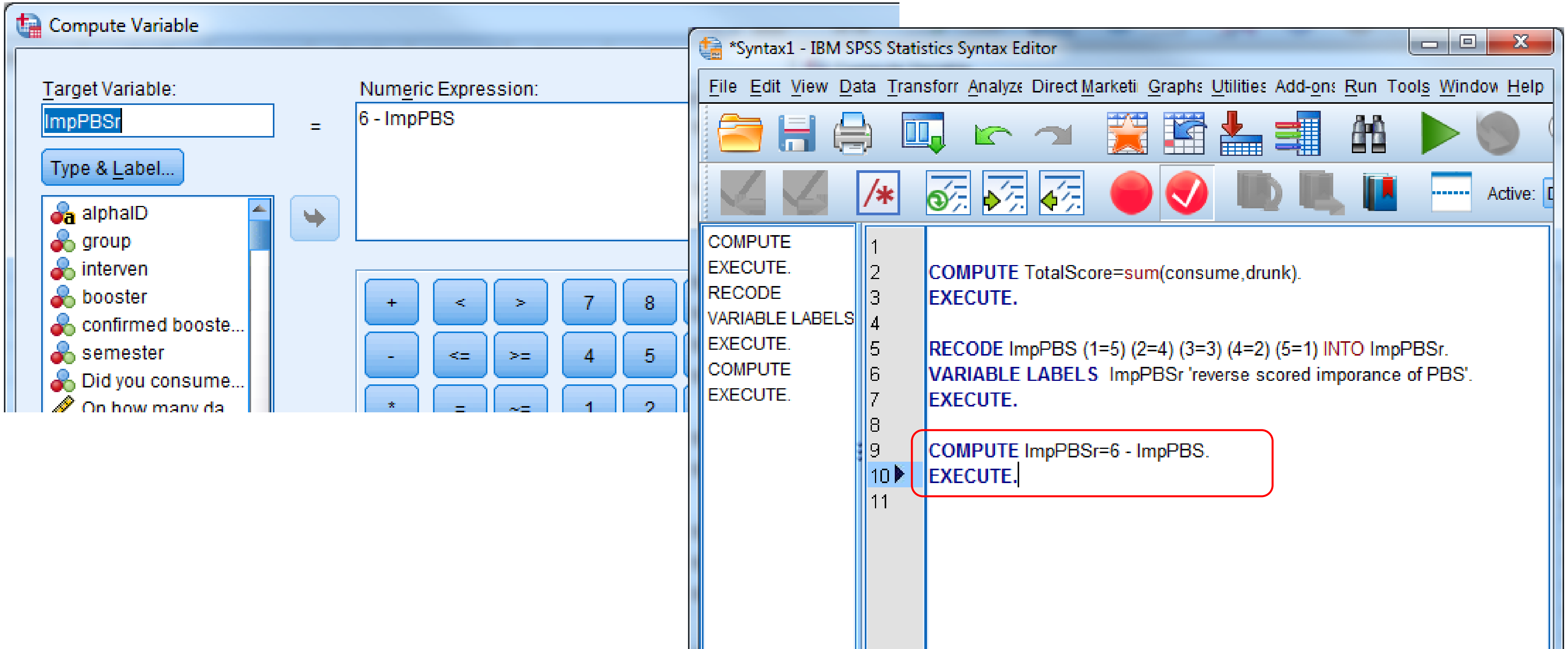
- Recode into different variables via syntax


*Syntax1 - IBM SPSS Statistics Syntax Editor
File Edit View Data Transforr Analyze Direct Marketi Graphs Utilities Add-ons Run Tools Window Help
Active:
COMPUTE
EXECUTE.
RECODE
VARIABLE LABELS
EXECUTE.
IBM SPSS Statistics Processor is ready
In 7 Col 8
NUM

```
COMPUTE TotalScore=sum(consume,drunk).
EXECUTE.

RECODE ImpPBS (1=5) (2=4) (3=3) (4=2) (5=1) INTO ImpPBSr.
VARIABLE LABELS  ImpPBSr 'reverse scored imporance of PBS'.
EXECUTE.
```

Reverse Scoring

- A better way!
- *If scaling starts at 1:* COMPUTE: (max+1) – value
- *If scaling starts at 0:* COMPUTE: max - value



Compute Variable

Target Variable: ImpPBSr = Numeric Expression: 6 - ImpPBS

*Syntax1 - IBM SPSS Statistics Syntax Editor

```
COMPUTE TotalScore=sum(consume,drunk).
EXECUTE.

RECODE ImpPBS (1=5) (2=4) (3=3) (4=2) (5=1) INTO ImpPBSr.
VARIABLE LABELS  ImpPBSr 'reverse scored imporance of PBS'.
EXECUTE.

COMPUTE ImpPBSr=6 - ImpPBS.
EXECUTE.
```

Reverse Scoring

- In many instances, yields the same results as RECODE

RECODE

(1=5) (2=4) (3=3) (4=2) (5=1)

- *1* becomes **5**
- *2* becomes **4**
- *3* becomes **3**
- *4* becomes **2**
- *5* becomes **1**

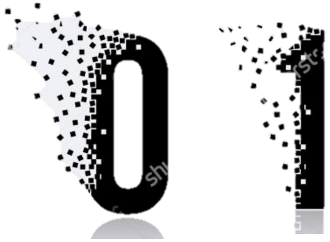
COMPUTE
= 6 - 1

- 6 – *1* = **5**
- 6 – *2* = **4**
- 6 – *3* = **3**
- 6 – *4* = **2**
- 6 – *5* = **1**

Reverse Scoring

- In some circumstances, COMPUTE option is superior
- COMPUTE version is faster with less room for human error
- Can incorporate non-whole numbers (like imputed values)
  - Imputed Value = 1.32
  - Using RECODE Into Different Variables, 1.32 becomes [.] (missing again)
  - Using COMPUTE, 1.32 becomes 6 – 1.32 = 4.68
  - Was between 1 and 2, now between 4 and 5
  - Successful recode

Dummy Codes



- Turning categorical variables into a series of dichotomous (0/1) variables
  - OR into one dichotomous variable
- Only necessary for linear models (e.g., regression, SEM, HLM)
  - Not for ANOVA, chi-square
- Set of new dummy variables if you have 3+ groups and want to keep 3+ groups
- Single new dummy variable if you have 2 groups, or want to turn 3+ groups into 2 groups
- Technically more possibilities than 0/1
  - [-1/+1], [-.5/+.5], etc.
  - Only worth exploring these if want to influence the interpretation of your intercept

Dummy Codes: Multi

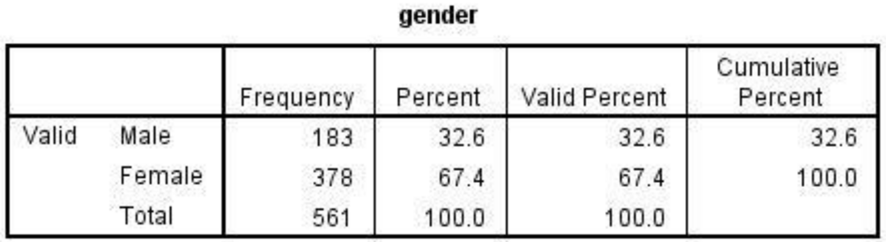
- Set of new dummy variables if you have 3+ groups and want to keep 3+ groups
- For *k* groups, will create *k* – 1 variables
- Choose reference group
  - Will have value of zero for all *k* – 1 variables
  - Will be the group represented by the intercept (or the default in the path model, etc.)
  - Often the most frequent group, or “default” such as control group
- Examples:
  - Race: 6 categories into 5 dichotomous variables
  - Treatment: 3 groups into 2 dichotomous variables
  - Marital Status: 5 categories into 4 dichotomous variables

Dummy Codes: Single

- If you have only 2 groups you want to analyze
    - Gender (male/female), treatment (control, active), Symptom (yes/no)
- Necessary if codes were [1,2], etc.
    - Common in survey software
- Choose which group is 0
    - Often most frequent group (females in SONA pool!) or default (control group, no symptom)
- Choose which group is 1
    - Other group
- Intercept represents “0”group
    - Mean for females in control group
- Variable represents how things change for “1” group
    - Change for males, change for receiving treatment, change for people who have symptom/diagnosis

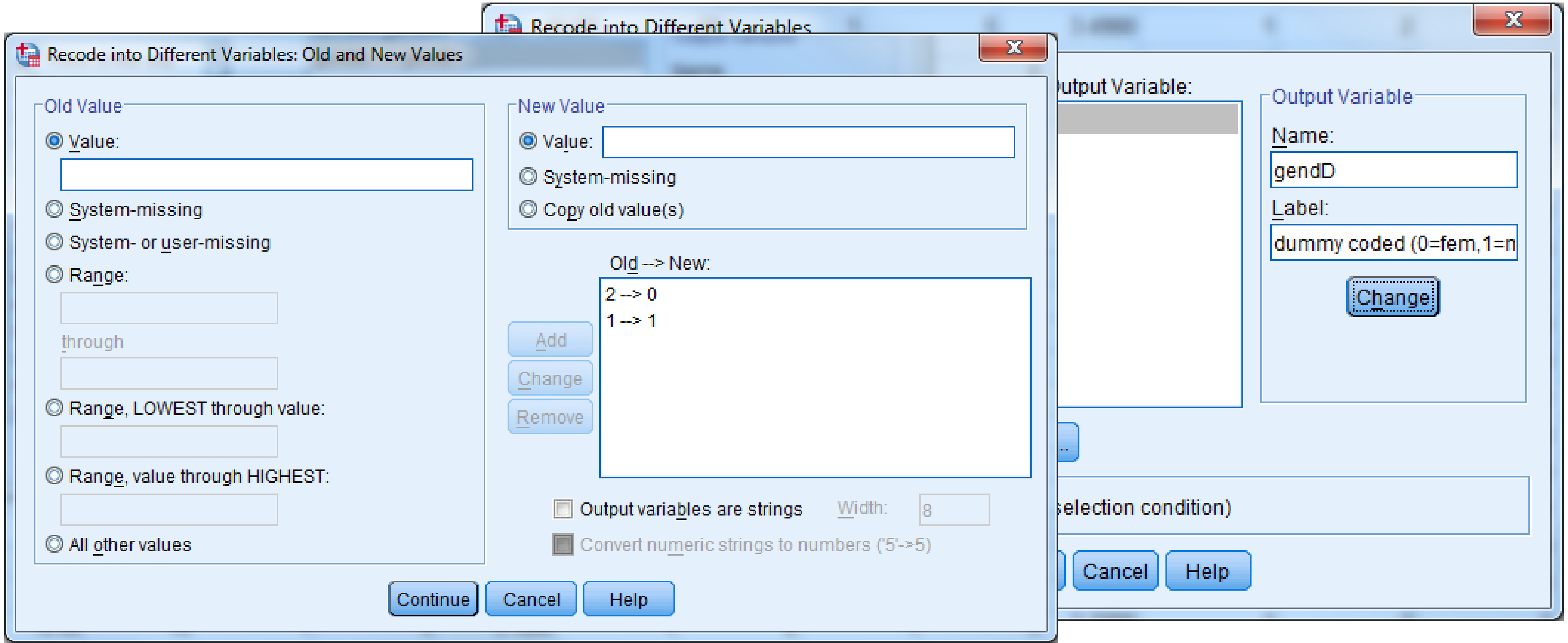
Dummy Codes: Single

- Gender example:
- Females are much more prevalent → 0
- RECODE into Different Variable



gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	183	32.6	32.6	32.6
	Female	378	67.4	67.4	100.0
	Total	561	100.0	100.0	



Recode into Different Variables: Old and New Values

Old Value

● Value:

○ System-missing

○ System- or user-missing

○ Range:

through

○ Range, LOWEST through value:

○ Range, value through HIGHEST:

○ All other values

New Value

● Value:

○ System-missing

○ Copy old value(s)

Old --> New:

2 --> 0
1 --> 1

□ Output variables are strings Width: 8

□ Convert numeric strings to numbers ('5'->5)

Recode into Different Variables

utput Variable:

Output Variable

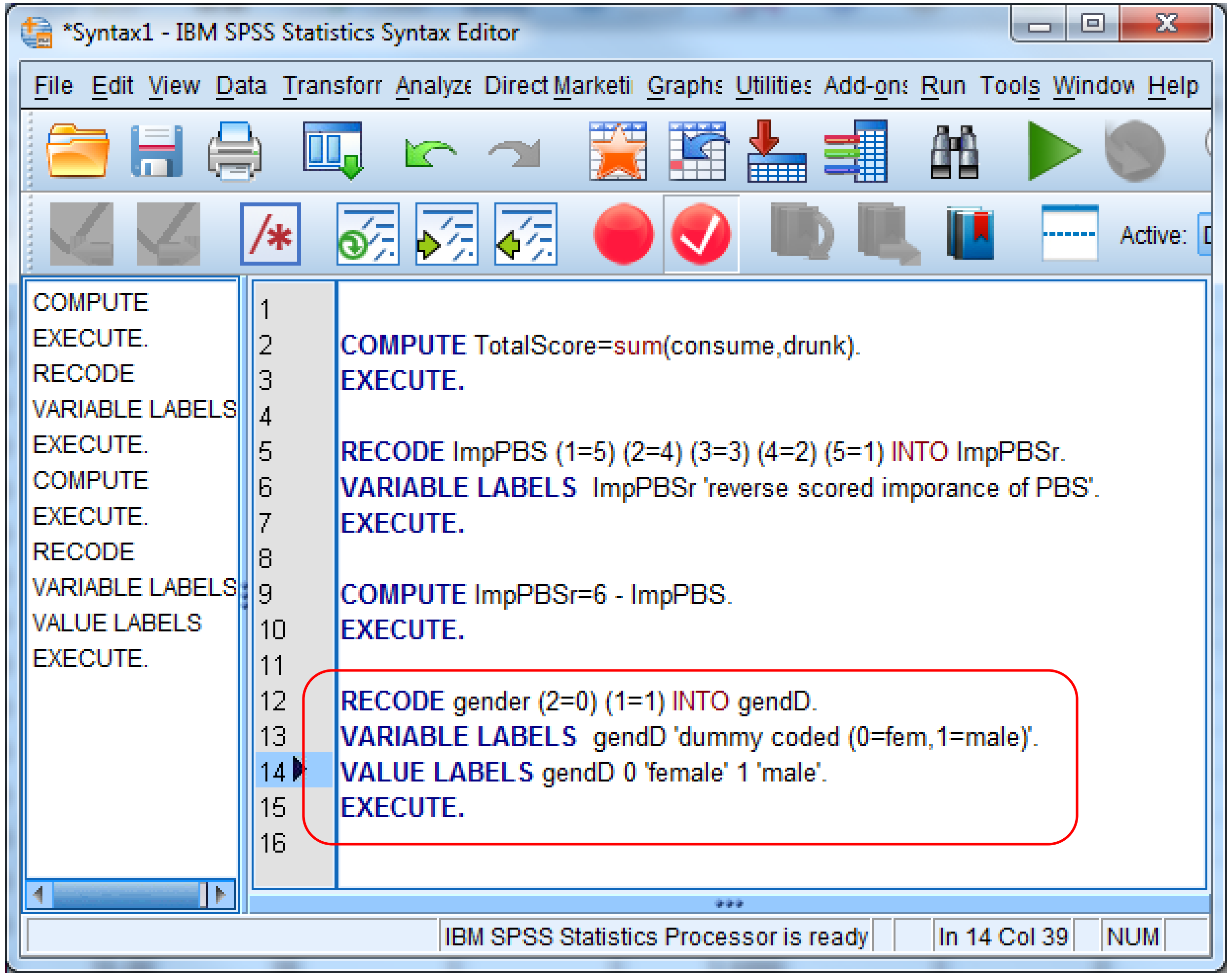
Name: gendD

Label: dummy coded (0=fem,1=n

election condition)

Dummy Codes: Gender

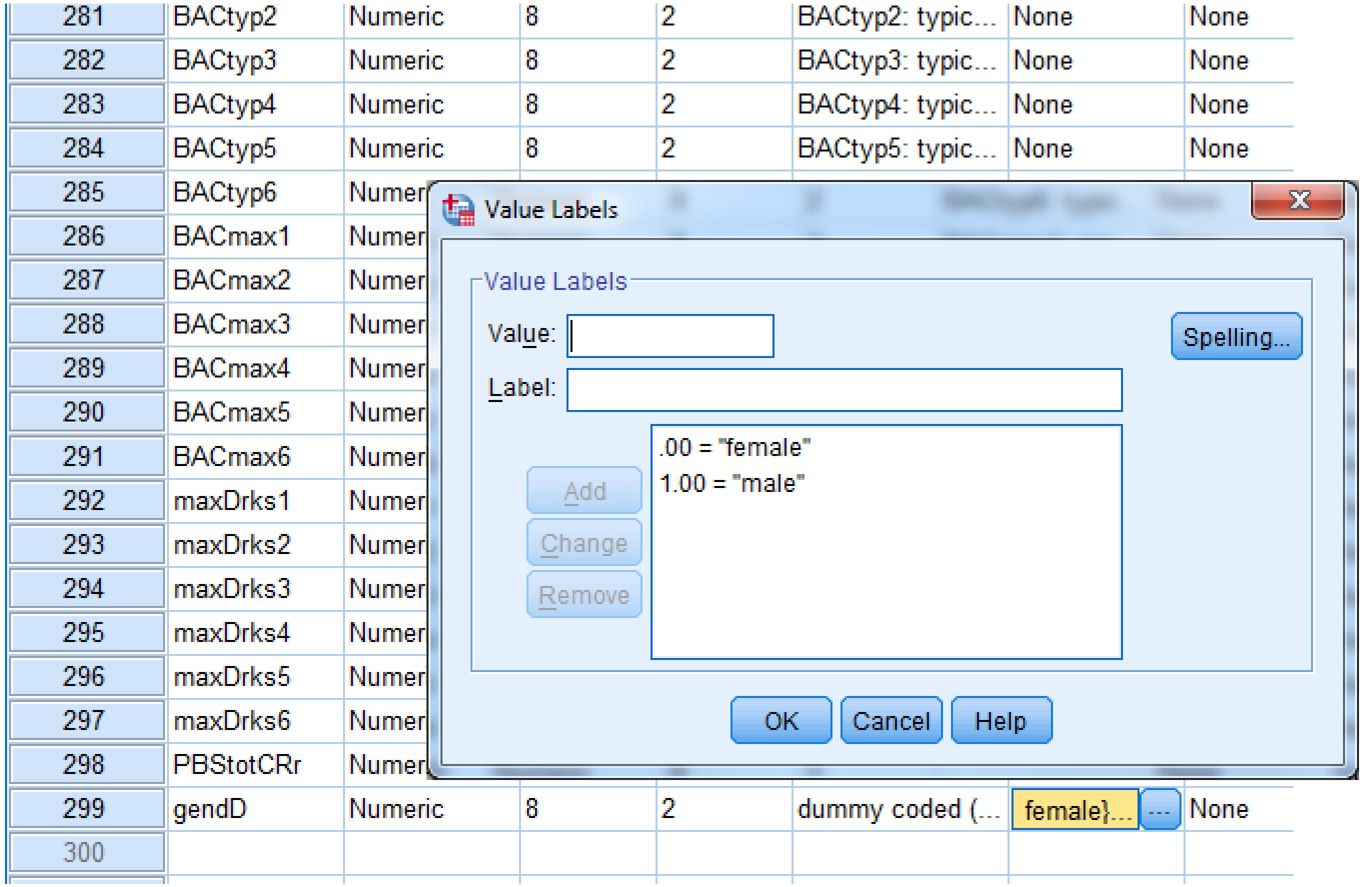
- Same as previous
- Added VALUE LABELS
  - “VALUE LABELS”
  - Variable name
  - Value
  - ‘label’
  - Next value
  - ‘label’
  - period


*Syntax1 - IBM SPSS Statistics Syntax Editor
File Edit View Data Transforr Analyze Direct Marketi Graphs Utilities Add-ons Run Tools Window Help
Active:
COMPUTE
EXECUTE.
RECODE
VARIABLE LABELS
EXECUTE.
COMPUTE
EXECUTE.
RECODE
VARIABLE LABELS
VALUE LABELS
EXECUTE.
IBM SPSS Statistics Processor is ready
In 14 Col 39
NUM

```

COMPUTE TotalScore=sum(consume,drunk).
EXECUTE.

RECODE ImpPBS (1=5) (2=4) (3=3) (4=2) (5=1) INTO ImpPBSr.
VARIABLE LABELS  ImpPBSr 'reverse scored imporance of PBS'.
EXECUTE.

COMPUTE ImpPBSr=6 - ImpPBS.
EXECUTE.

RECODE gender (2=0) (1=1) INTO gendD.
VARIABLE LABELS  gendD 'dummy coded (0=fem,1=male)'.
VALUE LABELS gendD 0 'female' 1 'male'.
EXECUTE.

