Data screening using SPSS

Natalie Loxton

7 July 2008

Recommended Texts

- Tabachnick & Fidell (2007) Using multivariate statistics (5th ed). Chpt 4
- Field (2003) Discovering statistics using SPSS for windows. Chpt 2
- Cohen, Cohen, West & Aiken (2003) Applied multiple regression/correlation analysis for the behavioural sciences (3rd ed.). Chpt 4

Key assumptions in GLM

1. Normality (see handout)
2. Homogeneity of variance (depends on analysis)
3. Interval level data (need to check this BEFORE design questionnaire or collect data, e.g., ask for actual age rather than provide age groups)
4. Independence of observations (part of data collection, research design)

General tips before we begin….

- Coding books !!!
- SPSS file types:
  *.sav = data file
  *.spo = output file
- *.sps = syntax file
- Switch on “Display commands in log”
- Set page length to “infinite” to reduce overall output

Data entry (Between subjects)

ID	Status	Friends	Alc	Income	Neurot
1	Lecturer	5	10	20000	10
2	Lecturer	2	15	40000	17
3	Lecturer	0	20	35000	14
4	Lecturer	4	5	22000	13
5	Lecturer	1	30	50000	21
6	Student	10	25	5000	7
7	Student	12	20	100	13
8	Student	15	16	3000	9
9	Student	12	17	10000	14
10	Student	17	18	10	13

Data entry (Within subjects)

ID	Pre-treatment	Post-Treatment	6mth FU
1	30	20	22
2	25	24	24
3	15	10	5
4	20	18	20
5	32	30	30

Data entry (Mixed design)

Client	Group	Pre	Post	6 mth FU
1	Treatment	30	20	22
2	Treatment	25	24	24
3	Treatment	15	10	5
4	Treatment	20	18	20
5	Treatment	32	30	30
6	Control	25	25	25
7	Control	20	20	22
8	Control	16	15	10
9	Control	17	20	25
10	Control	18	18	20

Screening

- Ensure collect enough data
- Drop down menus versus syntax
- Assigning ID numbers to cases (after the fact)

```
compute id= $casenum.
Execute.
```

Open dataset: “missing data Tabachnick.sav”

Add ID numbers

Missing Data

- Check amount of missing data

  compute miss= nmiss (var1 – var K).

  Execute.

- Types of missing data and estimation

Missing completely at random (MCAR)

Missing data is independent of any other measured variable (y2) and independent of the variable itself (y1).

- I.e., SES=y2; depression=y1.
- If participants dropped out across a range of SES levels, then the missing on depression would be independent of SES.
- Little's MCAR test in MVA indicates whether MCAR or not (want ns)

Missing at random (MAR)

Missing data may be dependent on another measured variable (y2), but is independent of the variable itself (y1).

- I.e., SES=y2; depression=y1.
- If participants only from high levels of SES dropped out , then the missing on depression would be dependent on SES.
- MAR can be inferred if Little's test is signif but missingness predictable from other vars (other than the variable itself) – tested by Separate Variance Test. MNAR indicated if this test reveals missingness related to the DV
- Consult supervisor or statistics advisor (not my area of expertise so check with someone who is)

Replacing Missing Values

Mean value replacement: NO - restricts the variance of the variables involved.

If <5% missing data

doesn't really matter what method

Tabachnick & Fidell (2007), pp. 62-72

Schafer, J. L., & Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods, 7*, 147-177.

Little & Rubin (1987) *Statistical analysis with missing data*. New York: John Wiley

Replacing Missing Values

Expectation Maximisation (EM):

Use “Missing values analysis” in SPSS and create new data set

```
MVA
  timedrs attdrug atthouse income  emplmnt mstatus race
  /MAXCAT = 25
  /CATEGORICAL = emplmnt mstatus race
  /NOUNIVARIATE
  /TTEST NOPROB PERCENT=5
  /MPATTERN DESCRIBE = timedrs attdrug atthouse income  emplmnt
    mstatus race
  /EM ( TOLERANCE=0.001 CONVERGENCE=0.0001 ITERATIONS=25
```

- OUTFILE='F:\Workshop_cleaning data\missing data tabachnick with EM‘+ ' substitution.sav') .