```

Dummy Codes: Gender

- Double check variable view



281	BACtyp2	Numeric	8	2	BACtyp2: typic...	None	None
282	BACtyp3	Numeric	8	2	BACtyp3: typic...	None	None
283	BACtyp4	Numeric	8	2	BACtyp4: typic...	None	None
284	BACtyp5	Numeric	8	2	BACtyp5: typic...	None	None
285	BACtyp6	Numer					
286	BACmax1	Numer					
287	BACmax2	Numer					
288	BACmax3	Numer					
289	BACmax4	Numer					
290	BACmax5	Numer					
291	BACmax6	Numer					
292	maxDrks1	Numer					
293	maxDrks2	Numer					
294	maxDrks3	Numer					
295	maxDrks4	Numer					
296	maxDrks5	Numer					
297	maxDrks6	Numer					
298	PBStotCRr	Numer					
299	gendD	Numeric	8	2	dummy coded (...	female}...	None
300							

Value Labels

Value Labels

Value:

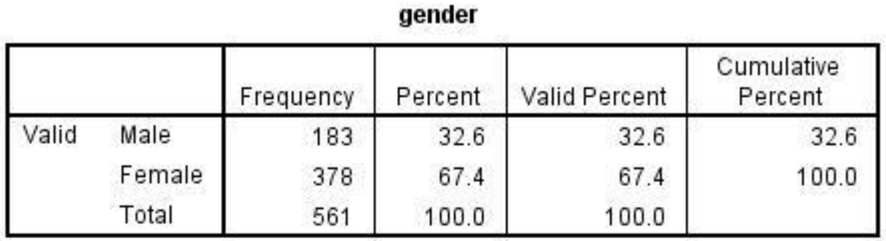
Label:

.00 = "female"
1.00 = "male"

Dummy Codes: Gender

- Double Check frequencies

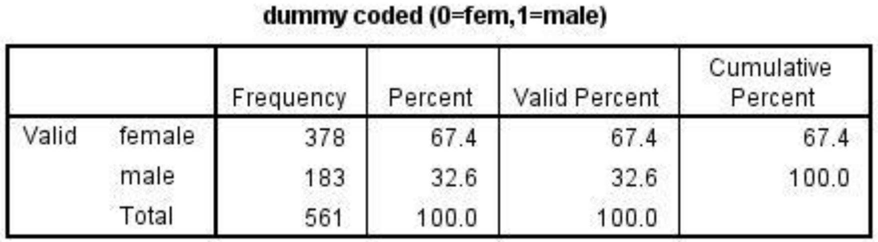
Original →



gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	183	32.6	32.6	32.6
	Female	378	67.4	67.4	100.0
	Total	561	100.0	100.0	

Dummy →

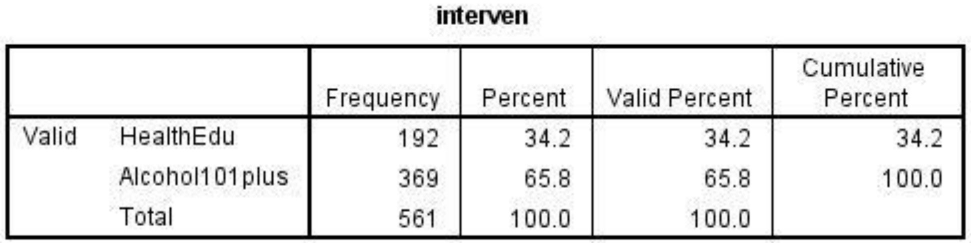


dummy coded (0=fem,1=male)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	female	378	67.4	67.4	67.4
	male	183	32.6	32.6	100.0
	Total	561	100.0	100.0	

Dummy Codes: Tx

- Tx example:
- Control Group (“HealthEdu”) is the default → 0
- Tx group (“Alcohol 101 plus”) → 1
- RECODE into Different Variable
    - RECODE interven (1=0) (2=1) INTO txD.
    - VARIABLE LABELS txD 'dummy coded tx (0=ctrl,1=alc101)'.
    - VALUE LABELS txD 0 ‘ctrl' 1 ‘alc101'.
    - EXECUTE.
- Double check in variable view and using frequencies

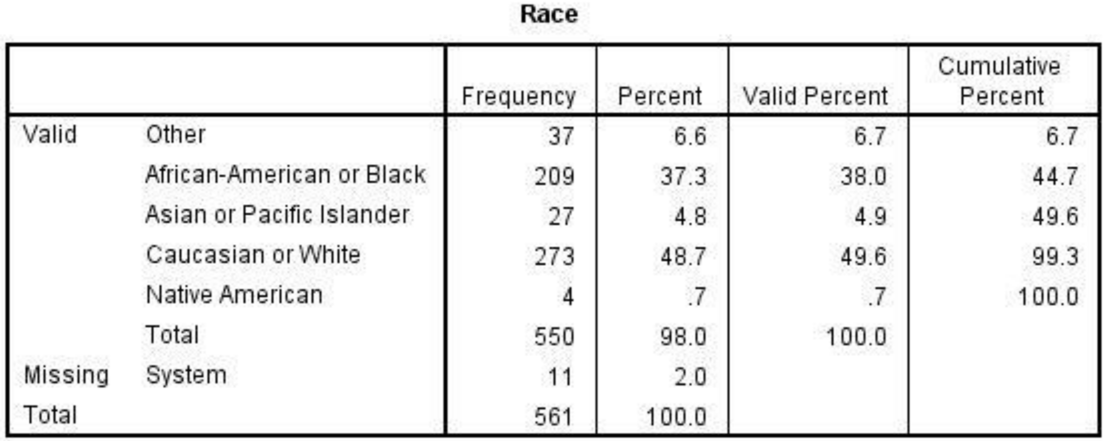


interven

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	HealthEdu	192	34.2	34.2	34.2
	Alcohol101plus	369	65.8	65.8	100.0
	Total	561	100.0	100.0	

Dummy Codes: Race

- Look at frequencies and choose reference group
- White and Black are both very frequent
- Is one group the "default"?
- No, so White → 0
  - Actually, *k* = 5, so *k*-1 = 4 new dummy variables
  - White → 0,0,0,0

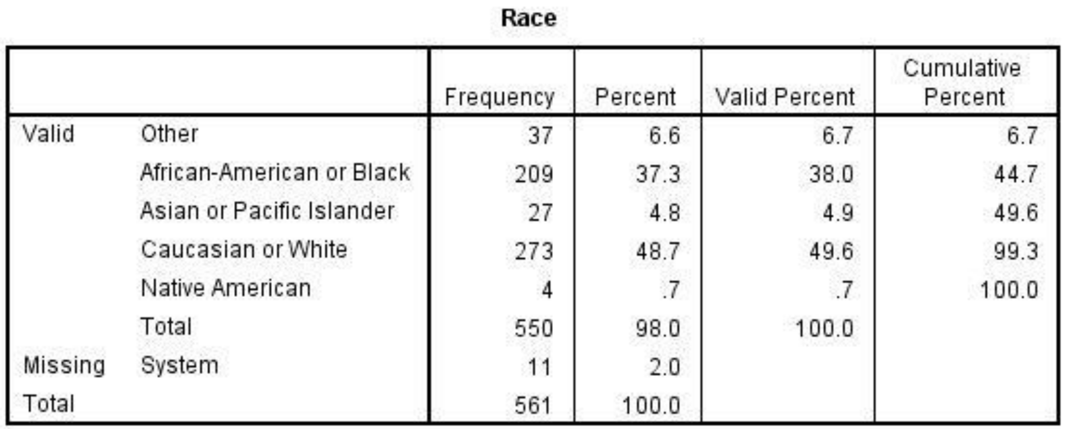


Race

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Other	37	6.6	6.7	6.7
	African-American or Black	209	37.3	38.0	44.7
	Asian or Pacific Islander	27	4.8	4.9	49.6
	Caucasian or White	273	48.7	49.6	99.3
	Native American	4	.7	.7	100.0
	Total	550	98.0	100.0	
Missing	System	11	2.0		
Total		561	100.0		

Dummy Codes: Race

- *k* = 5, so *k*-1 = 4 new dummy variables
- White → 0,0,0,0
- Each non-reference group gets its own dummy variable
- raceD4 → Other =1
- raceD3 → African-American or Black =1
- raceD2 → Asian or Pacific Islander =1
- raceD1 → Native American =1



Race

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Other	37	6.6	6.7	6.7
	African-American or Black	209	37.3	38.0	44.7
	Asian or Pacific Islander	27	4.8	4.9	49.6
	Caucasian or White	273	48.7	49.6	99.3
	Native American	4	.7	.7	100.0
	Total	550	98.0	100.0	
Missing	System	11	2.0		
Total		561	100.0		

Dummy Codes: Race

```
RECODE race (1=0) (2=0) (3=0) (4=0) (5=1) INTO raceD1.
RECODE race (1=0) (2=0) (3=1) (4=0) (5=0) INTO raceD2.
RECODE race (1=0) (2=1) (3=0) (4=0) (5=0) INTO raceD3.
RECODE race (1=1) (2=0) (3=0) (4=0) (5=0) INTO raceD4.
VARIABLE LABELS raceD1 'dummy coded race (1=other)'
   raceD2 'dummy coded race (1=Black)'
   raceD3 'dummy coded race (1=Asian/PI)'
   raceD4 'dummy coded race (1=NativeAmer)'.
VALUE LABELS raceD1  1 ‘other’
   raceD2  1 ‘American-American or Black’
   raceD3  1 ‘Asian or Pacific Islander’
   raceD4  1 ‘Native American’.
EXECUTE.
```

VARIABLE LABELS [variable name] [‘label’]
Period after last one ends command.

VALUE LABELS [variable name] [#] [‘label’]
Period after last one ends command.

Dummy Codes: Race

ID	Race	Original	raceD1	raceD2	raceD3	raceD4
1	Asian	3	0	1	0	0
2	White	4	0	0	0	0
3	Black	2	0	0	1	0
4	Native American	5	1	0	0	0
5	White	4	0	0	0	0
6	Other	1	0	0	0	1
7	Black	2	0	0	1	0
8	Black	2	0	0	1	0
9	Other	1	0	0	0	1
10	White	4	0	0	0	0

Dummy Codes: Race

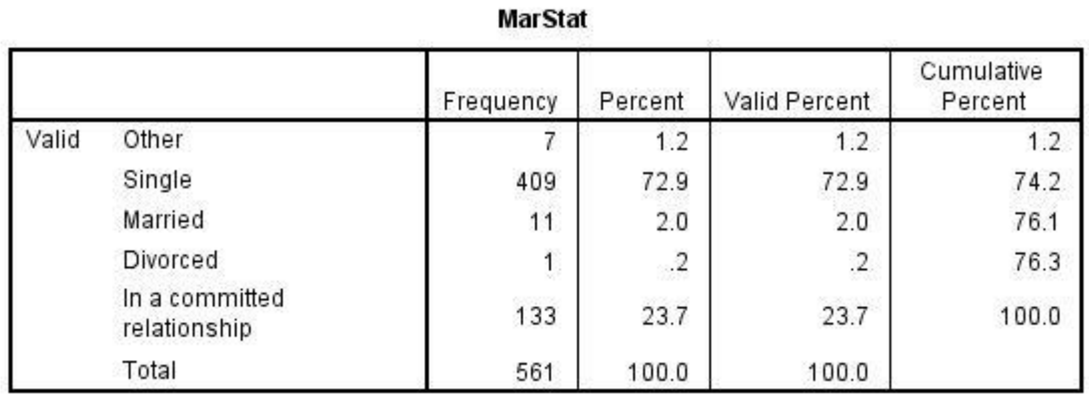
raceD4 → Other =1
raceD3 → African-American or Black =1
raceD2 → Asian or Pacific Islander =1
raceD1 → Native American =1

ID	Race	Original	raceD1	raceD2	raceD3	raceD4
1	Asian	3	0	**1**	0	0
2	White	4	0	0	0	0
3	Black	2	0	0	**1**	0
4	Native American	5	**1**	0	0	0
5	White	4	0	0	0	0
6	Other	1	0	0	0	**1**
7	Black	2	0	0	**1**	0
8	Black	2	0	0	**1**	0
9	Other	1	0	0	0	**1**
10	White	4	0	0	0	0

Dummy Codes

- How do you choose between two methods (one dummy code vs series)?
- After data collection is finished, you can see if all non-majority categories are similar (or at least not very different) from the majority category
- For example, if “On-campus” is the majority category for *residence*, and you want to collapse across all other participants to represent “other residence”, then if “Fraternity/sorority house”, “Off-campus with friends”, “Off-campus with family”, etc. means are all lower then “On-campus” means for alcohol quantity, this makes sense to do
- But if one mean is higher than “On-campus” and one mean is lower than “On-campus”, you can't collapse because you'd hide the effects (by combining increases with decreases)

Dummy Code: Marital Status



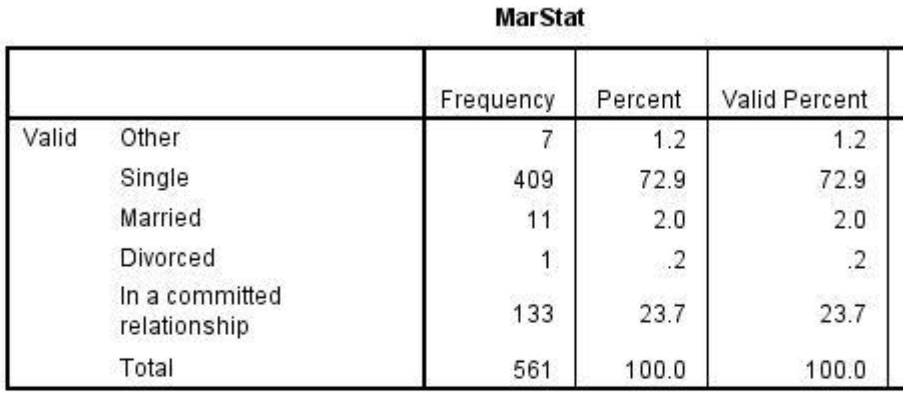
MarStat

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Other	7	1.2	1.2	1.2
	Single	409	72.9	72.9	74.2
	Married	11	2.0	2.0	76.1
	Divorced	1	.2	.2	76.3
	In a committed relationship	133	23.7	23.7	100.0
	Total	561	100.0	100.0	

- Most frequent group Single→0
- Other groups are VERY small (≤ 2%)
- Option 1
  - 4 variables for 5 groups
  - Three of those groups are VERY small, so those variables do not add much
- Option 2
  - 1 variable for 5 groups
  - Single (73%) versus everyone else
- Option 3
  - Meaningful groupings based on frequency AND definitions
  - Single people (72.9%) versus committed OR married (23.7 + 2.0 = 25.7%)
  - Excludes/deletes *n*=8 (“other” or divorced; 1.4%)

Dummy Code: Marital Status

- Option 1: 4 variables
  - marD1 → Other =1
  - marD2 → Married =1
  - marD3 → Divorced =1
  - marD4 → Committed =1
  - Single is 0,0,0,0
- Option 2: 1 variable for 5 groups
  - Single (73%) versus everyone else
  - Single → 0, everyone else → 1
- Option 3: 1 variable for meaningful groupings
  - MaritalD: 2 → 0 (single); 3 → 1, 5 → 1 (Committed or Married = 1)
  - Excludes/deletes *n*=8 (“other” or divorced)



MarStat

		Frequency	Percent	Valid Percent
Valid	Other	7	1.2	1.2
	Single	409	72.9	72.9
	Married	11	2.0	2.0
	Divorced	1	.2	.2
	In a committed relationship	133	23.7	23.7
	Total	561	100.0	100.0

Recoding Blanks

- There is one last type of “missing” data that is not a true missing
- For some “check all that apply” questions, survey software will generate a variable of 1’s and blanks rather than 1’s and 0’s
    - e.g., race, symptoms, problems, etc.
- [1,.] vs [1,0]
- Not necessary if creating totals using SUM or MEAN then deleting items
    - IS necessary if imputing item variables (blanks ARE data, not missing)
    - IS necessary if using “+” coding approach

Recoding Blanks

x1	*x2*	*x3*	*x4*	*x5*	*x6*
.	2	.	.	.	.
1	.	3	4	.	.
.	.	.	.	.	6
.	2	.	4	5	.
1	2	.	.	5	.

x1	*x2*	*x3*	*x4*	*x5*	*x6*
0	1	0	0	0	0
1	0	1	1	0	0
0	0	0	0	0	1
0	1	0	1	1	0
1	1	0	0	1	0

- [1,.] vs [1,0]

```
RECODE x1 x2 x3 x4 x5 x6 (2 thru 6=1) (MISSING=0).
EXECUTE.
RECODE x1 x2 x3 x4 x5 x6 (MISSING=0).
EXECUTE
```

Recoding Blanks

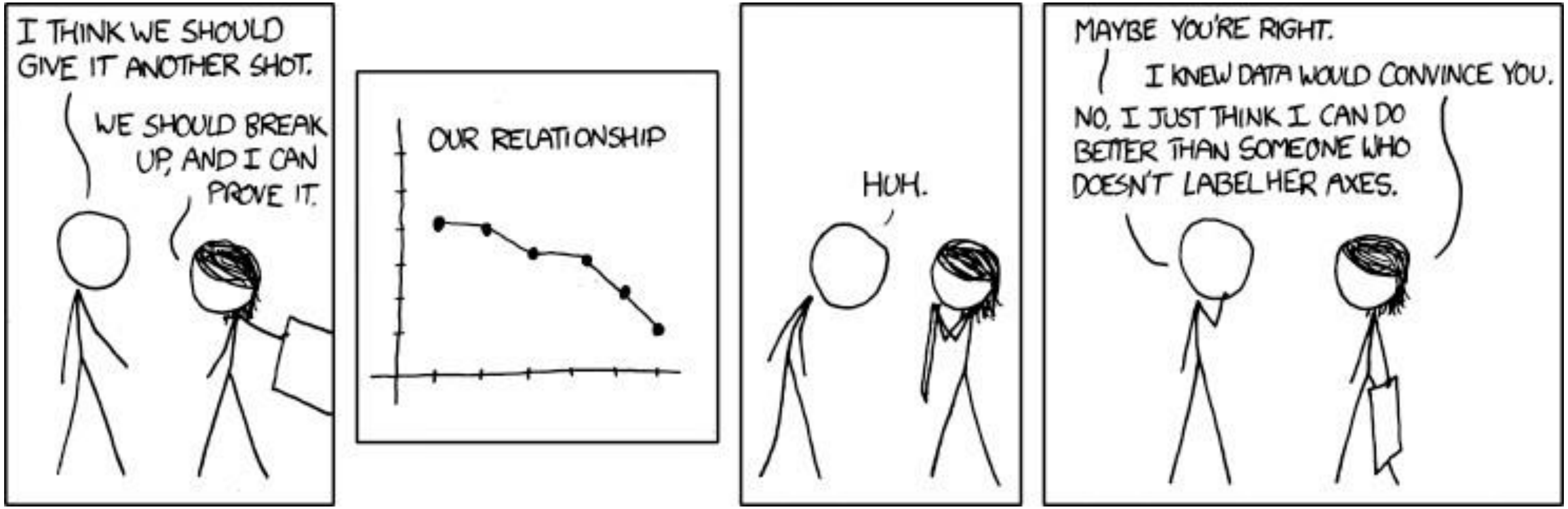
- Want to make sure you are not giving someone all 0’s if they skipped the questionnaire
    - Did they experience NO alcohol-related problems? Or did they just skip the YAACQ?
- Can sum responses to see if they endorsed any item at all
- Can recode only for those who completed the questionnaire (DO IF command)
- What if they really didn’t experience problems?
    - Can be helpful to either have a preliminary item (did you seek any form of treatment? If so, select all that apply) or a “none of the above” option so SOMETHING is selected
    - If you didn’t do either of those things? Less certainty, but can see if they answered items later in the survey

When to Impute/Recode

- Impute then Recode?
  - Total scores can be created with all datapoints (including imputed values)
  - Recoding can be more difficult (numbers often have decimal points)
    - Particularly relevant for reverse scoring and dummy codes
  - Nominal variables are not contributing to imputation estimations, whereas they could if they were recoded first
- Recode then Impute?
  - Can include more variables in imputation analysis
    - Categorical variables are excluded, but not dummy codes
  - Total scores are already made
    - Already falsely deflated or mean value without random error

When to Impute/Recode

- Recode then Impute then Recode
    - Dummy code and reverse score first
        - Now all variables are continuous, item-level
    - Impute
        - Now all values are unbiased (no deflated sums or reduced SE’s)
    - Create composites/total scores with imputed variables
        - Means/Totals reflect unbiased estimates
            - At least, not due to bad imputation
- Recode then use ML estimation for analysis
    - If doing true SEM and using item-level analysis, then no composites are required
    - For path analyses, etc., might create composites first
    - OR can create composites in Mplus using “DEFINE”, which still uses ML


I THINK WE SHOULD GIVE IT ANOTHER SHOT.
WE SHOULD BREAK UP, AND I CAN PROVE IT.
OUR RELATIONSHIP
HUH.
MAYBE YOU'RE RIGHT.
I KNEW DATA WOULD CONVINCE YOU.
NO, I JUST THINK I CAN DO BETTER THAN SOMEONE WHO DOESN'T LABEL HER AXES.

Outline for Today

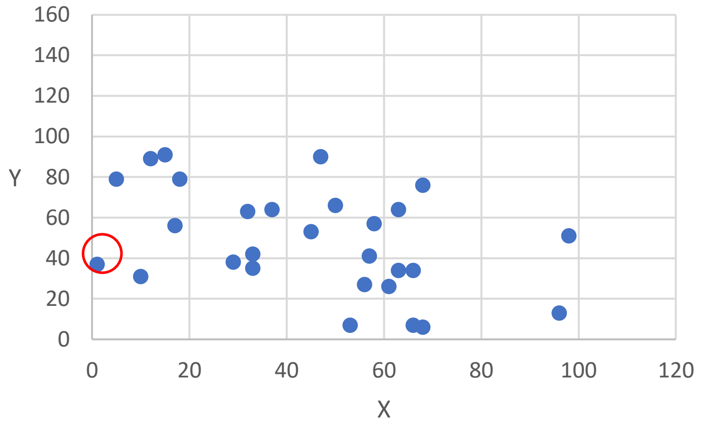
- Missing Data
  - Identifying, assessing type, imputation options
- Composite Scores
  - Total scores, recoding, dummy coding
- **Outliers**
  - **Identifying and addressing univariate and multivariate outliers**
- Normality
  - Assessing and addressing (e.g., transformations, analysis specifications)
- Bivariate Linearity
  - Reading scatterplots and what to do about them
- Documentation
  - The importance of codebooks and data logs

Outliers

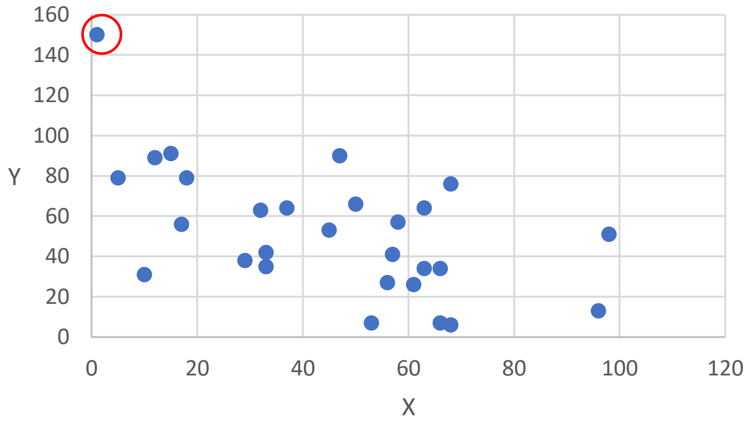
- How to identify them
- What to do about them
- Start with univariate (one variable at a time)
- Touch on multivariate

Outliers

- Why do we care?
  - Have a stronger influence on the data
  - Can influence results of study



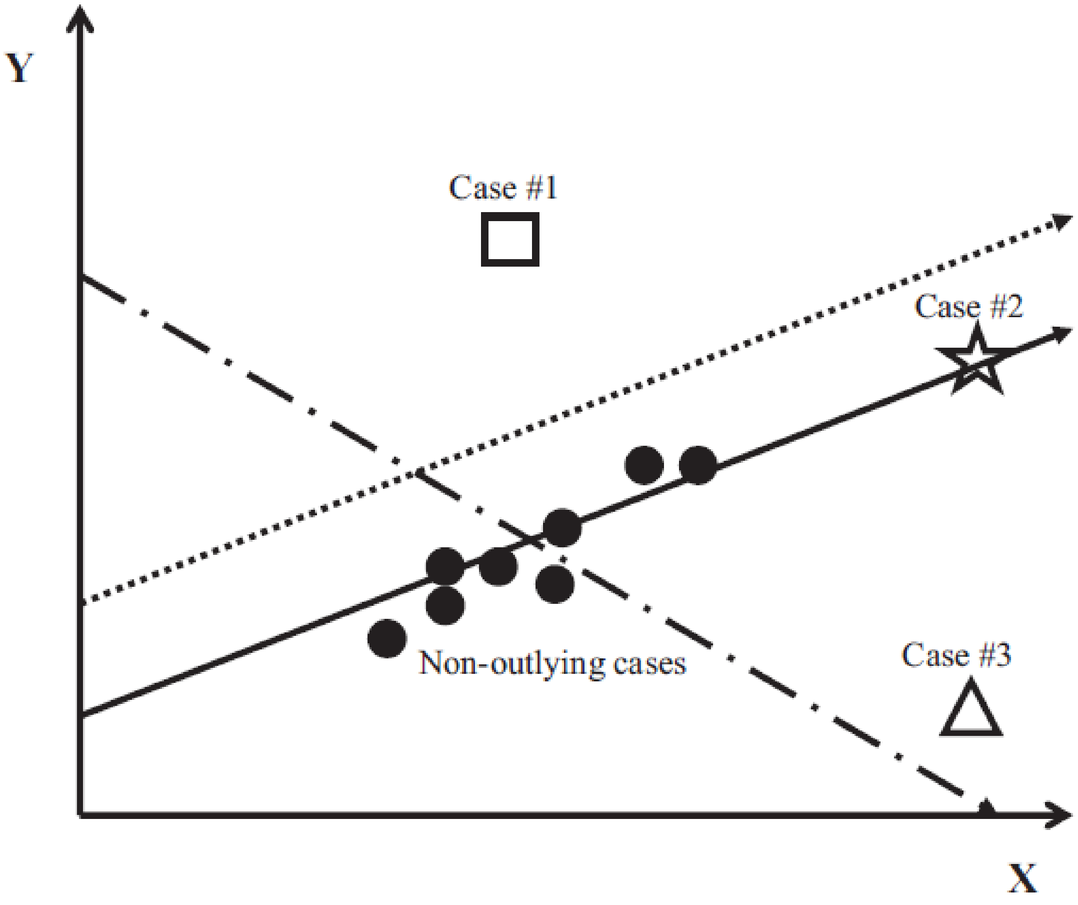
r = -.426



r = -.567

- Same data, changed one value from 37 to 150

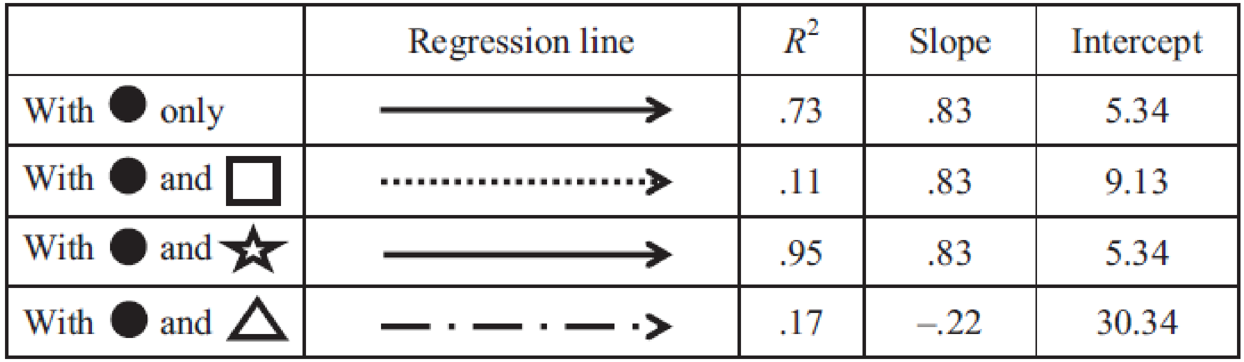
Outliers



Case 1 (□) = Large residual value, reasonable X

Case 2 (☆) = Small residual, extreme X

Case 3 (△) = Large residual value, extreme X



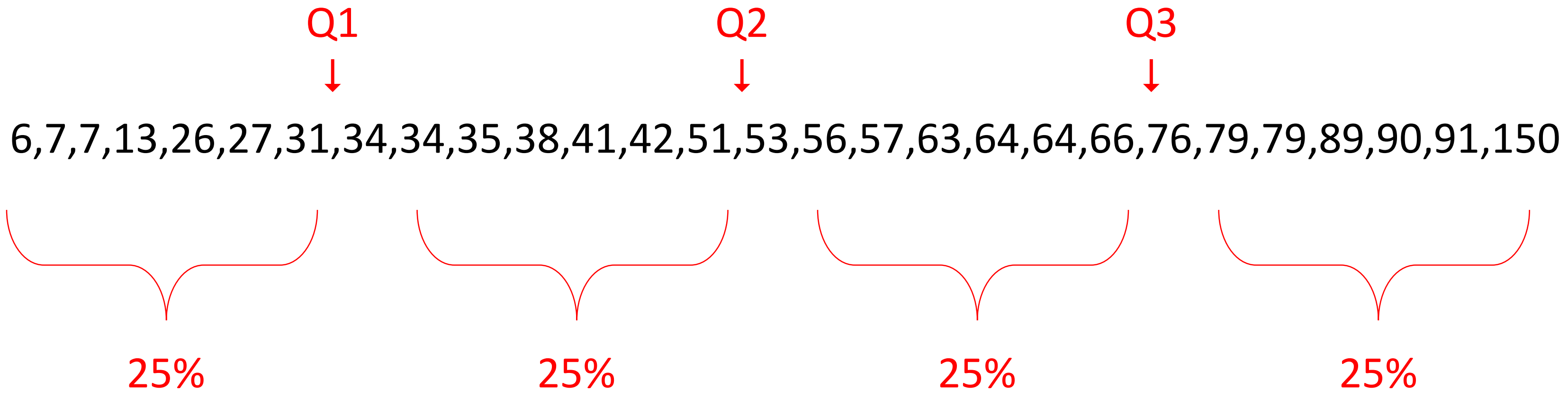
	Regression line	R^2	Slope	Intercept
With ● only	⟶	.73	.83	5.34
With ● and □	⋯⋯⋯→	.11	.83	9.13
With ● and ☆	⟶	.95	.83	5.34
With ● and △	— · — · — →	.17	–.22	30.34

Identifying Outliers

- Do this on the final version of your variables
    - After composites have been created
    - After imputation (if imputing)
- Standard deviations?
    - Themselves influenced by extreme values
- *SD* of [34,35,41,56,71,75] = 16.53
    - *M* = 52.0; *M* + *SD* = 52.0 + 16.53 = 68.53
    - *M* + 2 *SD*'s = 85.06
- *SD* of [34,35,41,56,71,150] = 40.37
    - *M* = 64.5; *M* + *SD* = 64.5 + 40.37 = 104.87
    - *M* + 2 *SD*'s = 145.24
    - *M* + 3 *SD*'s = 185.61
- So having extreme values makes it harder to detect extreme values

Identifying Outliers

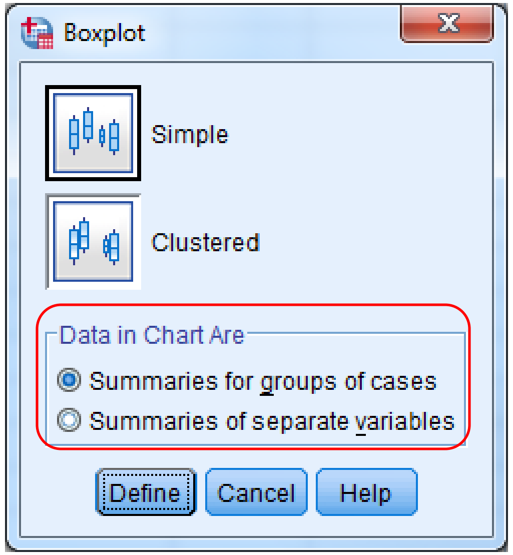
- What do you do? Boxplots!
  - Rely on Q1, Q2, Q3 (from IQR [InterQuartile Range])
  - Q2 = median
  - Q1 equals sub-median below lowest and median
  - Q3 equals sub-median below highest and median



6,7,7,13,26,27,31,34,34,35,38,41,42,51,53,56,57,63,64,64,66,76,79,79,89,90,91,150

Boxplots

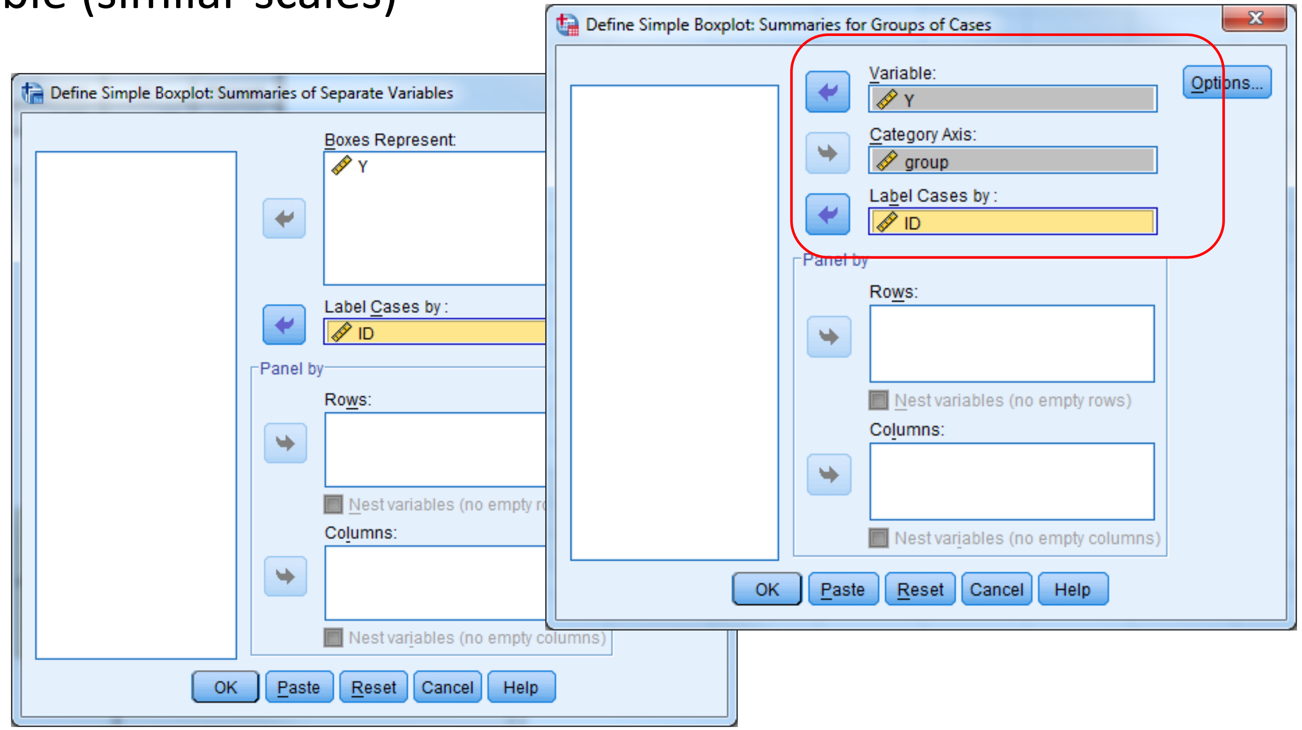
- Graphs > Legacy Dialogs > Boxplots
- Groups? Or no groups?
  - Groups makes sense if group matters to your design, and if they are balanced
  - Can be overly sensitive if some groups have a small *n*
- Can explore more than one variable (similar scales)



Data in Chart Are

● Summaries for groups of cases

○ Summaries of separate variables



Define Simple Boxplot: Summaries of Separate Variables
OK
Paste
Reset
Cancel
Help

Boxes Represent: Y

Label Cases by: ID

Panel by

Rows:

□ Nest variables (no empty r

Columns:

□ Nest variables (no empty columns)

Define Simple Boxplot: Summaries for Groups of Cases
Options...
OK
Paste
Reset
Cancel
Help

Variable: Y

Category Axis: group

Label Cases by: ID

Panel by

Rows:

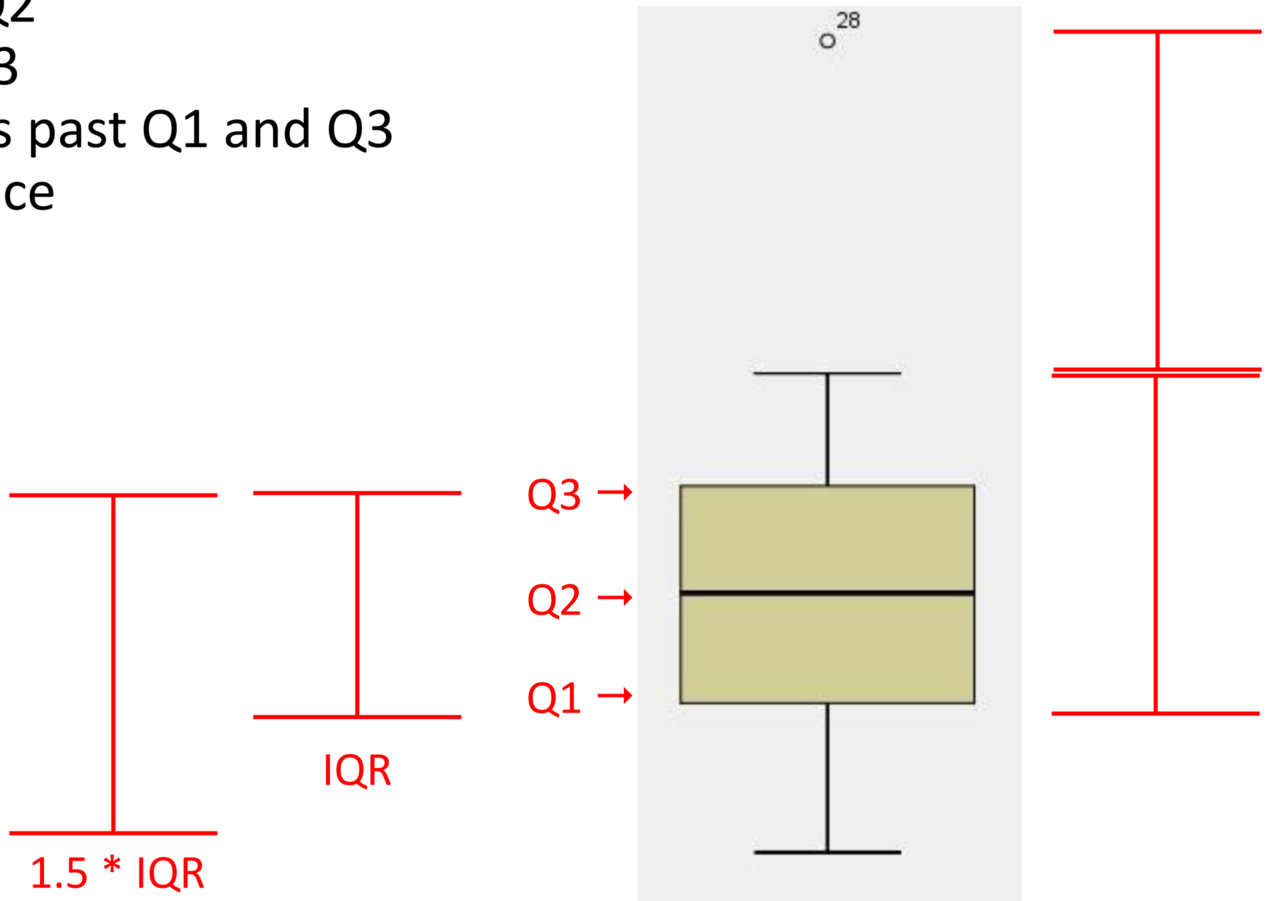
□ Nest variables (no empty rows)

Columns:

□ Nest variables (no empty columns)

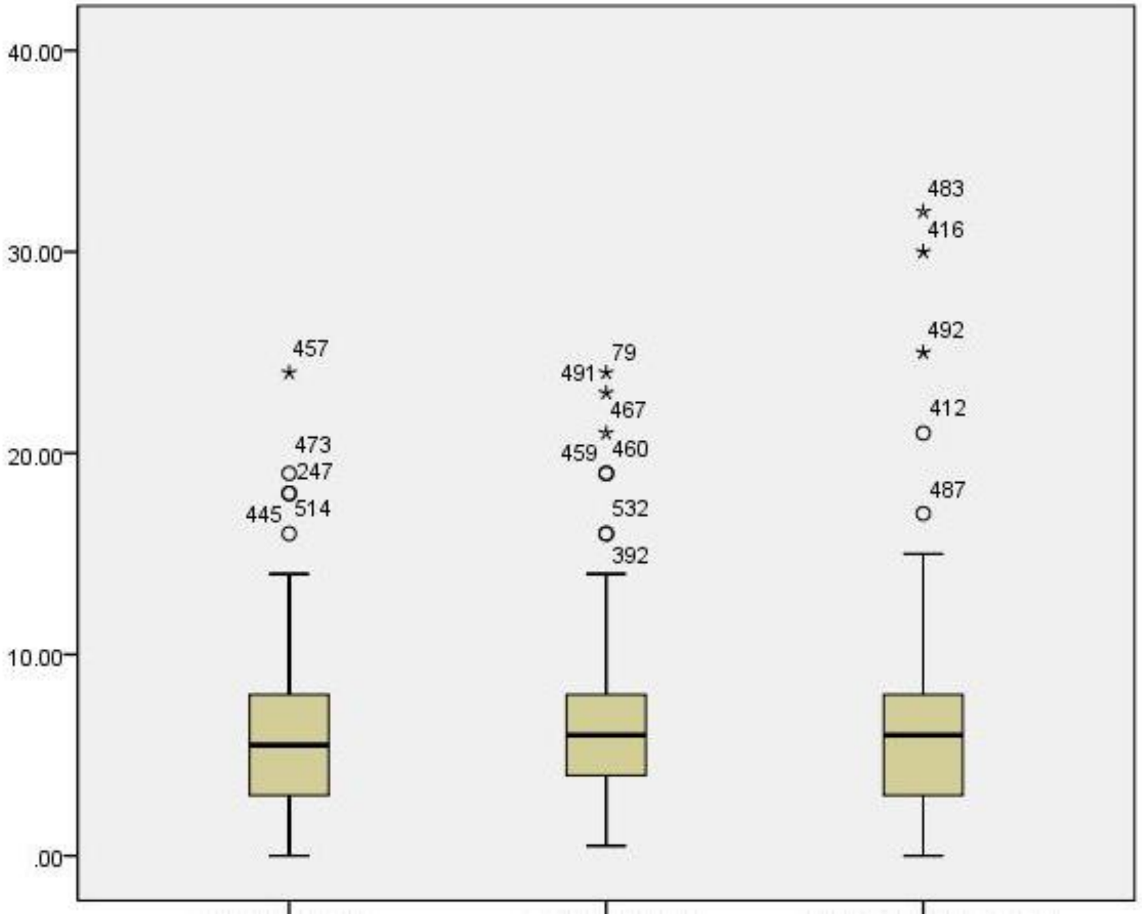
Boxplots

- Box and “whiskers” or “fence”
  - Line/middle = median = Q2
  - Box = IQR = from Q1 to Q3
  - Whiskers/fence = 1.5 IQRs past Q1 and Q3
  - Circles = 1.5 IQRs past fence
  - Asterisks = beyond circles
  - What’s extreme??
    - Circles? Asterisks?



Boxplots

- How many outliers do we have?
- What do we want to do with them?