See Table 4.1 in T&F (2007) for output and discussion of limitations of EM substitution

Repeat analysis with EM sub and complete cases and check for similarity

SPSS Exam data (Field, 2000)

- 100 stats students
- ***Exam*** = first year statistics exam scores
- ***Computer*** = measure of computer literacy
- ***Lecture*** = percentage of statistics lectures attended
- ***Numeracy*** = measure of student's numeracy out of 15

Give ID numbers to cases

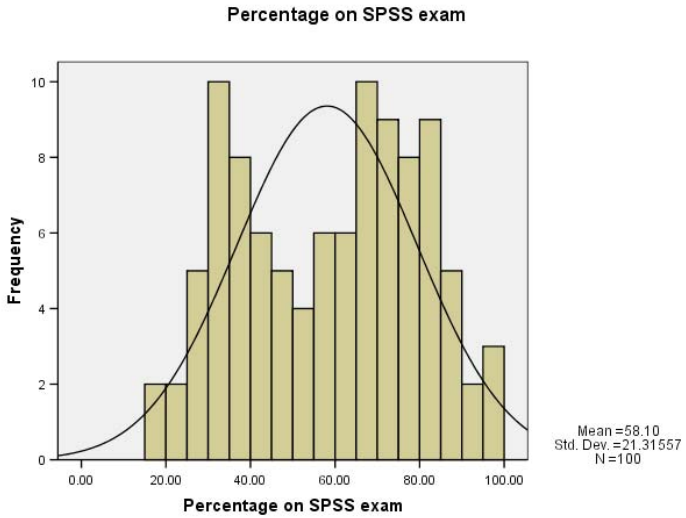
Check min & max scores within range

Types of distributions

Refer to handout for details

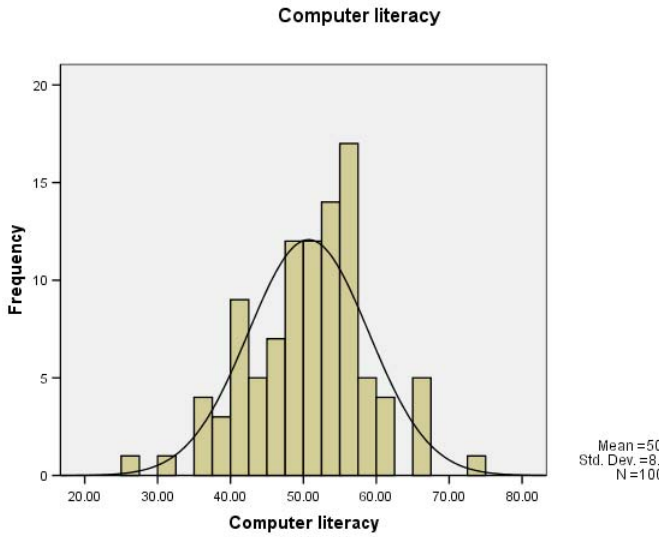
freq [VARIABLE LIST]
  /stat= def skew seskew kurt sekurt /hist= norm.