Addressing Outliers

- Deletion or “Trimming”
  - Deleting values or removing entire cases deemed an outlier
- Good: Indicates noncompliance or a “bad” participant
  - Delete whole case
  - Reaction time indicates participant fell asleep (too long) or wasn’t paying attention (too short)
- Bad: Just an extreme value
  - Bill Gates salary is valid/accurate
  - He is someone with average education and a very high salary
  - Deleting valid cases biases your sample

Addressing Outliers

- Winsorize
  - Make the value less extreme
  - Find not extreme value, and go beyond it
    - Maintain rank among multiple outliers
- Examples:
- One outlier: [34,35,41,56,71,150]

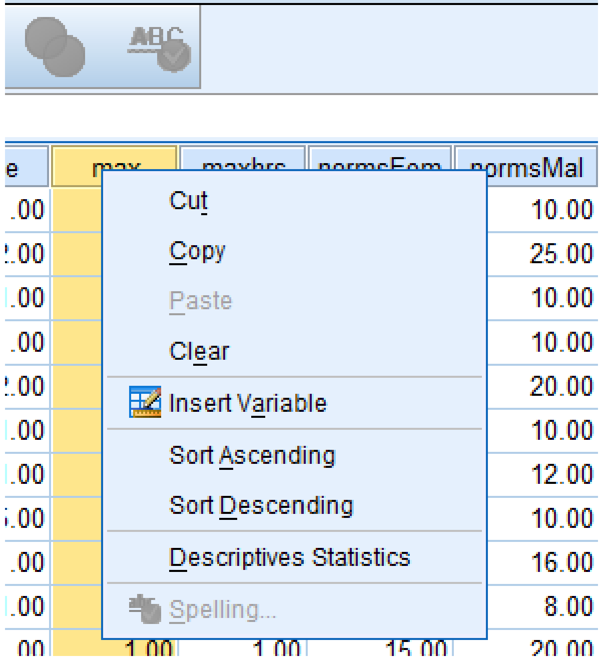
→[34,35,41,56,71,72]

- Two outliers: [34,35,41,56,71,150]

→[34,35,41,56,57,58]

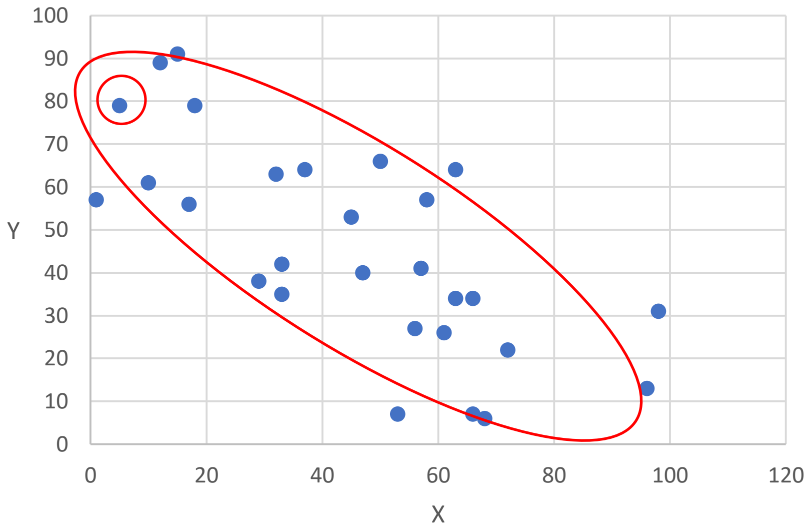
Addressing Outliers

- Winsorize in SPSS
  - No fancy syntax
- Sort cases
  - Right click on relevant variable
  - Sort ascending if looking for low outliers
  - Sort “descending” if looking for high outliers
  - Repeat for each variable
- Use discretion and common sense
  - If values are 1.2, 1.3, 1.3, 1.4
  - Don’t change outliers to 2 and 3
  - Change to 1.5 and 1.6

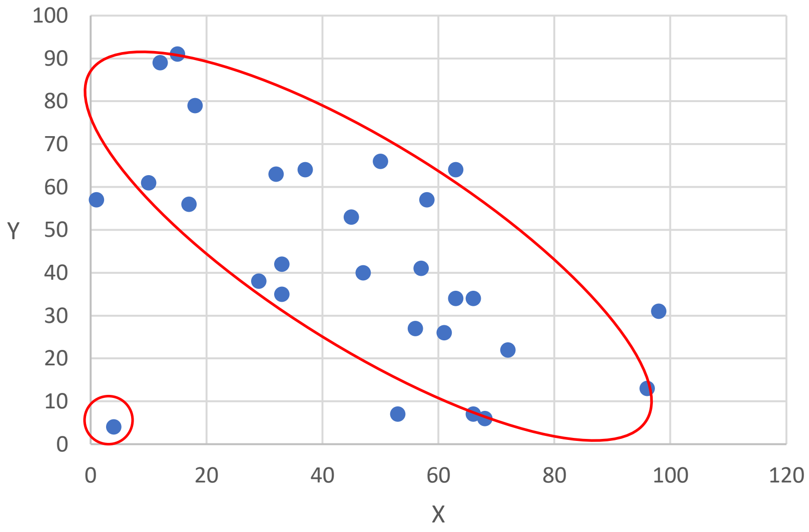


Multivariate Outliers

- Regression (linear analysis) specific
  - Requested in workshop survey; not for every analysis plan
- Value may be reasonable within its variable range
  - Not a univariate outlier
- Pattern across multiple variables indicates the combination is extreme
  - Can greatly influence analysis



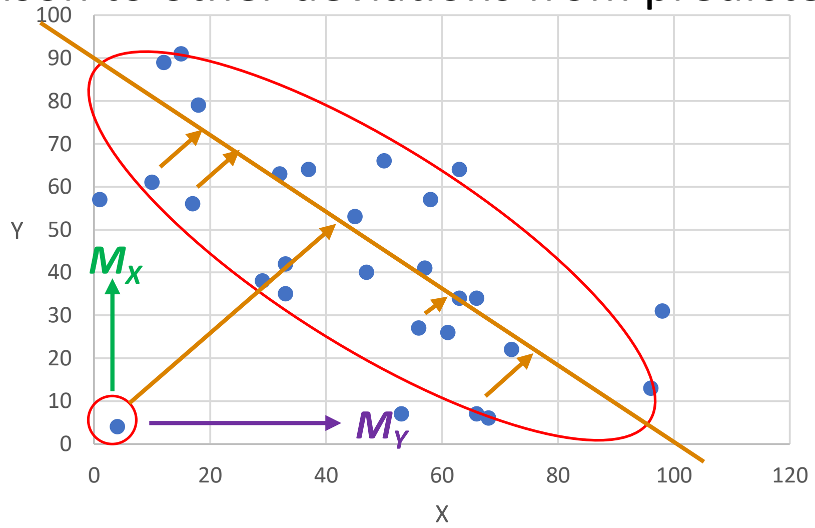
r = -.710



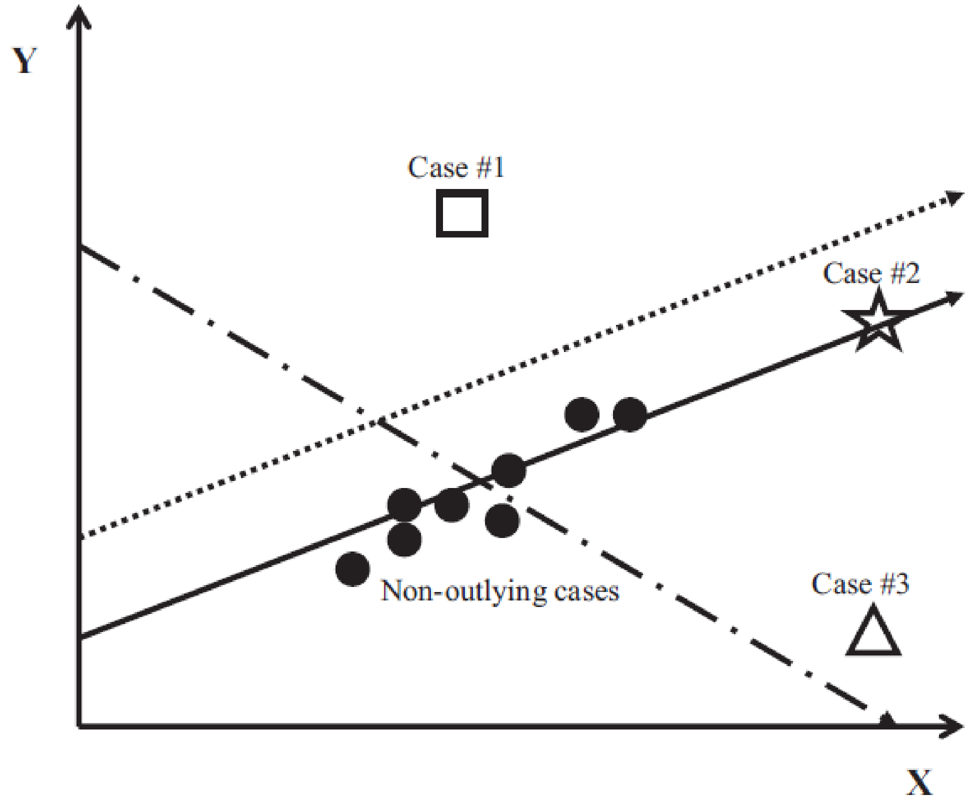
r = -.522

Multivariate Outliers

- Assumption of regression (and some other analyses)
- How can you examine?
  - Leverage, discrepancy, and influence
  - Rely on residuals
    - Distance from mean may not be large, but distance from predicted value may be huge
    - Especially in comparison to other deviations from predicted value



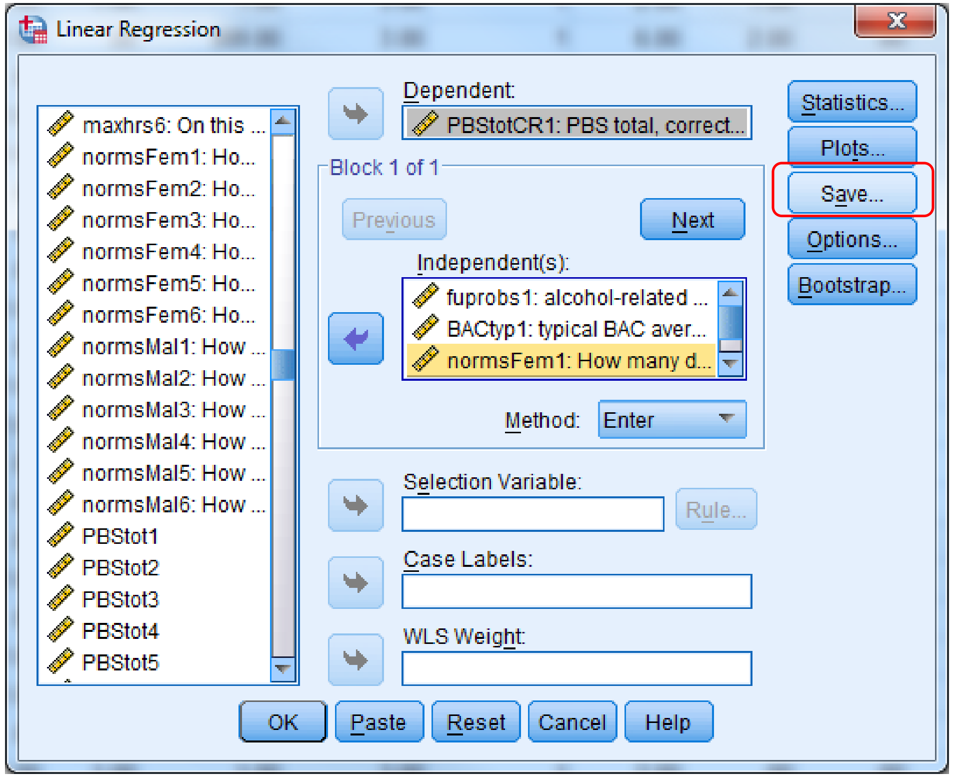
Multivariate Outliers



- 3 approaches to asses
  - Multiple indicators for each
- Leverage
  - Distance from centroid
    - Like a multivariate mean
      (center of data across variables)
  - Leverage, Mahalanobis Distance
- Discrepancy
  - Distance from regression line
  - “studentized residual”, “studentized deleted residual”
- Influence
  - How much a case affects the regression line
  - Compare regression coefficients with and without this case
  - DFFITStandardized, Cook’s D

Multivariate Outliers

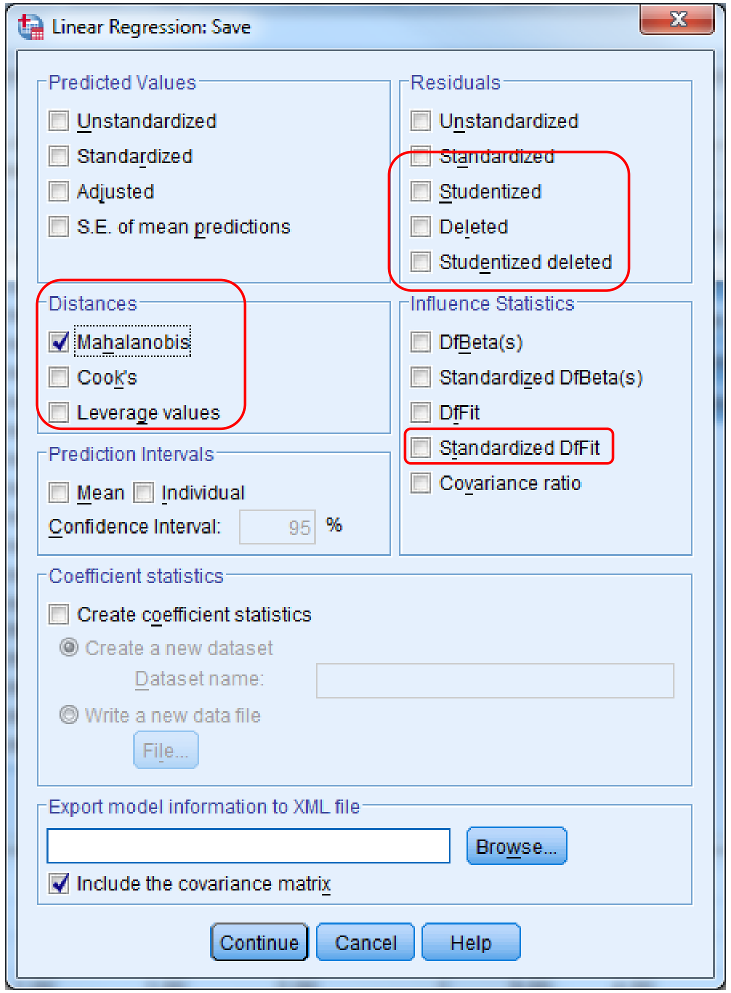
- The SPSS Regression command allows you to save the indicators for each case



Dependent: PBStotCR1: PBS total, correct...

Independent(s):
furprobs1: alcohol-related ...
BACtyp1: typical BAC aver...
normsFem1: How many d...

Method: Enter



Predicted Values
□ Unstandardized
□ Standardized
□ Adjusted
□ S.E. of mean predictions

Residuals
□ Unstandardized
□ Standardized
□ Studentized
□ Deleted
□ Studentized deleted

Distances
☑ Mahalanobis
□ Cook's
□ Leverage values

Influence Statistics
□ DfBeta(s)
□ Standardized DfBeta(s)
□ DfFit
□ Standardized DfFit
□ Covariance ratio

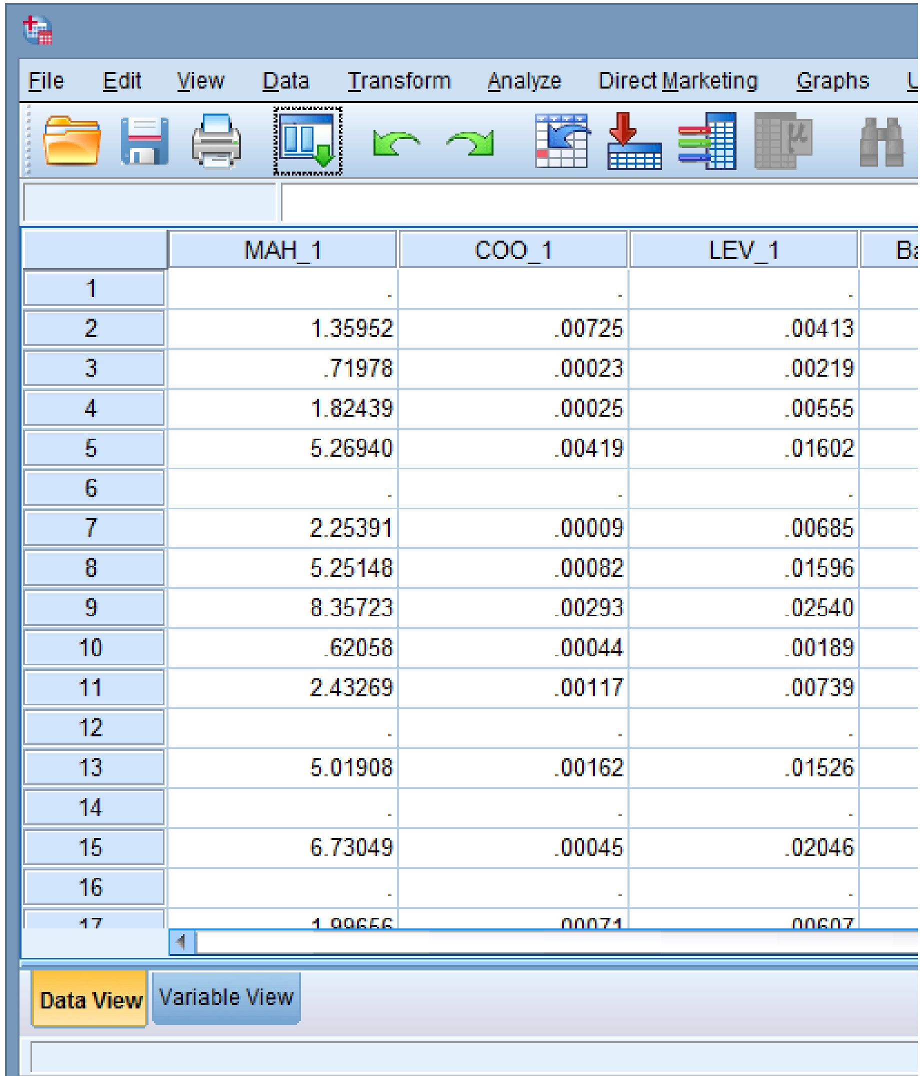
Prediction Intervals
□ Mean □ Individual
Confidence Interval: 95 %

Coefficient statistics
□ Create coefficient statistics
● Create a new dataset
Dataset name:
○ Write a new data file

Export model information to XML file
☑ Include the covariance matrix

Multivariate Outliers

- Saved as new variables in the dataset
- Scroll to the end of existing variables
- Sort descending (or ascending) to view potential outliers based on cutoffs



	MAH_1	COO_1	LEV_1	Ba
1	.	.	.	
2	1.35952	.00725	.00413	
3	.71978	.00023	.00219	
4	1.82439	.00025	.00555	
5	5.26940	.00419	.01602	
6	.	.	.	
7	2.25391	.00009	.00685	
8	5.25148	.00082	.01596	
9	8.35723	.00293	.02540	
10	.62058	.00044	.00189	
11	2.43269	.00117	.00739	
12	.	.	.	
13	5.01908	.00162	.01526	
14	.	.	.	
15	6.73049	.00045	.02046	
16	.	.	.	
17	1.99656	.00071	.00607	

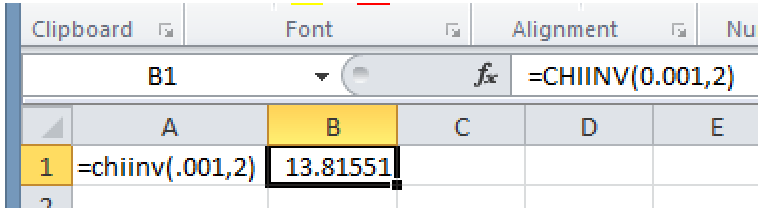
Multivariate Outliers

- Other software packages allow you to examine multivariate outliers as well (in contexts outside of regression, like SEM)
- In mplus:

```
PLOT:
OUTLIERS ARE  MAHALANOBIS;
              LOGLIKELIHOOD;
              INFLUENCE;
              COOKS;
```

Multivariate Outliers

- Cutoffs (excel can be helpful for critical values for distributions)
- Leverage
  - Leverage: Large samples= 2*k*/*n*, Small/medium samples: 3*k*/*n*
    - *k* represents constructs, not necessarily variables
    - Example: you are doing a regression with race and age as predictors, 200 cases
      - Even though race is 4 dummy variables, *k* is 2 (race + age), not 5 (raceD1 + raceD2 + raceD3 + raceD4 + age)
      - Cutoff = 3k/n = 3*2/200 = 0.03
  - Mahalanobis Distance: = $\chi^2_{CRIT}(k)$ [note that this is a chi-square critical value]
  - α = .001 or .01 is appropriate when assessing many assumptions
  - In excel: "=chiinv(α,*k*)"
- Discrepancy
  - "studentized residual" or "studentized deleted residual":
  - Use t_{CRIT} with a Bonferroni correction for alpha: α /*n*
  - In excel: "=tinv(α/*n*,*df*)"
- Influence
  - Standardized DFFIT: small/medium sample: >1, large sample: 2*sqrt([*k*+1]*n*)
  - Cook's D: In excel: "=finv(.5,*k*+1,*n*-*k*-1)" [not a typo… really use .5 as alpha]



=CHIINV(0.001,2)

	A	B
1	=chiinv(.001,2)	13.81551

Multivariate Outliers

- What do you do if you have multivariate outliers?
- Can delete cases
    - Only justified if the observations are contaminated in some way
- Model respecification (change IVs, etc.)
    - Can control for more covariates, or drop *ns* covariates
    - **Add moderators**! (*requires examination*)
- Variable transformation
    - Any border variables for skewness or kurtosis can be re-examined (next section)
- Robust approaches
    - Other than OLS (ML estimation, weighted least squares, etc.)
- Reducing extreme scores
    - i.e., Windsorizing, but only really works for univariate outliers
- Bootstrapping
    - Can be done in SPSS, Mplus, SAS, etc.

When to Exclude Multivariate Outliers

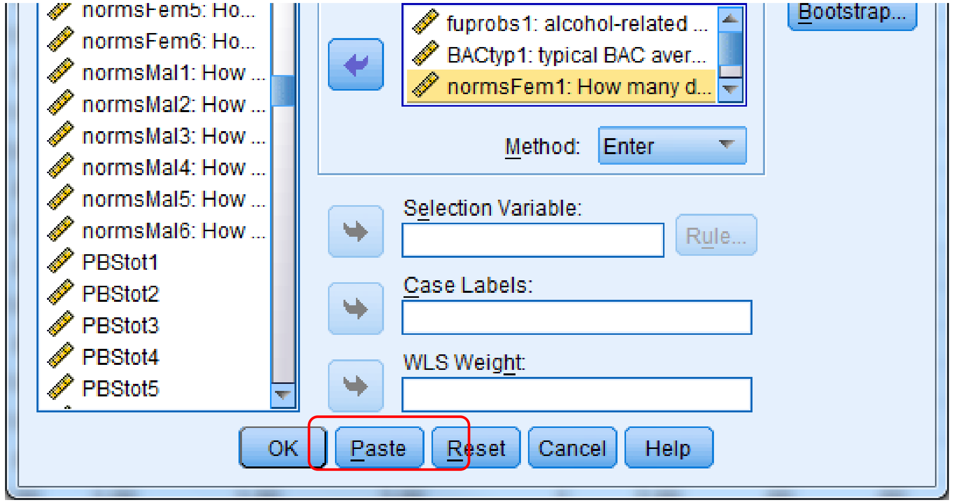
- Need to determine case-by-case (to see if illogical, or incompatible)
- Values on two or more variables are logically, or physically, incompatible (e.g., weighting lbs. 100 and being 6’ 5 tall or expressing support for a positively worded proposition and for the same, negatively worded, proposition).
    - Plausible univariate values, but implausible in multivariate context
- Responses on control questions aimed at verifying participants' attention should also be inspected. If the respondent is detected as a multivariate outlier and has also failed such a question, it may raise suspicion as to the validity of his/her responses.
    - Always include attention checks!
- In online surveys especially, participants may respond mechanically, not paying attention to the questions. As an alternative or supplement to control questions, the presence of systematic patterns (e.g., answering systematically at the extremes) should be checked and, if confirmed, can justify excluding outliers.
    - Check for patterns of all 5s, or all 5s and 1s, etc. Again, only catch in multivariate context.

When to Exclude Multivariate Outliers

- If outliers are associated with a specific condition or stimulus, rather than being randomly distributed among conditions, this suggests that an unknown factor was confounded with the manipulation and the problem may be greater than just the outliers. In such a situation, excluding them may not be appropriate, because it would violate random allocation.
  - Need to investigate procedure, identify problems, possibly include moderator
- All examples from:
  - Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology*, *74*, 150-156. doi: 10.1016/j.jesp.2017.09.011

Multivariate Outliers

- Unfortunately, requesting the necessary information to detect multivariate outliers requires running the analysis
- If you have to make a change, you have to re-run the analysis after your update
- Saving your syntax is helpful
  - Use the “paste” button in ALL command windows

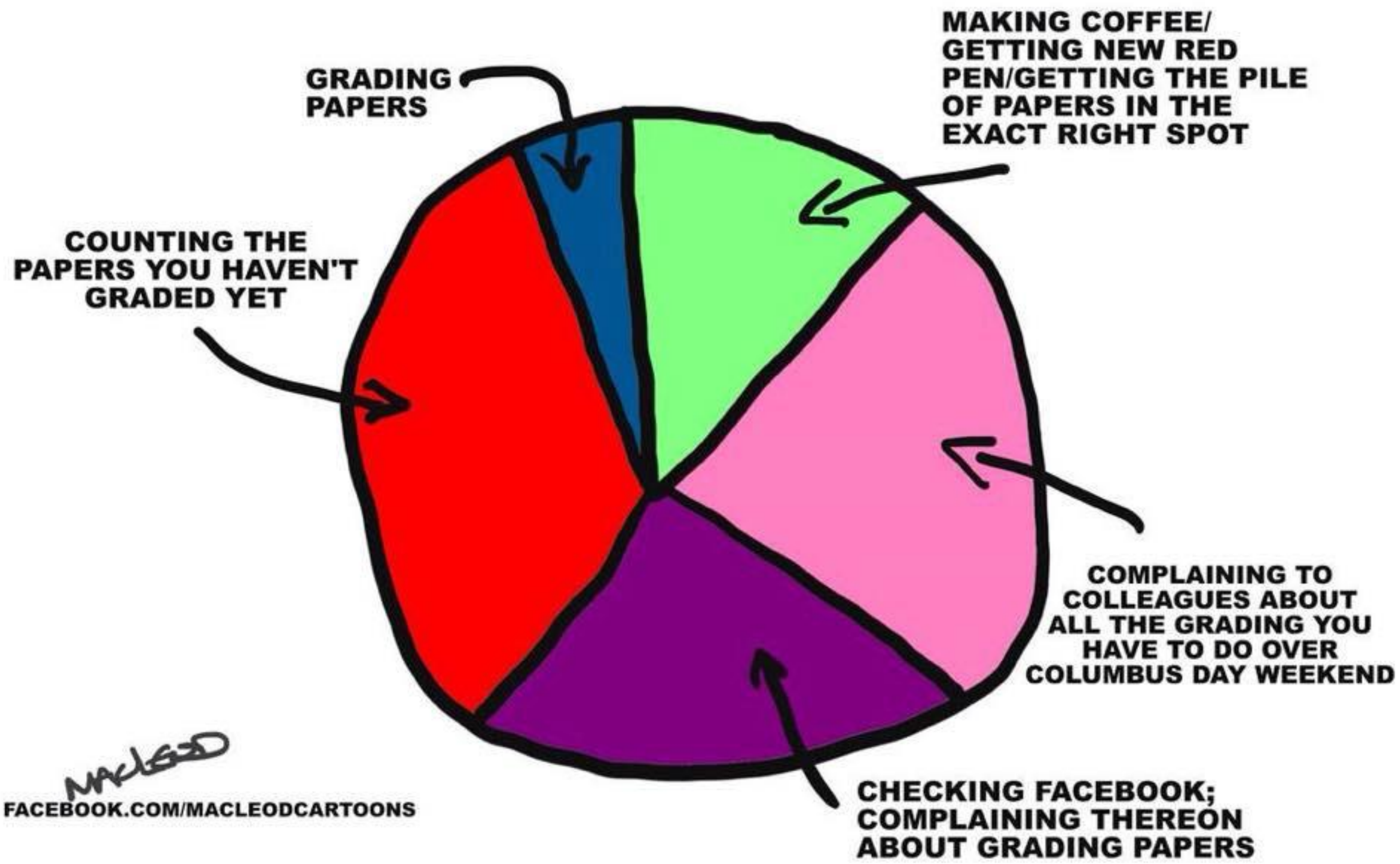


Method: Enter

Want to Learn More?

- Leys, C., Delacre, M., Mora, Y. L., Lakens, D., & Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. *International Review of Social Psychology, 32*(1). doi:10.5334/irsp.289
- Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology, 74*, 150-156. doi:10.1016/j.jesp.2017.09.011
- Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764-766. doi:10.1016/j.jesp.2013.03.013

ANALYSIS OF FACULTY TIME USE WHILE "GRADING PAPERS"



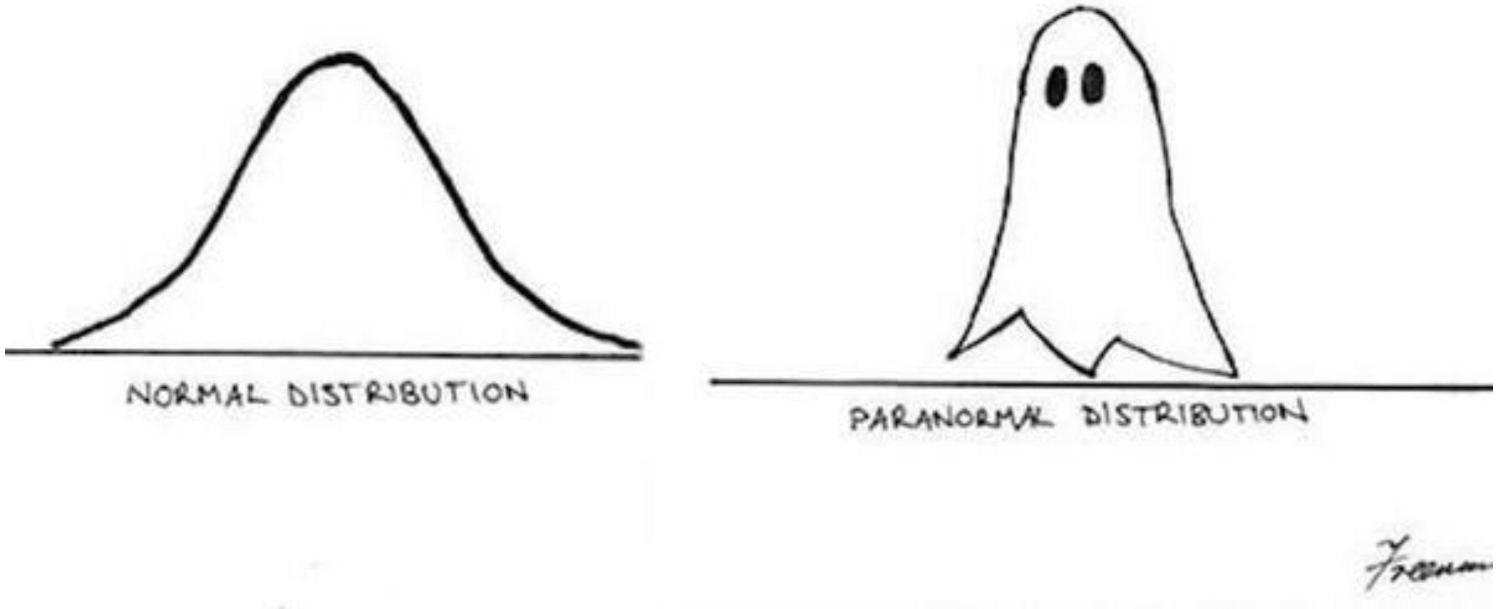
GRADING PAPERS
MAKING COFFEE/ GETTING NEW RED PEN/GETTING THE PILE OF PAPERS IN THE EXACT RIGHT SPOT
COUNTING THE PAPERS YOU HAVEN'T GRADED YET
COMPLAINING TO COLLEAGUES ABOUT ALL THE GRADING YOU HAVE TO DO OVER COLUMBUS DAY WEEKEND
CHECKING FACEBOOK; COMPLAINING THEREON ABOUT GRADING PAPERS
MACLEOD
FACEBOOK.COM/MACLEODCARTOONS

Outline for Today

- Missing Data
  - Identifying, assessing type, imputation options
- Composite Scores
  - Total scores, recoding, dummy coding
- Outliers
  - Identifying and addressing univariate and multivariate outliers
- **Normality**
  - **Assessing and addressing (e.g., transformations, analysis specifications)**
- Bivariate Linearity
  - Reading scatterplots and what to do about them
- Documentation
  - The importance of codebooks and data logs

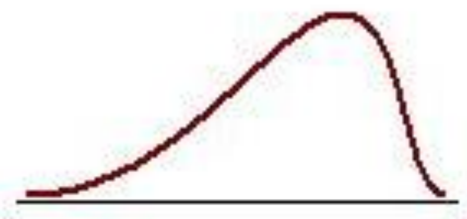
Normality

- How to assess/detect it
- What to do if you have non-normal data

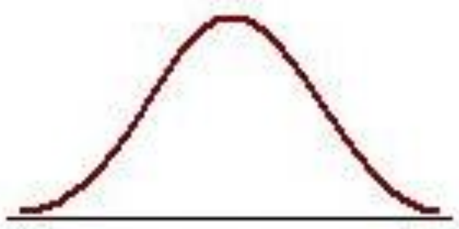


Assessing Normality

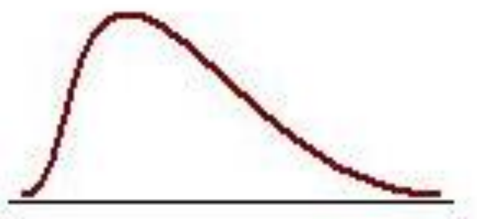
- Skewness and Kurtosis



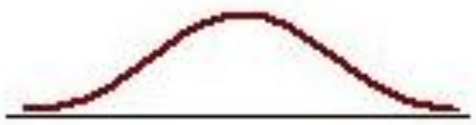
Negatively
Skewed



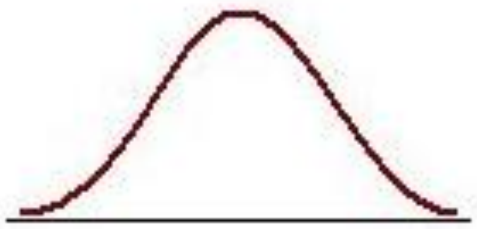
Normal



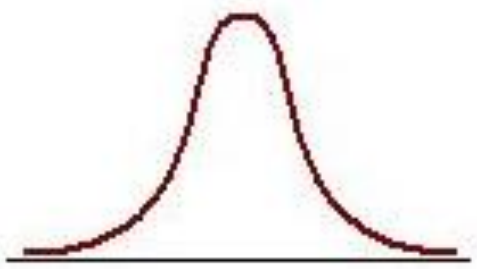
Positively
Skewed



Platykurtic



Normal

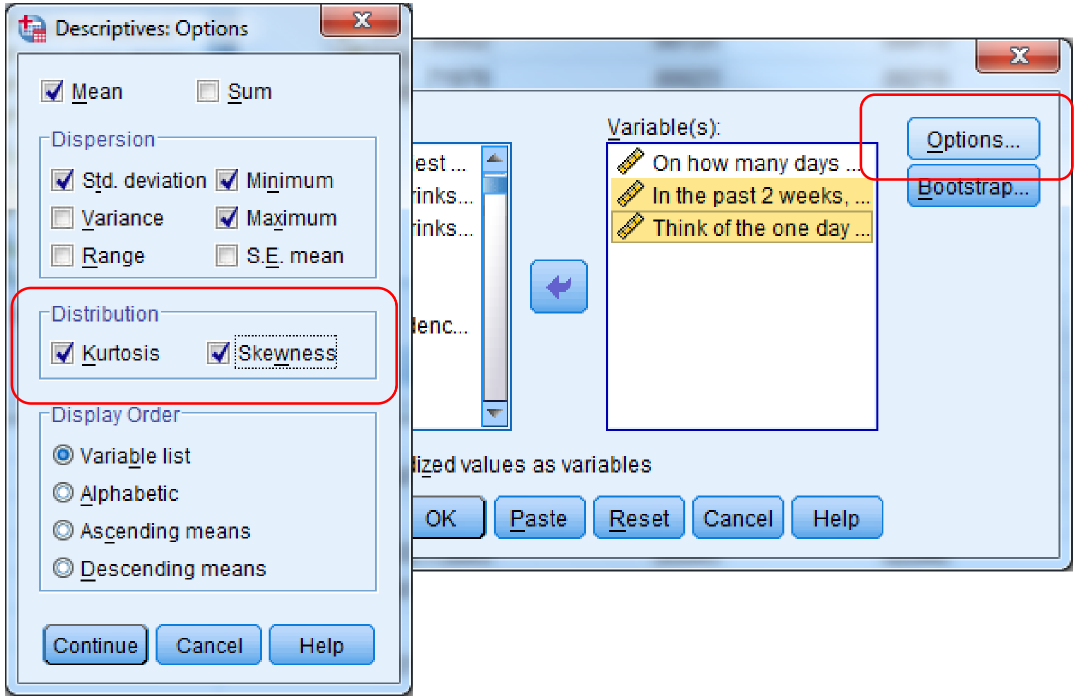


Leptokurtic

Assessing Normality

- Skewness is often more rigid for analyses than kurtosis
- Assess via requesting skewness and kurtosis for each variable
  - Analyze > Descriptive Statistics > Descriptives

- How much is too much?
  - Skewness:
    - Some say 2, some say 3
    - Use your best judgment
    - Matching skew for predictors and outcomes might be less bad
  - Kurtosis: Can be much higher (wouldn't bat an eye at 5)



Descriptives: Options

☑ Mean □ Sum

Dispersion

☑ Std. deviation ☑ Minimum

□ Variance ☑ Maximum

□ Range □ S.E. mean

Distribution

☑ Kurtosis ☑ Skewness

Display Order

● Variable list

○ Alphabetic

○ Ascending means

○ Descending means

Variable(s):

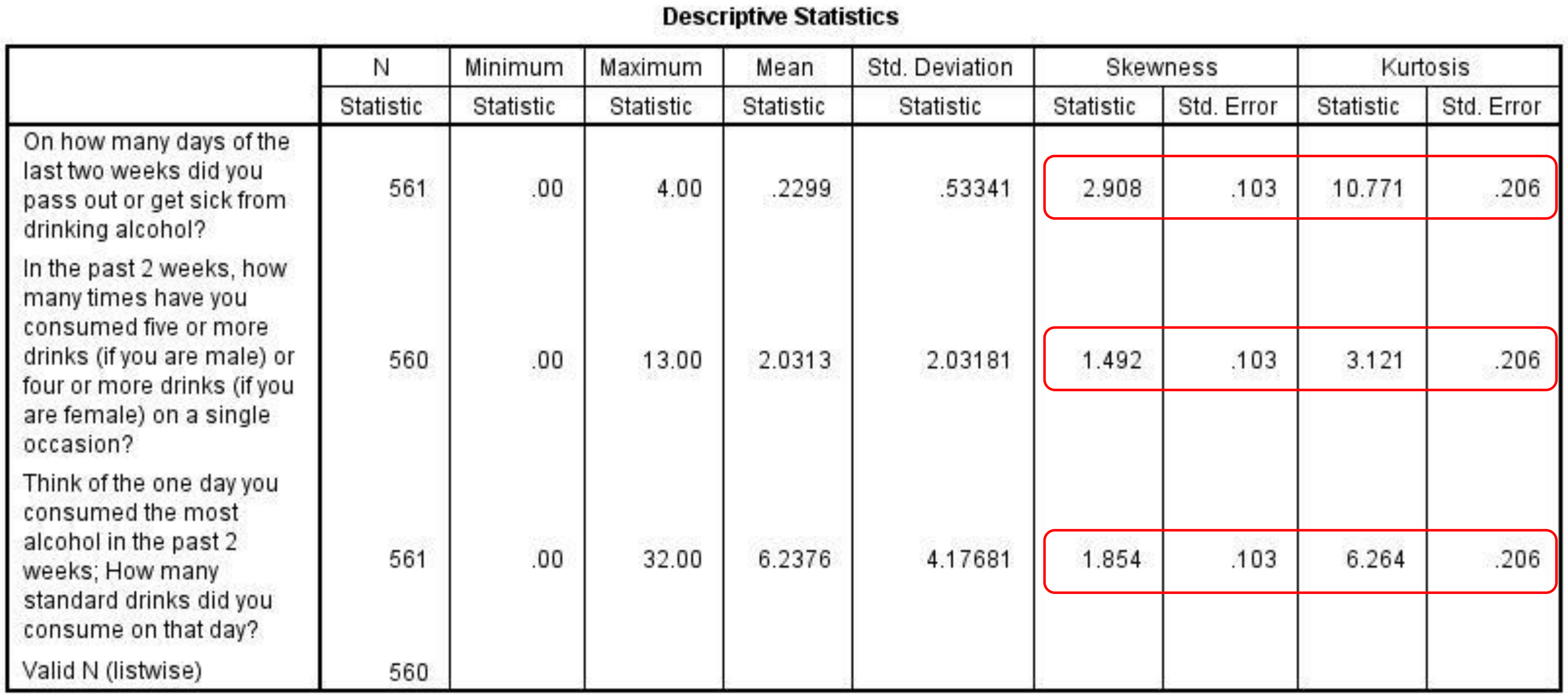
On how many days ...

In the past 2 weeks, ...

Think of the one day ...