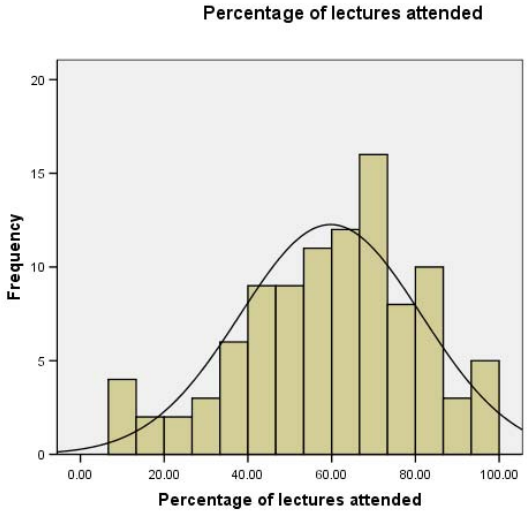
Distributions

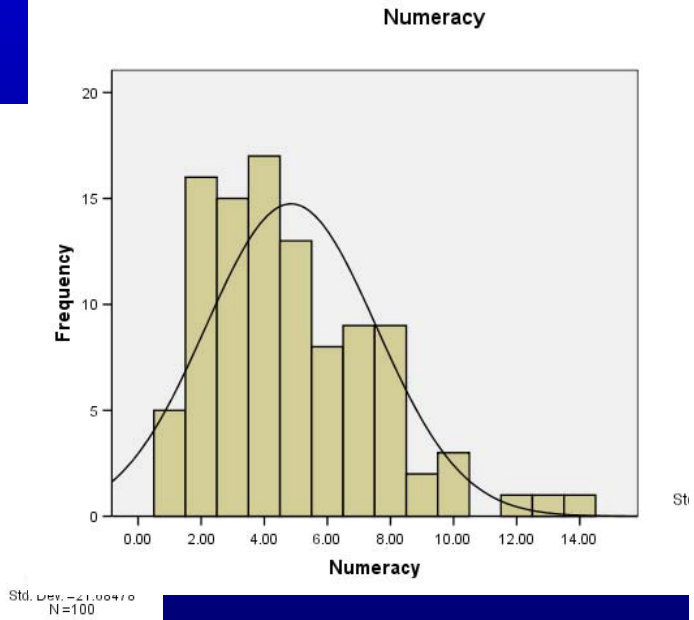


Statistics

		Percentage on SPSS exam	Computer literacy	Percentage of lectures attended	Numeracy
N	Valid	100	100	100	100
	Missing	0	0	0	0
Mean		58.1000	50.7100	59.7650	4.8500
Std. Deviation		21.31557	8.26004	21.68478	2.70568
Skewness		-.107	-.174	-.422	.961
Std. Error of Skewness		.241	.241	.241	.241
Kurtosis		-1.105	.364	-.179	.946
Std. Error of Kurtosis		.478	.478	.478	.478
Minimum		15.00	27.00	8.00	1.00
Maximum		99.00	73.00	100.00	14.00



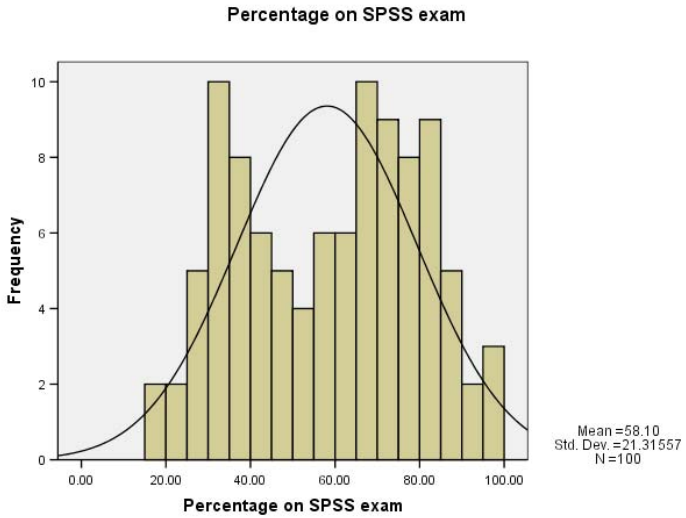




- [x] Give ID numbers to cases
- [x] Check min & max scores within range
- [x] Types of distributions
- [x] Check skewness & kurtosis
- [x] Identify univariate outliers