Assessing Normality

- Binge frequency looks okay
    - Kurtosis is over 3, but that’s not terrible
- Max drinks is surprisingly okay for skew
    - Kurtosis is a little strong
- Days passed out/sick is bad

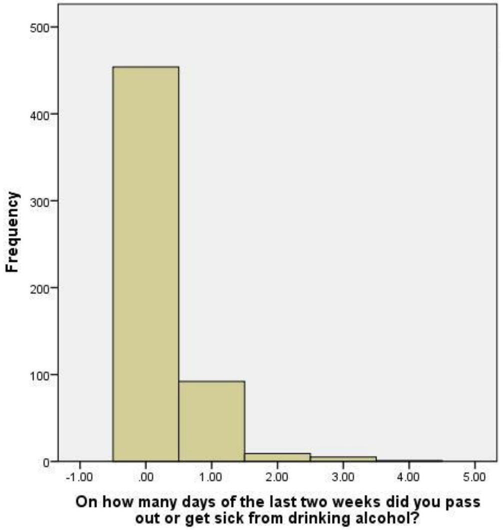


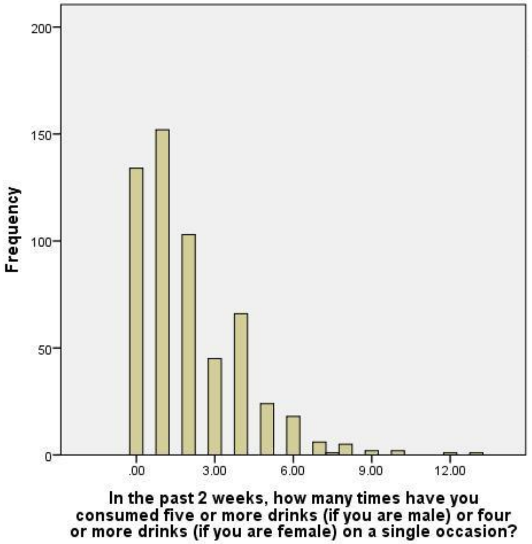
Descriptive Statistics

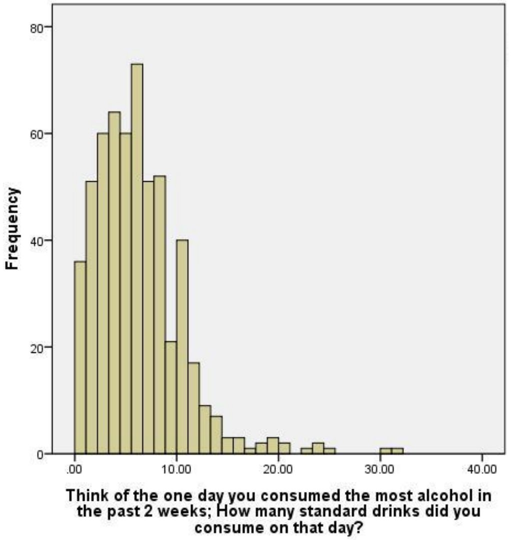
	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
On how many days of the last two weeks did you pass out or get sick from drinking alcohol?	561	.00	4.00	.2299	.53341	2.908	.103	10.771	.206
In the past 2 weeks, how many times have you consumed five or more drinks (if you are male) or four or more drinks (if you are female) on a single occasion?	560	.00	13.00	2.0313	2.03181	1.492	.103	3.121	.206
Think of the one day you consumed the most alcohol in the past 2 weeks; How many standard drinks did you consume on that day?	561	.00	32.00	6.2376	4.17681	1.854	.103	6.264	.206
Valid N (listwise)	560								

Assessing Normality

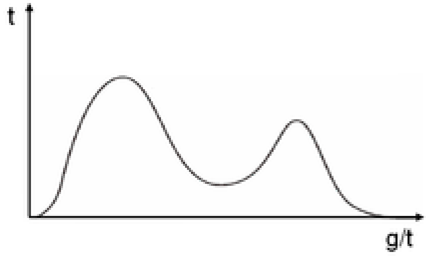
- Enough?
  - Needs to be Unimodal!
  - Check with Histograms (don’t overlay normal curve)
    - Graphs > Legacy Dialogs > Historam







- Uniform (flat) is not terrible
- Don’t want bi/tri/quadmodal



Non-Normality

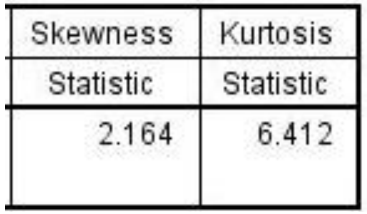
- What do you do if you find non-normality?
- Make sure you already addressed outliers
  - If not, winsorizing outliers may remove the problem
- Transform data
  - Particularly helpful for addressing skewness and kurtosis
    - Most transformations help both simultaneously
  - Most common solution
- Specify a different type of analysis
  - Can specify Poisson distribution
  - Logistic regression
  - Zero-inflated Poisson
  - Each type could be its own workshop or class
    - Point you in the right direction

Non-Normality: Transformation

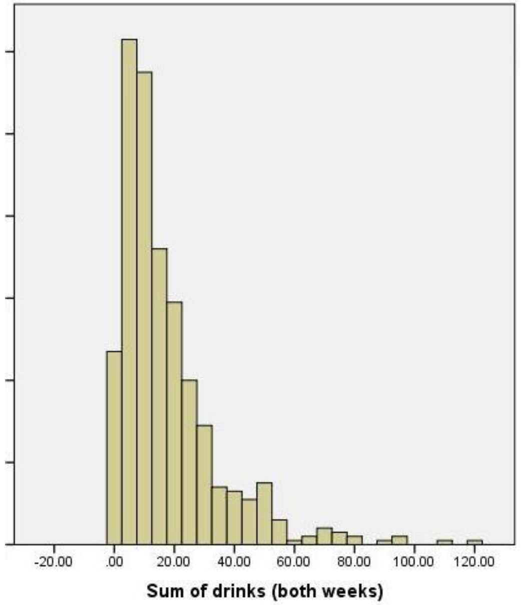
- Transform the offending variables to change their distribution
- Suggestions
- Positively skewed:
  - Log transform
    - Cannot log zero
    - SPSS: *NewVar* = Ln(*OldVar*)
  - Square root : *NewVar* = sqrt(*OldVar*)
- Negatively skewed:
  - Raise the power (** in SPSS)
  - $X^{1.5}$: *NewVar* = *OldVar***1.5
  - X^2 : *NewVar* = *OldVar***2
  - X^3 : *NewVar* = *OldVar***3
- Affects the interpretation of coefficients
  - To get back to raw metric, need to reverse the transformation
  - Reverse of log is exponentiation (e^X)
  - Reverse of square root is to square

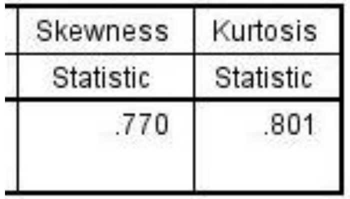
Non-Normality: Transformation

- Sum of drinks past 2 weeks: positively skewed
- Original square root log

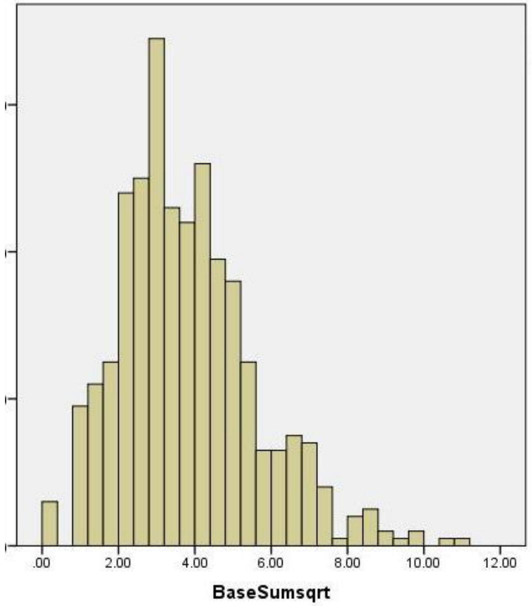


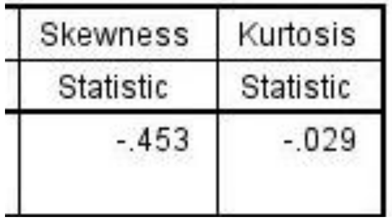
Skewness	Kurtosis
Statistic	Statistic
2.164	6.412



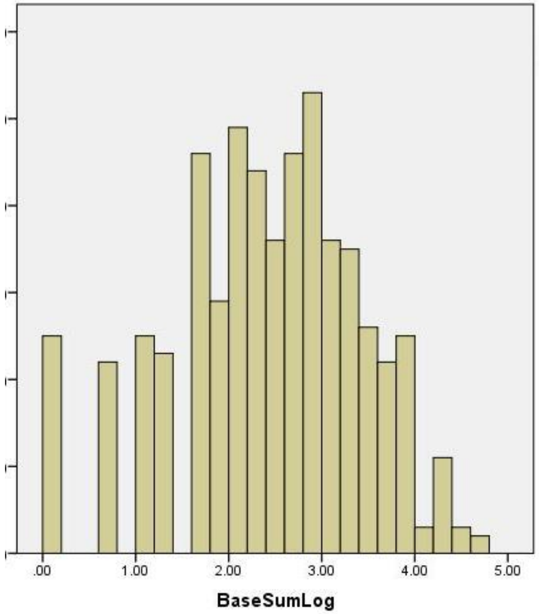


Skewness	Kurtosis
Statistic	Statistic
.770	.801



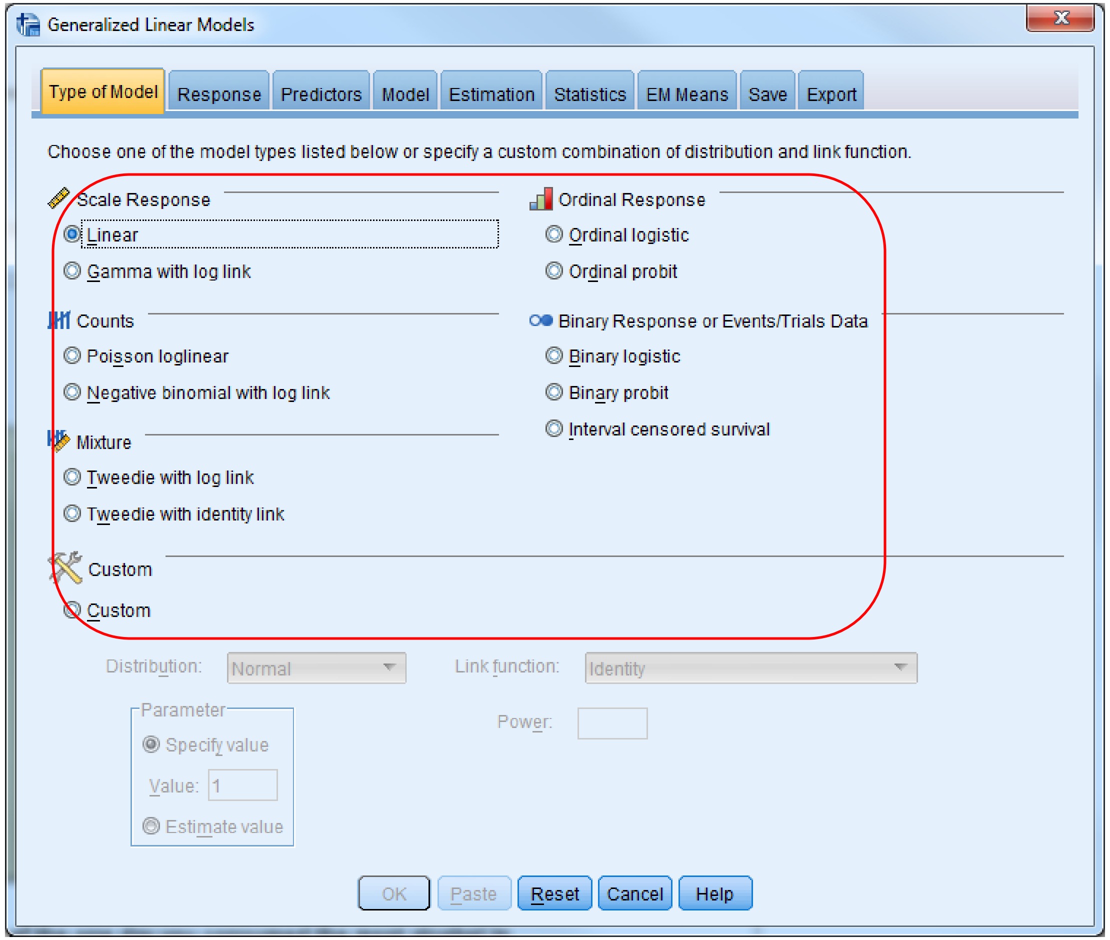


Skewness	Kurtosis
Statistic	Statistic
-.453	-.029



Non-Normality: Analysis

- Poisson
  - “COUNT” in Mplus
  - Button in HLM
  - Choose “Generalized Linear Model” in SPSS
    - Analyze
    - Generalized Linear Model
    - “d” is crucial
  - Choose the appropriate link function



Choose one of the model types listed below or specify a custom combination of distribution and link function.

Scale Response
- ● Linear
- ○ Gamma with log link

Counts
- ○ Poisson loglinear
- ○ Negative binomial with log link

Mixture
- ○ Tweedie with log link
- ○ Tweedie with identity link

Ordinal Response
- ○ Ordinal logistic
- ○ Ordinal probit

Binary Response or Events/Trials Data
- ○ Binary logistic
- ○ Binary probit
- ○ Interval censored survival

Custom
- ○ Custom

Distribution: Normal

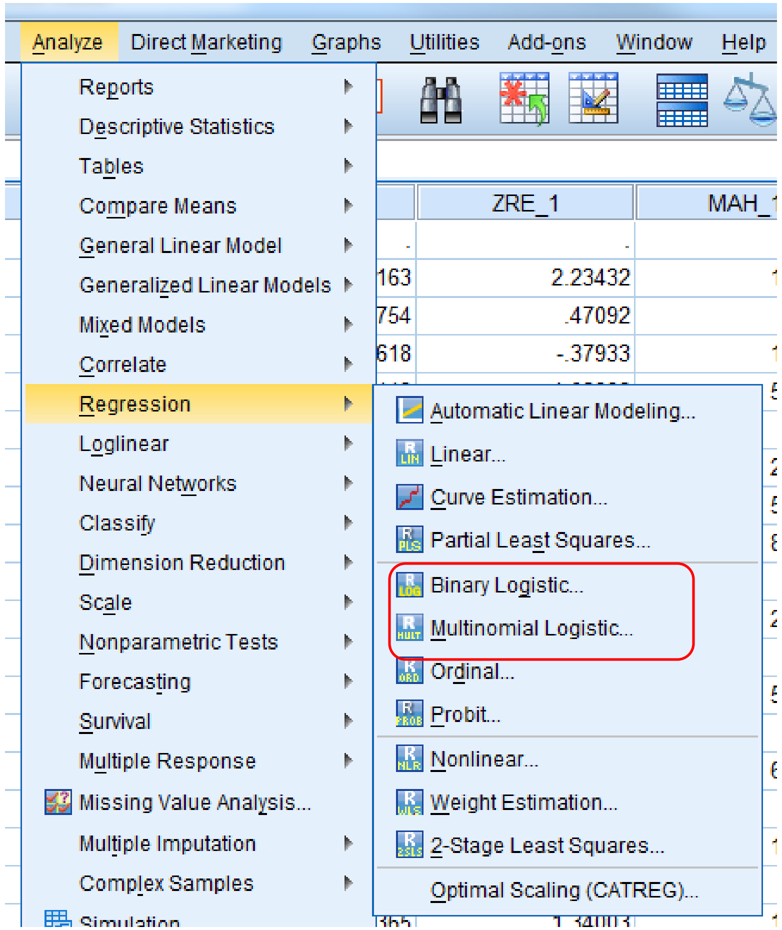
Link function: Identity

Parameter
- ● Specify value
- Value: 1
- ○ Estimate value

Power:

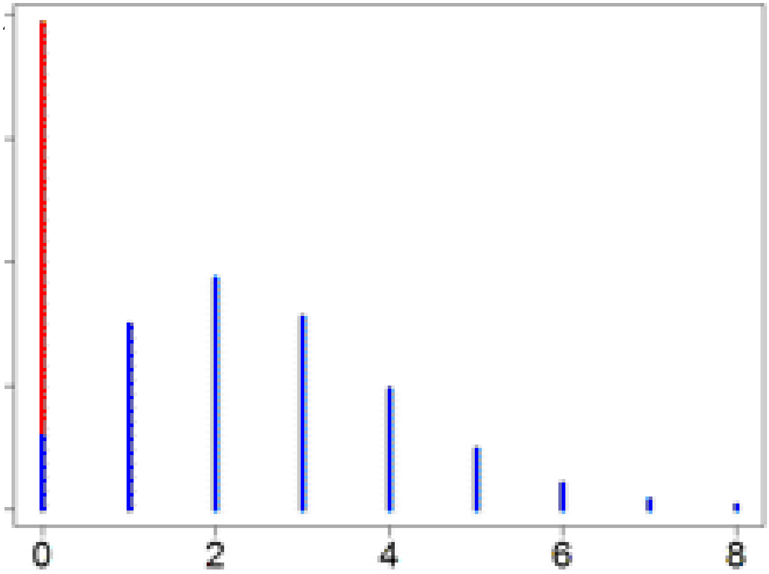
Non-Normality: Analysis

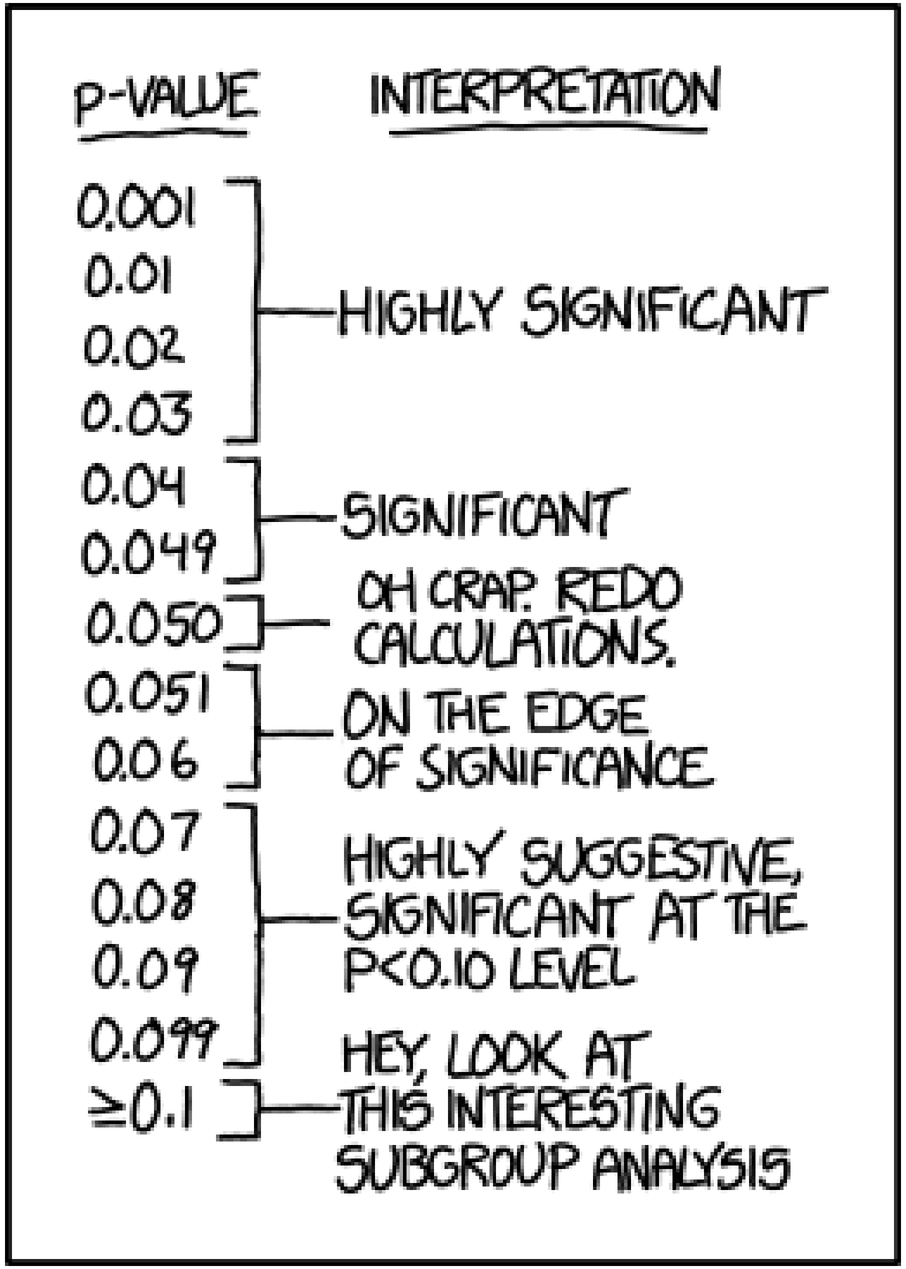
- Logistic regression
- Binary: Predict occurrence of an outcome
  - Probability of yes (compared to no)
    - Diagnosis?
    - Relapse?
    - Occurrence of violence?
  - If not naturally dichotomous, dummy code outcome variable
    - RECODE values >0 into 1
- Multinomial: Predict membership across a small number of groups
  - Probability of being medium as opposed to low, or high compared to low
  - Probability of being divorced compared to married, or single compared to married



Non-Normality: Analysis

- Zero-inflated Poisson
  - Relevant if you have a Poisson distribution, combined with a very high number of zeroes
    - Blue = Poisson; Red = zero inflation
  - Two simultaneous analyses
    - 1: probability of being a yes compared to no
    - 2: If yes,




P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

Outline for Today

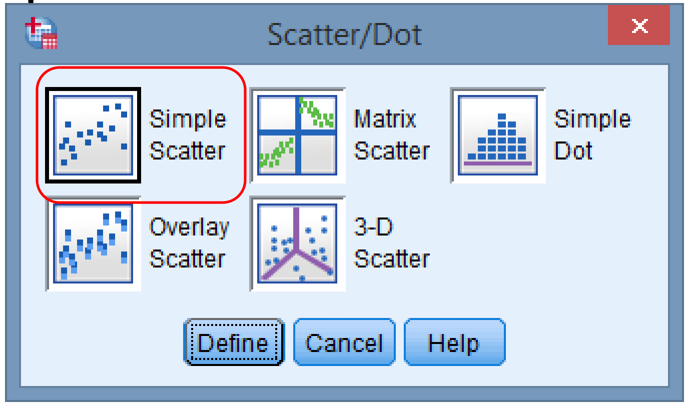
- Missing Data
  - Identifying, assessing type, imputation options
- Composite Scores
  - Total scores, recoding, dummy coding
- Outliers
  - Identifying and addressing univariate and multivariate outliers
- Normality
  - Assessing and addressing (e.g., transformations, analysis specifications)
- **Bivariate Linearity**
  - **Reading scatterplots and what to do about them**
- Documentation
  - The importance of codebooks and data logs

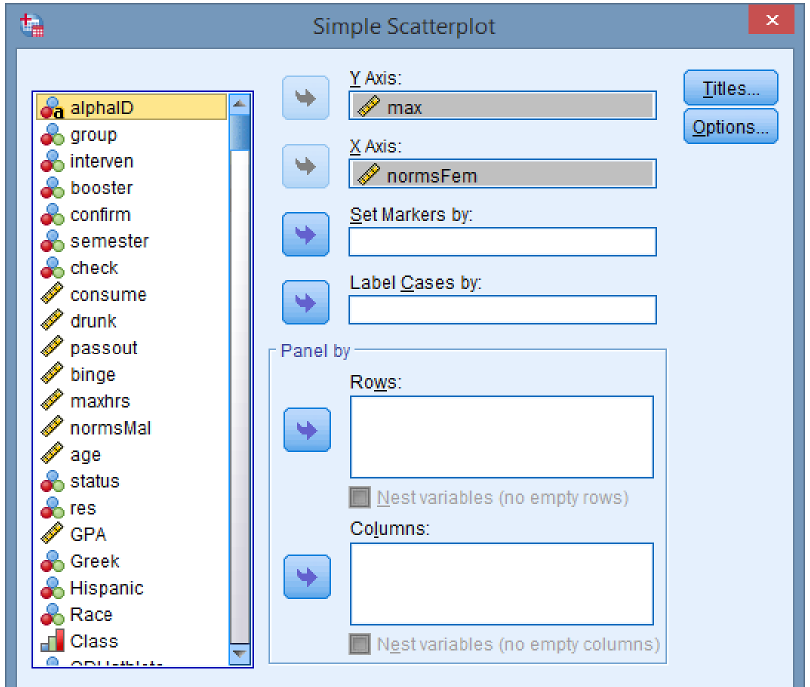
Bivariate Linearity

- How to assess
- What to do

Assessing Linearity

- Another assumption of regression and other linear approaches
  - Comparing how *X* changes with *Y*
- Asses it with scatterplots!
- Graphs > Legacy Dialogs > Scatter/dot
  - Simple Scatter





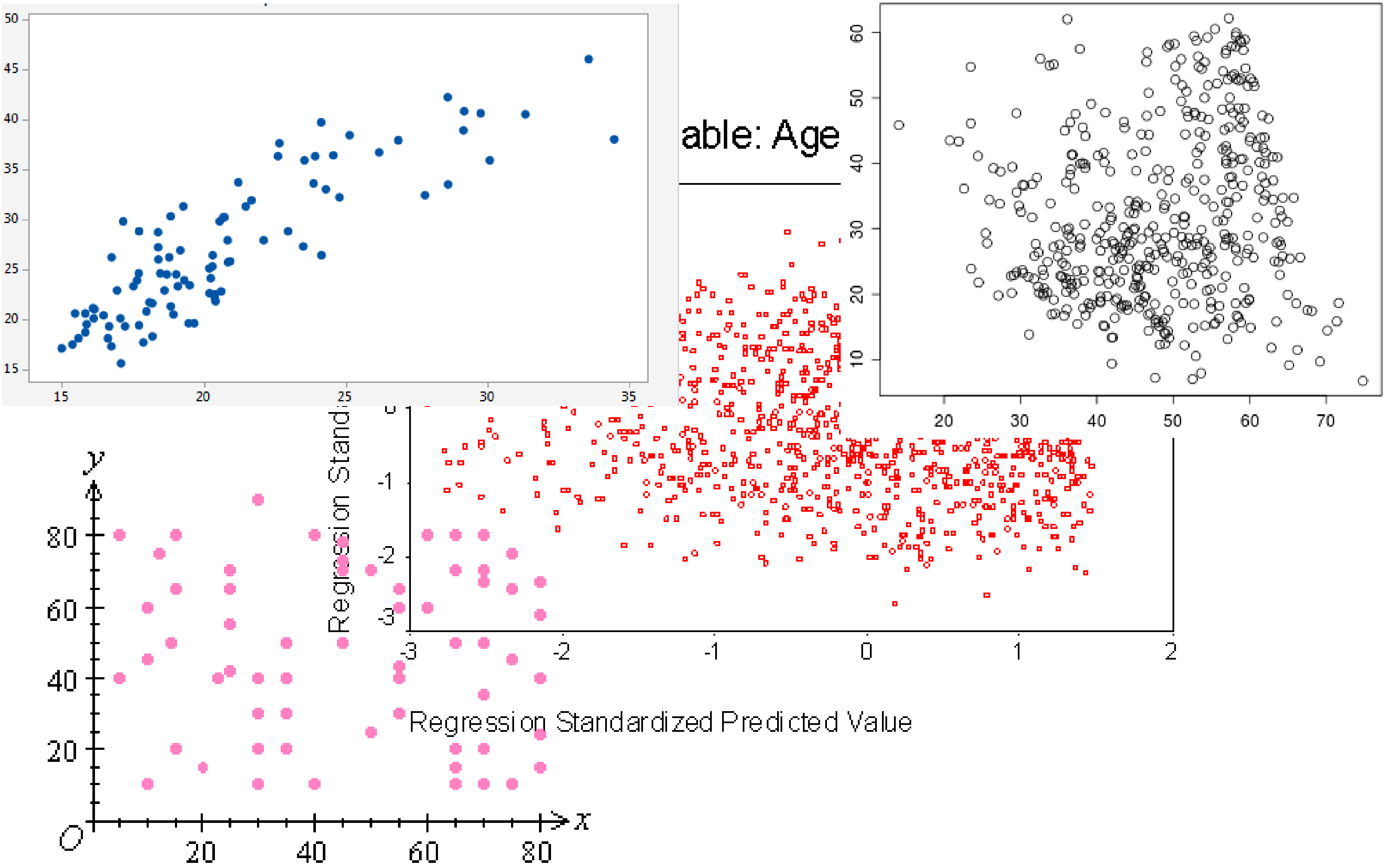
Y Axis: max

X Axis: normsFem

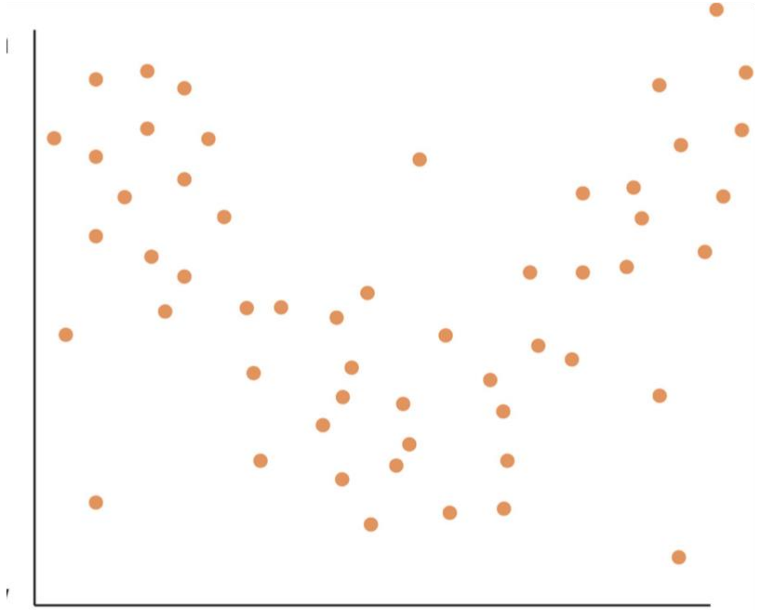
Assessing Linearity

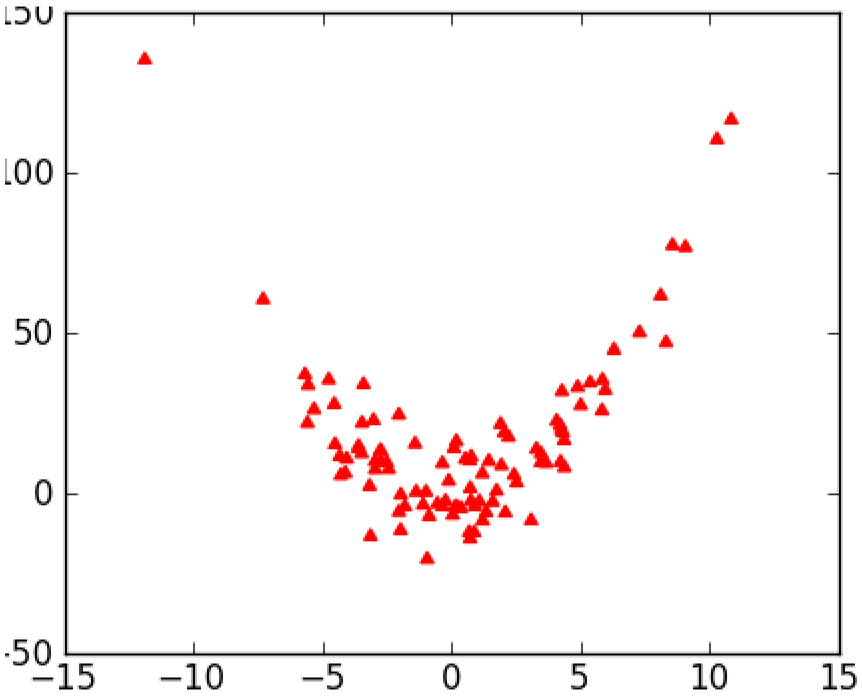
- What’s normal?
  - Line
    - Most obvious
  - Football/ellipse
    - More likely
  - True randomness
    - Pretty common among less correlated data
- What’s bad?
  - Curvilinear
  - Frown or smile
  - Convex or concave

Assessing Linearity



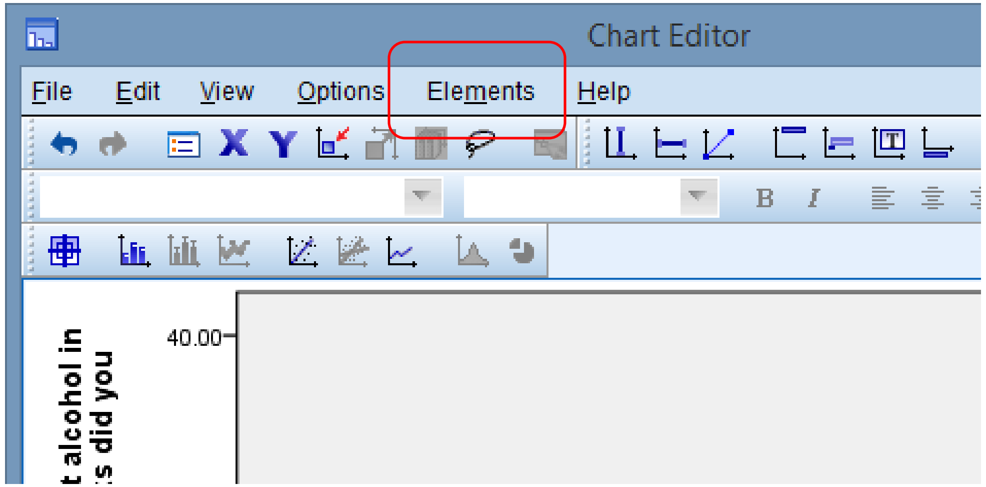
Assessing Linearity

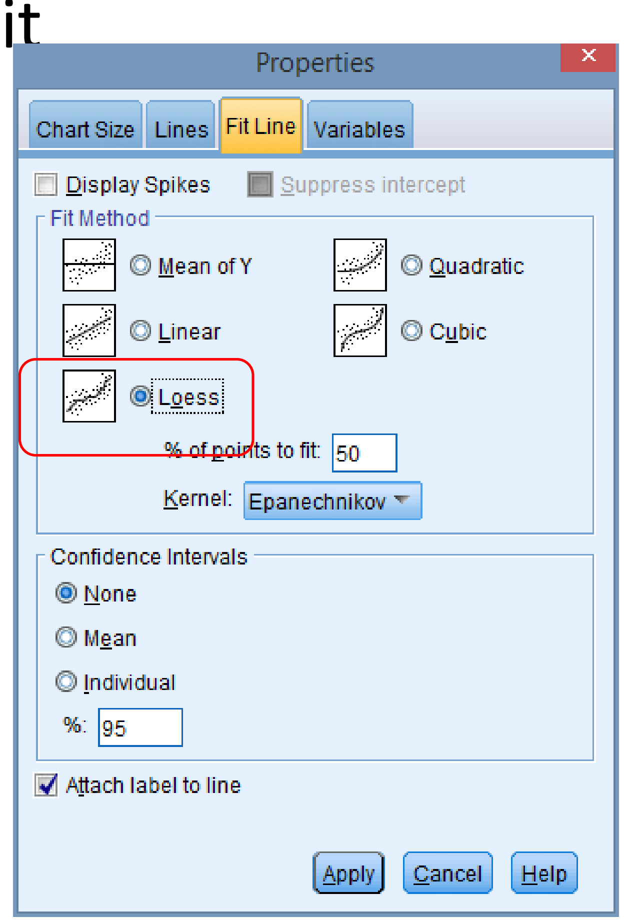




Assessing Linearity

- Having trouble? Use Lowess (Loess) lines!
- Double-click the Scatterplot to Activate it
- Elements > Fit Line at Total





□ Display Spikes ▨ Suppress intercept

Fit Method

○ Mean of Y ○ Quadratic

○ Linear ○ Cubic

● Loess

% of points to fit: 50

Kernel: Epanechnikov

Confidence Intervals

● None

○ Mean

○ Individual

%: 95

☑ Attach label to line

Assessing Linearity

- A curvy Loess line indicates a problem
- Mild bumps are okay
- Be on the lookout for smile or frown
- U = bad
- ∩ = bad

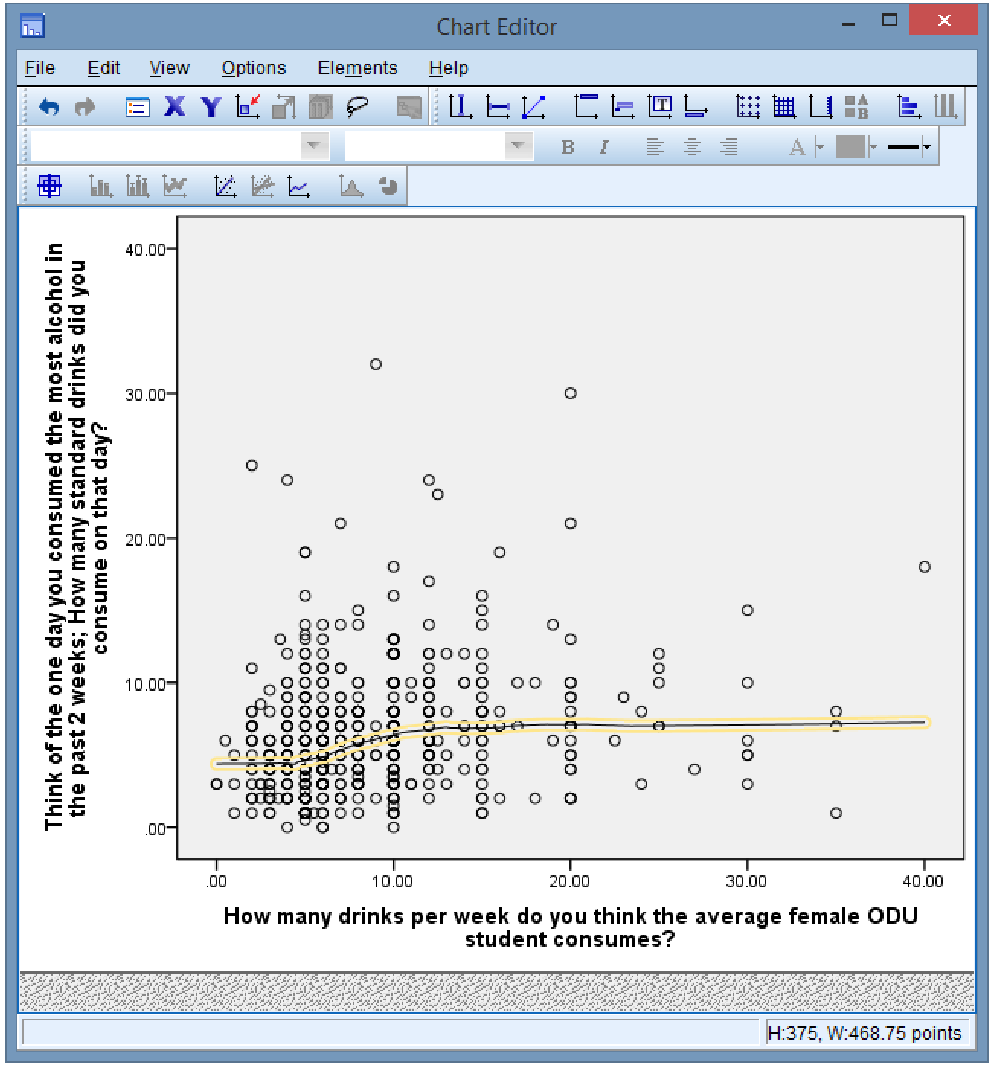
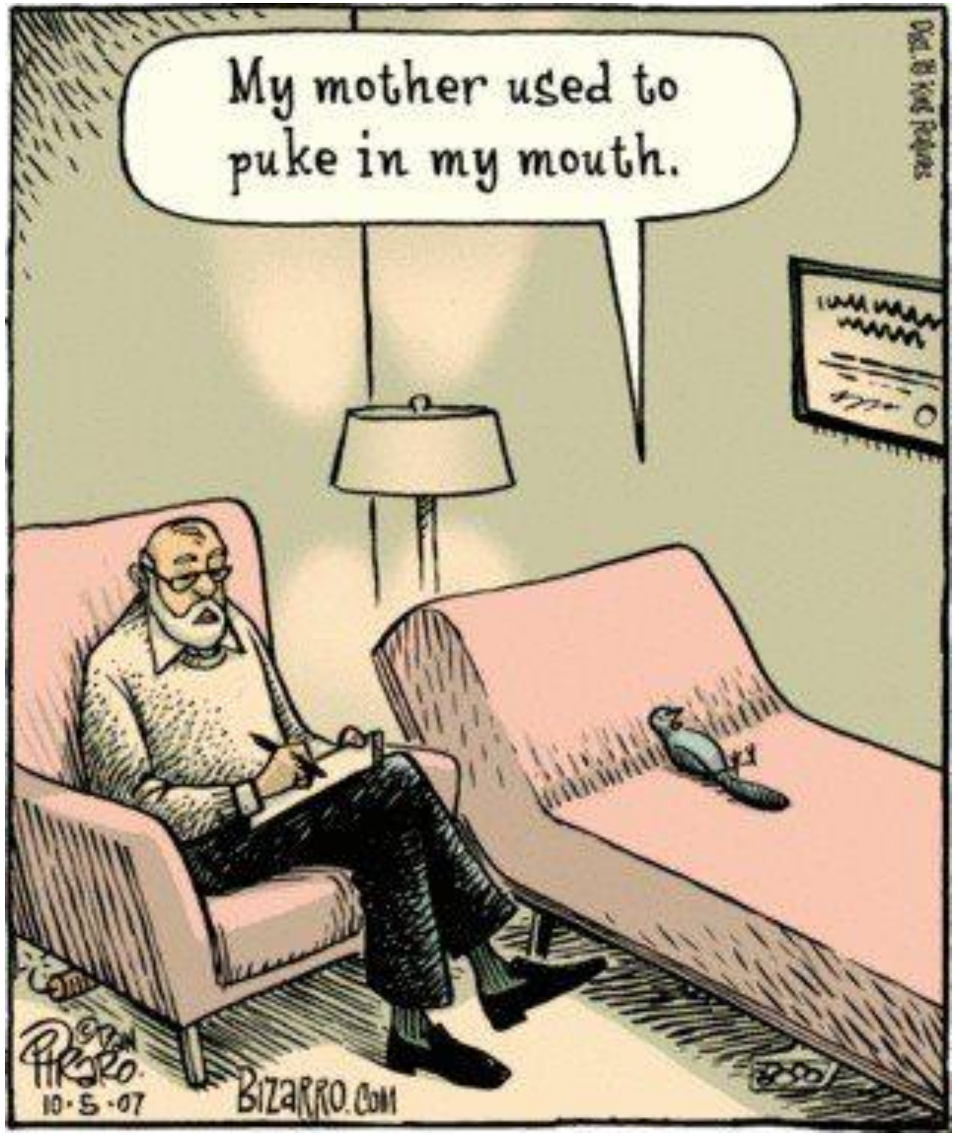


Chart Editor
File
Edit
View
Options
Elements
Help
Think of the one day you consumed the most alcohol in the past 2 weeks; How many standard drinks did you consume on that day?
40.00
30.00
20.00
10.00
.00
.00
10.00
20.00
30.00
40.00
How many drinks per week do you think the average female ODU student consumes?
H:375, W:468.75 points

Addressing Nonlinearity

- What if your data are curvilinear?
- Split your data
  - For younger… for older…
- Transformations might help
  - Not for extreme curvilinearity
- Include polynomial predictors (regression)
  - X^2, X^3
  - Linear AND quadratic relationships


My mother used to puke in my mouth.
©2007 Dan Piraro
10-5-07
BIZARRO.COM
Dist. by King Features

Outline for Today

- Missing Data
  - Identifying, assessing type, imputation options
- Composite Scores
  - Total scores, recoding, dummy coding
- Outliers
  - Identifying and addressing univariate and multivariate outliers
- Normality
  - Assessing and addressing (e.g., transformations, analysis specifications)
- Bivariate Linearity
  - Reading scatterplots and what to do about them
- **Documentation**
  - **The importance of codebooks and data logs**

Documenting Your Activities

- Codebooks
- Datalogs
- Syntax for composites and recoding
  - This may happen multiple times
- Save multiple versions of dataset
  - Data.sav
  - Data_dummy.sav
  - Data_dummy_imputed.sav
  - Data_dum_imp_composites.sav
  - Data_dum_imp_com_nooutliers.sav
  - Data_dum_imp_com_noout_small.sav
  - Data_dum_imp_com_noout_small_999.sav
  - Mplus.sav
- If lots of versions, can keep a version document

Documenting Your Activities

- Why?
- Committee members
  - Request edits
  - Readers may want more info
  - Want to ensure you’re a skilled scientist/researcher
- Journal Reviewers/Editors
  - “How many outliers?”
- Collaborators

Codebooks

- For each scale:
  - Citations and references
  - Response options from survey (note if different from validating article)
  - Instructions provided to participants (note if different)
  - Coding schemes
    - Subscales? Is a total score appropriate? Reverse scoring? Other recoding?
    - Means or sums?
  - If scoring is unusual (e.g., DDQ), note all possible values created (e.g., quantity, frequency, binge, peak, BAC, etc., with definitions of each)
  - Note any deviations from original/validated scale
    - E.g., did not include subjective evaluation items from CEOA
    - E.g., changed response options for SQ).
    - Note why. Provided additional citation/reference if relevant

Codebooks

- For each item:
    - Variable names (short SPSS version), included recoded versions
    - Variable label (longer description of variable)
        - Item text, or description of recode, etc.
    - Response scale
    - Any comments
        - Could indicate if item was recoded into something else, or is a recoded version of an earlier variable
        - Could include “DO NOT USE” if variable is too skewed or otherwise inappropriate for analysis
    - New variables should be new entries
        - Total scores or subscales
        - Transformed variables, etc.