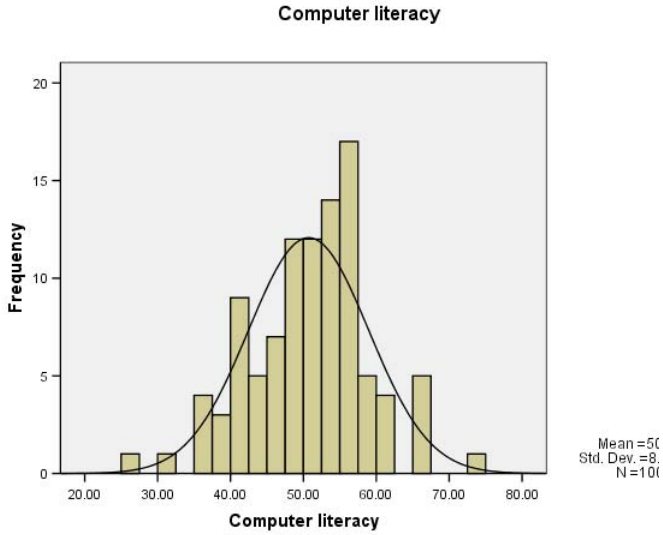
Distributions

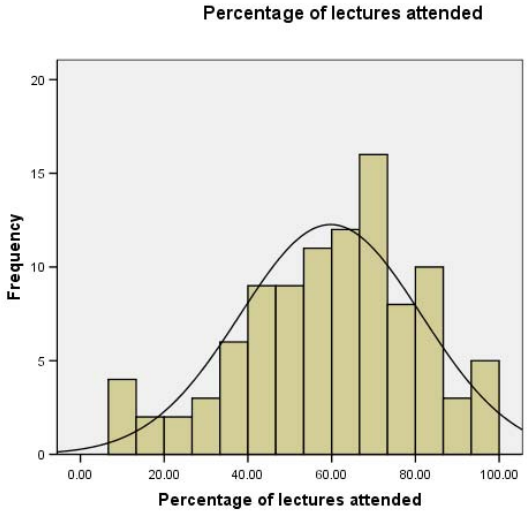
.241* 3 = .723

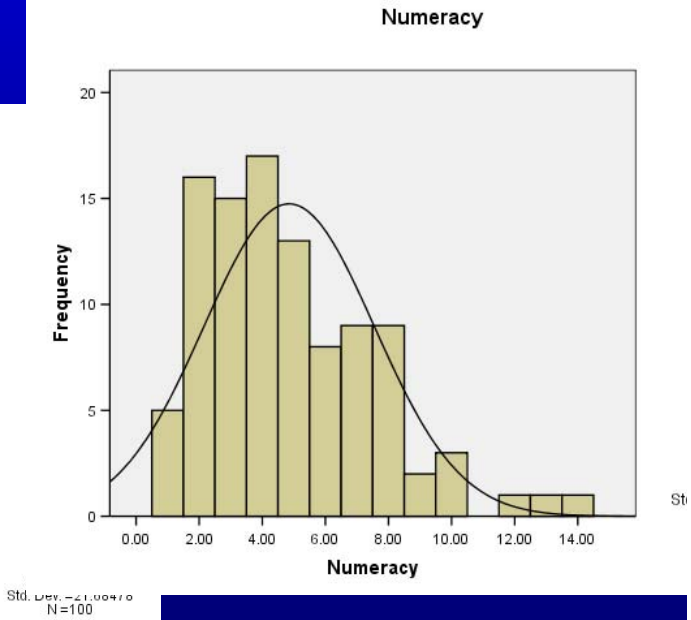


Statistics

		Percentage on SPSS exam	Computer literacy	Percentage of lectures attended	Numeracy
N	Valid	100	100	100	100
	Missing	0	0	0	0
Mean		58.1000	50.7100	59.7650	4.8500
Std. Deviation		21.31557	8.26004	21.68478	2.70568
Skewness		-.107	-.174	-.422	.961
Std. Error of Skewness		.241	.241	.241	.241
Kurtosis		-1.105	.364	-.179	.946
Std. Error of Kurtosis		.478	.478	.478	.478
Minimum		15.00	27.00	8.00	1.00
Maximum		99.00	73.00	100.00	14.00







Outliers

- Using “Descriptives” ask for Z scores
- Identify data points >3.29 or < -3.29

Using syntax:

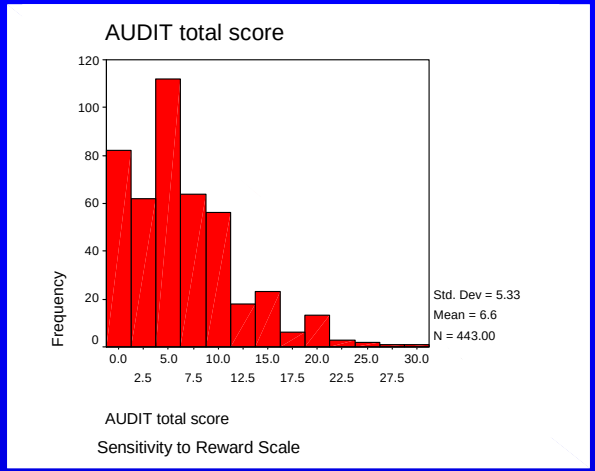
```
Desc [var1 var2 var3]  /SAVE.

*by default, var called z[oldvarname].

*check if there are any greater than Zabs = 3.29.
freq [ZVAR]
  /stat= min max /format=notable.

*list the offenders.
temp.
sel if ( [ZVAR] > 3.29 or [ZVAR] < - 3.29 ).
list id [ZVAR].
```

Outliers



Reasons for outliers

1. Data entry error
2. Failure to specify 99 or 999 as missing data
3. Outlier not a true member of population of interest
4. Outlier is a true member of population of interest with an extreme score

What to do?

1. Look at histogram
    1. Sometimes transforming data can “pull in” the outlier
    2. Censoring outliers
    3. May need to delete case/s and run with and without outlier

- [x] Give ID numbers to cases
- [x] Check min & max scores within range
- [x] Types of distributions
- [x] Check skewness & kurtosis
- [x] Identify univariate outliers
- [x] Making a record of decisions

Screening Sheets

Variable	Outliers?	Skewed? +/-?	Kurtosis?	Transformation?	Result of Transformation?	Additional Information

Break

"Eating & Alcohol data"

- 430 female university students
- ***Age*** = age in years
- ***Bul*** = Bulimia Scale
- ***AUDIT*** = Measure of hazardous drinking
- ***SR*** = Sensitivity to Reward
- ***FES_Coh*** = Family Cohesion
- *Your task – run thru the previous checklist and assign an ID and check for skew, kurtosis, outliers and record on screening sheet (missing data already sorted)*

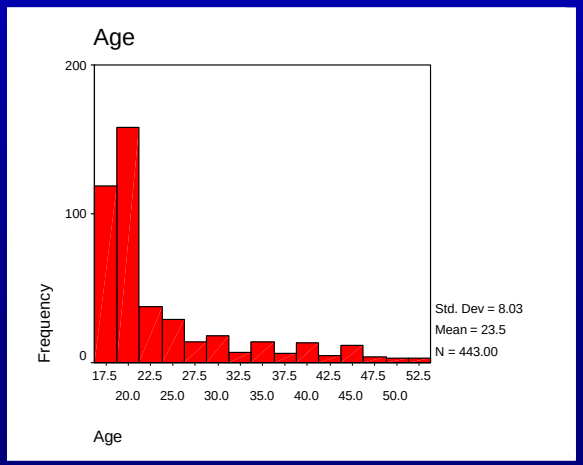
- [x] Give ID numbers to cases
- [x] Check min & max scores within range
- [x] Types of distributions
- [x] Check skewness & kurtosis
- [x] Identify univariate outliers
- [x] Making a record of decisions

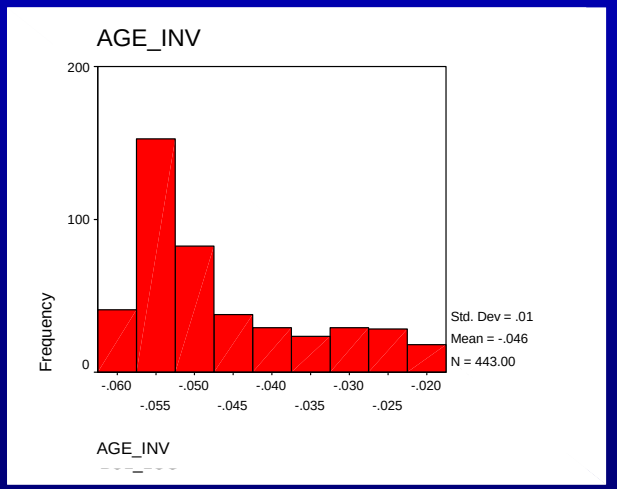
Transforming data

- Positively skewed data
- Moderate skew – use squareroot (SQRT(var))
- More severe skew – use log (lg10(var))
- Horribly skewed – try inverse (-1/(var))
- Note. If start at zero need to add a constant

COMPUTE var_inv = -1/(var+1) .

EXECUTE



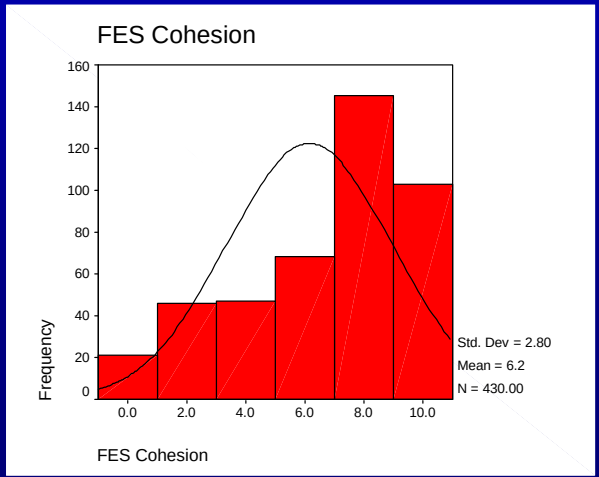


Transforming data

- Negatively skewed data
- Need to reflect and transform, then flip
- COMPUTE var_rsq = -1*SQRT(K-var).
- When K = greatest value plus 1
- E.g., COMPUTE coh_rsq = -1*SQRT(10-coh)

Statistics

		FES Cohesion
N	Valid	430
	Missing	13
Mean		6.1500
Std. Deviation		2.79551
Skewness		-.822
Std. Error of Skewness		.118
Kurtosis		-.567
Std. Error of Kurtosis		.235
Minimum		.00
Maximum		9.00



Transforming data

- Bimodal data
- Check whether 2 underlying pop.s
- Split into dichotomous var around the break and test using analyses that handle dichotomous variables (depends on whether the dichotomous var is an IV or a DV)

Multivariate screening

Need to run a MR

Multivariate outliers – use Mahalanobis Distance (the distance of a point for a participant from the centre of the distribution for all participants)

- use Chi Sq table to see if so distant as to be a statistically significant multivariate outlier

(df = number of IVs; *p* always set to .001)

Regression results

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	4.8239	9.1463	6.6698	.88096	430
Std. Predicted Value	-2.095	2.811	.000	1.000	430
Standard Error of Predicted Value	.290	1.227	.543	.179	430
Adjusted Predicted Value	4.7844	9.4943	6.6678	.88577	430
Residual	-8.15579	21.92994	.00000	5.27521	430
Std. Residual	-1.539	4.138	.000	.995	430
Stud. Residual	-1.552	4.178	.000	1.002	430
Deleted Residual	-8.29190	22.35648	.00193	5.34797	430
Stud. Deleted Residual	-1.554	4.261	.002	1.006	430
Mahal. Distance	.285	21.991	3.991	3.576	430
Cook's Distance	.000	.068	.003	.007	430
Centered Leverage Value	.001	.051	.009	.008	430

a. Dependent Variable: AUDIT total score

TABLE C.4 CRITICAL VALUES OF CHI SQUARE (χ^2)