Codebooks

- If longitudinal
    - Timeframe for each construct
    - At which time points each construct was assessed
        - Baseline only versus follow-ups only versus all

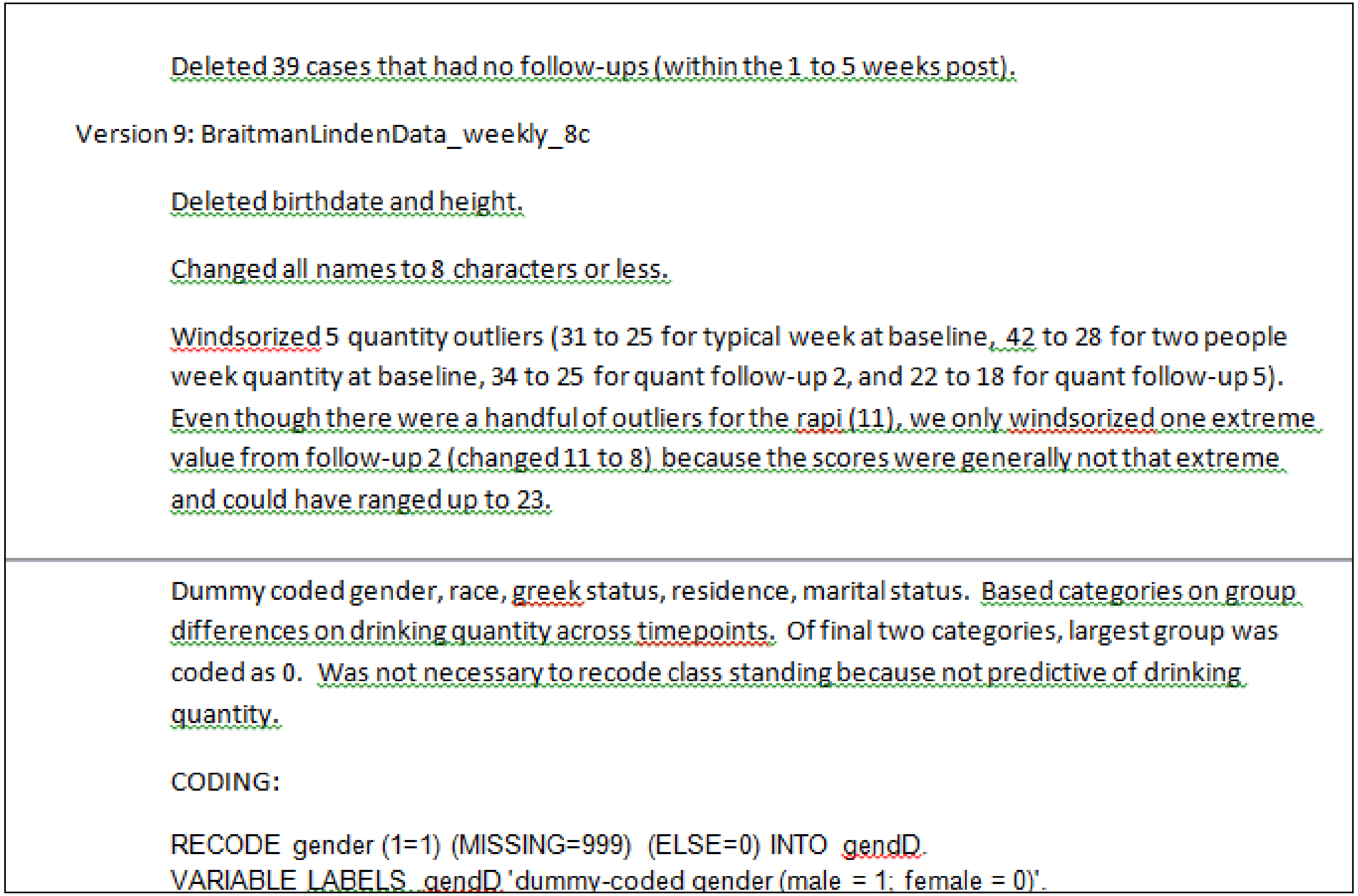
Codebook Examples

- See files

Datalogs

- Date can be helpful, but not required
  - Who did it an be very helpful if there are multiple hands in the project
- ACTIVITIES are required, with specific details
  - What you did
  - Why you did it
- Missing data
  - What percentage for data? Each variable?
  - How did you address it?
- Outliers
  - How many for each variable?
  - Old and new values?
- Recoding
  - What dummy codes did you create? Why?
  - What composites scores did you create? Means or sums or something else?
    - Did you remember to reverse score??
- Did you check linearity? Normality?
  - Confirmed for which variables?
  - What adjustments were made for which variables (if any)?

Datalog Example



Deleted 39 cases that had no follow-ups (within the 1 to 5 weeks post).

Version 9: BraitmanLindenData_weekly_8c

Deleted birthdate and height.

Changed all names to 8 characters or less.

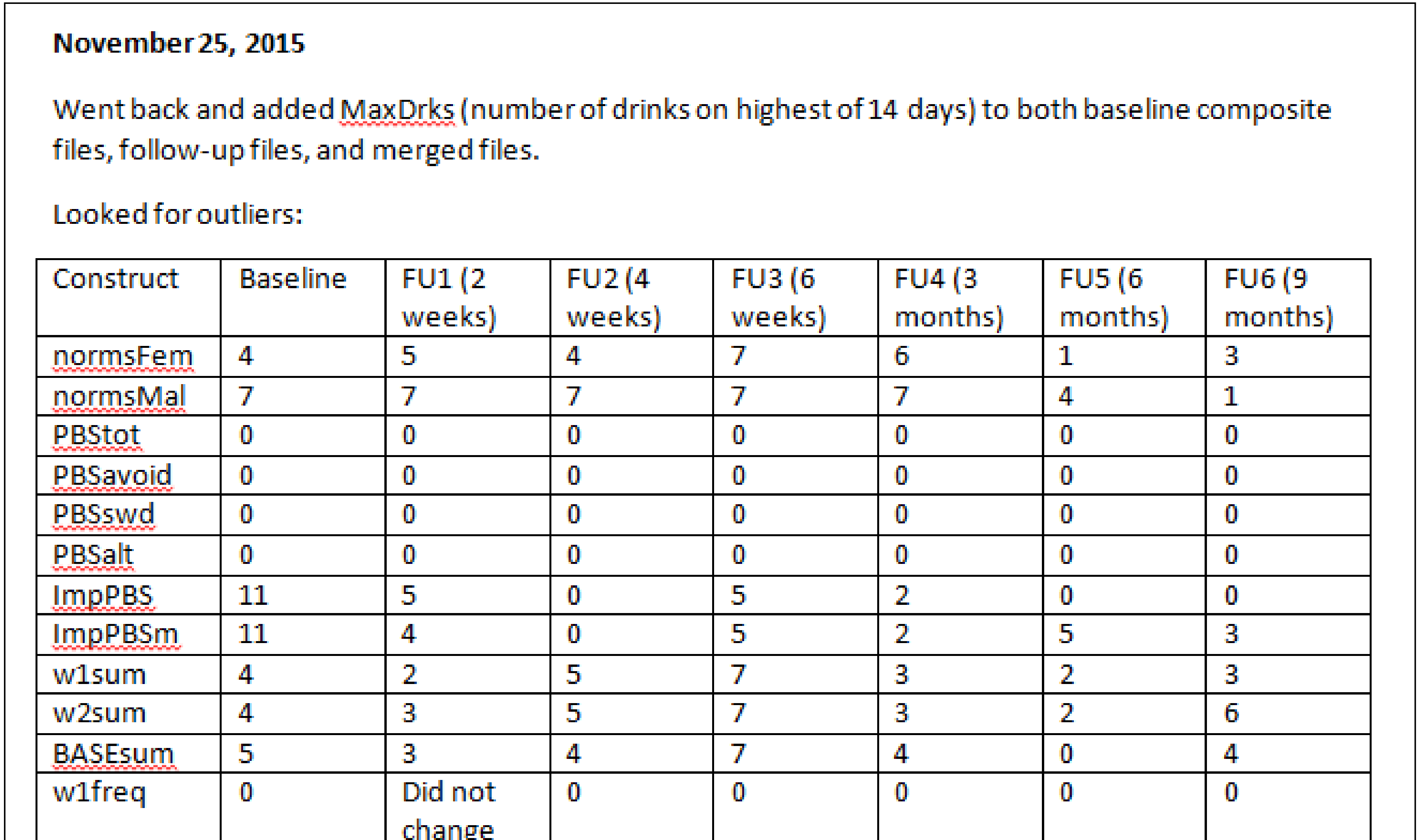
Windsorized 5 quantity outliers (31 to 25 for typical week at baseline, 42 to 28 for two people week quantity at baseline, 34 to 25 for quant follow-up 2, and 22 to 18 for quant follow-up 5). Even though there were a handful of outliers for the rapi (11), we only windsorized one extreme value from follow-up 2 (changed 11 to 8) because the scores were generally not that extreme and could have ranged up to 23.

Dummy coded gender, race, greek status, residence, marital status. Based categories on group differences on drinking quantity across timepoints. Of final two categories, largest group was coded as 0. Was not necessary to recode class standing because not predictive of drinking quantity.

CODING:

```
RECODE gender (1=1) (MISSING=999) (ELSE=0) INTO gendD.
VARIABLE LABELS gendD 'dummy-coded gender (male = 1; female = 0)'.
```

Datalog Example



November 25, 2015

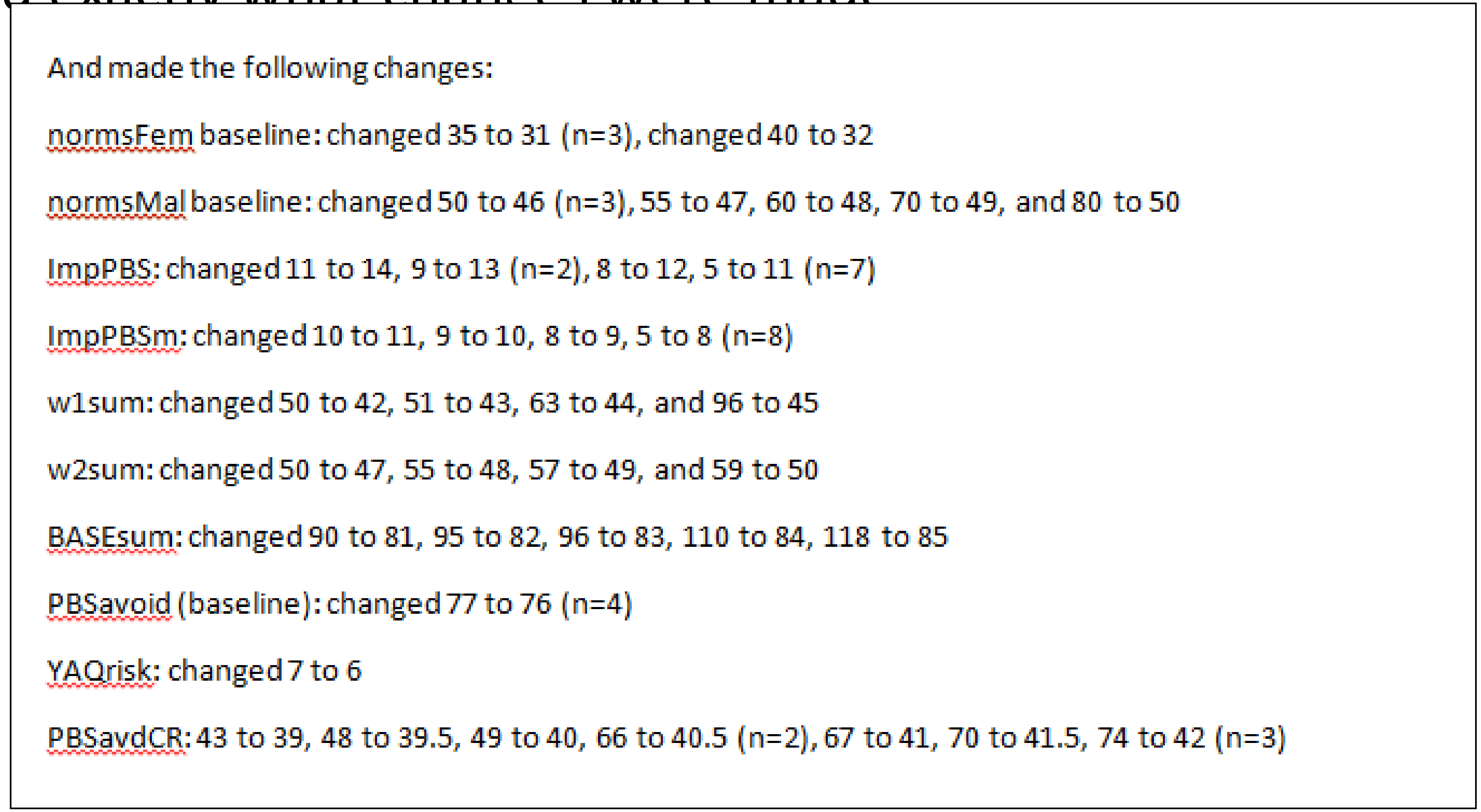
Went back and added MaxDrks (number of drinks on highest of 14 days) to both baseline composite files, follow-up files, and merged files.

Looked for outliers:

Construct	Baseline	FU1 (2 weeks)	FU2 (4 weeks)	FU3 (6 weeks)	FU4 (3 months)	FU5 (6 months)	FU6 (9 months)
normsFem	4	5	4	7	6	1	3
normsMal	7	7	7	7	7	4	1
PBStot	0	0	0	0	0	0	0
PBSavoid	0	0	0	0	0	0	0
PBSswd	0	0	0	0	0	0	0
PBSalt	0	0	0	0	0	0	0
ImpPBS	11	5	0	5	2	0	0
ImpPBSm	11	4	0	5	2	5	3
w1sum	4	2	5	7	3	2	3
w2sum	4	3	5	7	3	2	6
BASEsum	5	3	4	7	4	0	4
w1freq	0	Did not change	0	0	0	0	0

Datalog Example

- Noted exactly what changes were made



And made the following changes:

normsFem baseline: changed 35 to 31 (n=3), changed 40 to 32

normsMal baseline: changed 50 to 46 (n=3), 55 to 47, 60 to 48, 70 to 49, and 80 to 50

ImpPBS: changed 11 to 14, 9 to 13 (n=2), 8 to 12, 5 to 11 (n=7)

ImpPBSm: changed 10 to 11, 9 to 10, 8 to 9, 5 to 8 (n=8)

w1sum: changed 50 to 42, 51 to 43, 63 to 44, and 96 to 45

w2sum: changed 50 to 47, 55 to 48, 57 to 49, and 59 to 50

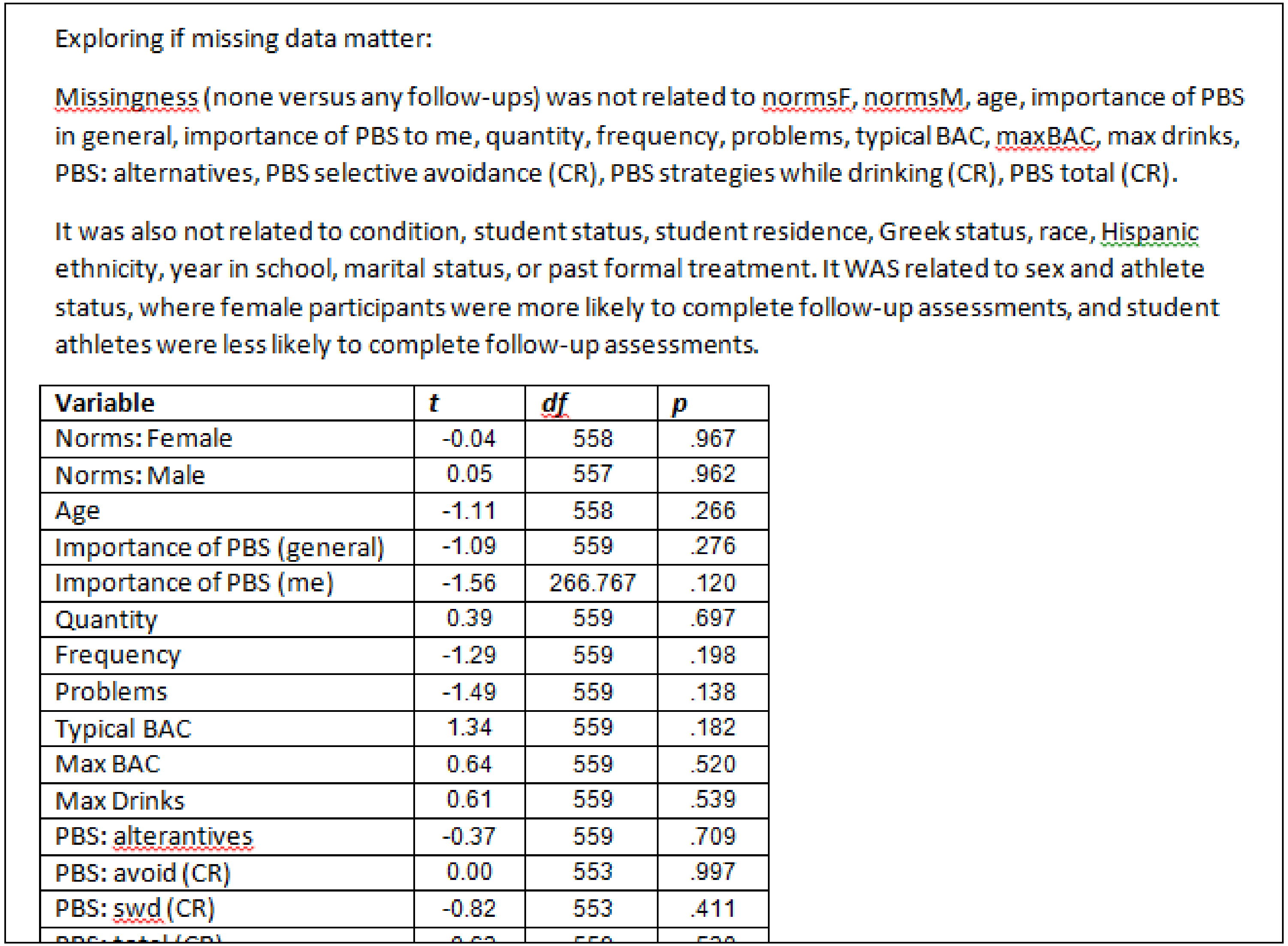
BASEsum: changed 90 to 81, 95 to 82, 96 to 83, 110 to 84, 118 to 85

PBSavoid (baseline): changed 77 to 76 (n=4)

YAQrisk: changed 7 to 6

PBSavdCR: 43 to 39, 48 to 39.5, 49 to 40, 66 to 40.5 (n=2), 67 to 41, 70 to 41.5, 74 to 42 (n=3)

Datalog Example



Exploring if missing data matter:

Missingness (none versus any follow-ups) was not related to normsF, normsM, age, importance of PBS in general, importance of PBS to me, quantity, frequency, problems, typical BAC, maxBAC, max drinks, PBS: alternatives, PBS selective avoidance (CR), PBS strategies while drinking (CR), PBS total (CR).

It was also not related to condition, student status, student residence, Greek status, race, Hispanic ethnicity, year in school, marital status, or past formal treatment. It WAS related to sex and athlete status, where female participants were more likely to complete follow-up assessments, and student athletes were less likely to complete follow-up assessments.

Variable	***t***	***df***	***p***
Norms: Female	-0.04	558	.967
Norms: Male	0.05	557	.962
Age	-1.11	558	.266
Importance of PBS (general)	-1.09	559	.276
Importance of PBS (me)	-1.56	266.767	.120
Quantity	0.39	559	.697
Frequency	-1.29	559	.198
Problems	-1.49	559	.138
Typical BAC	1.34	559	.182
Max BAC	0.64	559	.520
Max Drinks	0.61	559	.539
PBS: alterantives	-0.37	559	.709
PBS: avoid (CR)	0.00	553	.997
PBS: swd (CR)	-0.82	553	.411
PBS: total (CR)	-0.63	553	.530

Thank you! Questions?