df	0.250	0.100	0.050	0.025	0.010	0.005	0.001
1	1.32330	2.70554	3.84146	5.02389	6.63490	7.87944	10.828
2	2.77259	4.60517	5.99147	7.37776	9.21034	10.5966	13.816
3	4.10835	6.25139	7.81473	9.34840	11.3449	12.8381	16.266
4	5.38527	7.77944	9.48773	1.1433	13.2767	14.8602	18.467
5	6.62568	9.23635	11.0705	12.8325	15.0863	16.7496	20.515
6	7.84080	10.6446	12.5916	14.4494	16.8119	18.5476	22.458
7	9.03715	12.0170	14.0671	16.0128	18.4753	20.2777	24.322
8	10.2188	13.3616	15.5073	17.5346	20.0902	21.9550	26.125
9	11.3887	14.6837	16.9190	19.0228	21.6660	23.5893	27.877
10	12.5489	15.9871	18.3070	20.4831	23.2093	25.1882	29.588
11	13.7007	17.2750	19.6751	21.9200	24.7250	26.7569	31.264
12	14.8454	18.5494	21.0261	23.3367	26.2170	28.2995	32.909
13	15.9839	19.8119	22.3621	24.7356	27.6883	29.8194	34.528
14	17.1770	21.0642	23.6848	26.1190	29.1413	31.3193	36.123
15	18.2451	22.3072	24.9958	27.4884	30.5779	32.8013	37.697
16	19.3688	23.5418	26.2962	28.8454	31.9999	34.2672	39.252
17	20.4887	24.7690	27.5871	30.1910	33.4087	35.7185	40.790
18	21.6049	25.9894	28.8693	31.5264	34.8053	37.1564	42.312
19	22.7178	27.2036	30.1435	32.8523	36.1908	38.5822	43.820
20	23.8277	28.4120	31.4104	34.1696	37.5662	39.9968	45.315
21	24.9348	29.6151	32.6705	35.4789	38.9321	41.4010	46.797
22	26.0393	30.8133	33.9244	36.7807	40.2894	42.7956	48.268
23	27.1413	32.0069	35.1725	38.0757	41.6384	44.1813	49.728
24	28.2412	33.1963	36.4151	39.3641	42.9798	45.5585	51.179
25	29.3389	34.3816	37.6525	40.6465	44.3141	46.9278	52.620
26	30.4345	35.5631	38.8852	41.9232	45.6417	48.2899	54.052
27	31.5284	36.7412	40.1133	43.1944	46.9630	49.6449	55.476
28	32.6205	37.9159	41.3372	44.4607	48.2782	50.9933	56.892
29	33.7109	39.0875	42.5569	45.7222	49.5879	52.3356	58.302
30	34.7998	40.2560	43.7729	46.9792	50.8922	53.6720	59.703
40	45.6160	51.8050	55.7585	59.3417	63.6907	66.7659	73.402
50	56.3336	63.1671	67.5048	71.4202	76.1539	79.4900	86.661
60	66.9814	74.3970	79.0819	83.2976	88.3794	91.9517	99.607
70	77.5766	85.5271	90.5312	95.0231	100.425	104.215	112.317
80	88.1303	96.5782	101.879	106.629	112.329	116.321	124.839
90	98.6499	107.565	113.145	118.136	124.116	128.299	137.208
100	109.141	118.498	124.342	129.561	135.807	140.169	149.449

Source: Adapted from Table 8 in *Biometrika Tables for Statisticians*, vol. 1, 2d ed., edited by E. S. Pearson and H. O. Hartley (New York: Cambridge University Press, 1958). Reproduced with the permission of the *Biometrika* trustees.

Residuals Statist

	Minimum	Maximum
Predicted Value	4.8239	9.1463
Std. Predicted Value	-2.095	2.811
Standard Error of Predicted Value	.290	1.227
Adjusted Predicted Value	4.7844	9.4943
Residual	-8.15579	21.92994
Std. Residual	-1.539	4.138
Stud. Residual	-1.552	4.178
Deleted Residual	-8.29190	22.35648
Stud. Deleted Residual	-1.554	4.261
Mahal. Distance	.285	21.991
Cook's Distance	.000	.068
Centered Leverage Value	.001	.051

a. Dependent Variable: AUDIT total score

Multivariate screening

Multicollinearity :

Correlation between IVs

If <.7 ok

Check bivariate correlations between IVs

Tolerance:

Amount of variance NOT predicted by the other IVs

>.01 ok

Place “tol” in syntax or check collinearity diagnostics when run MR

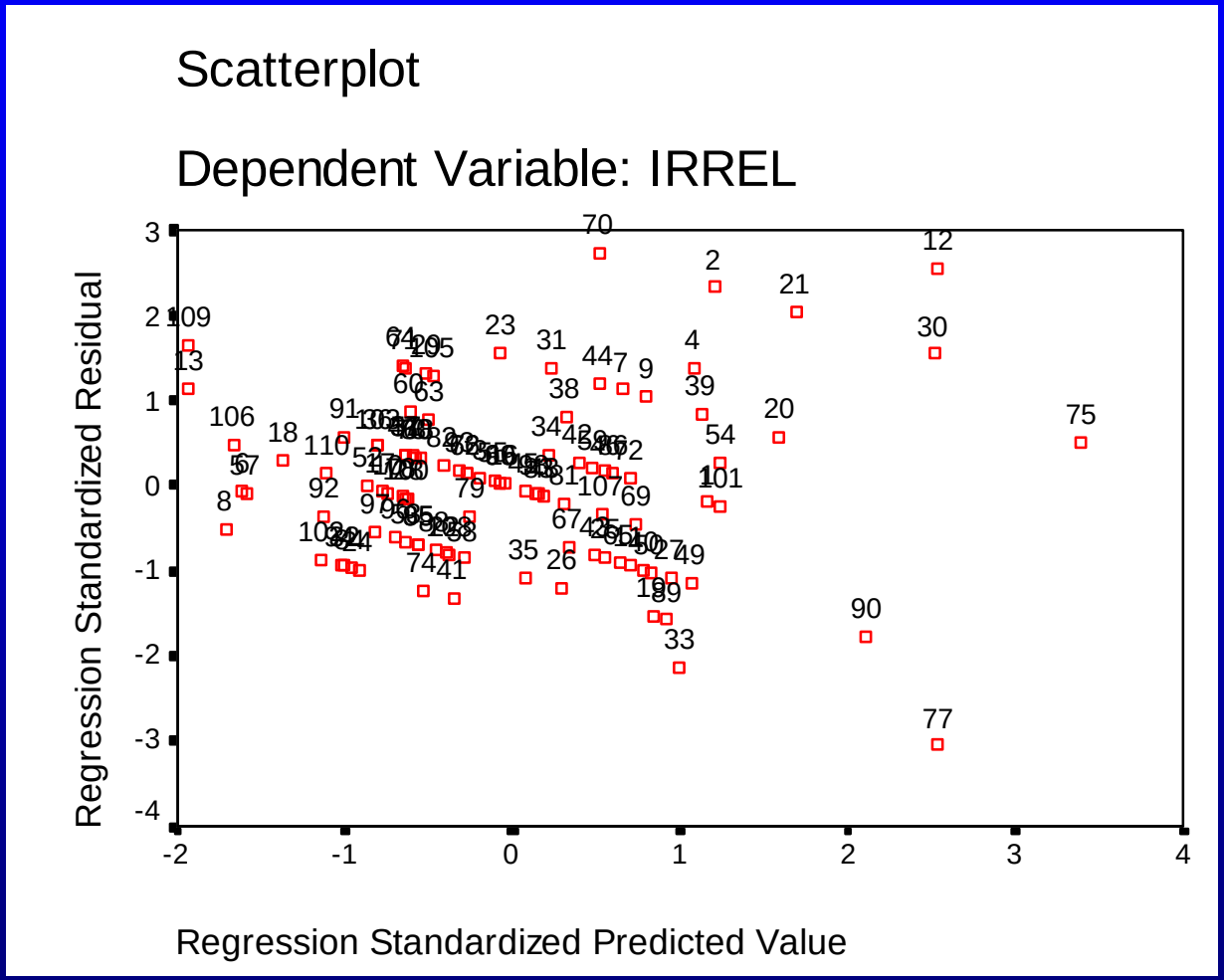
Regression results

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	8.709	1.603		5.431	.000		
	Age	.000	.035	-.001	-.013	.990	.868	1.152
	Bulimia 6 point scale	.045	.042	.054	1.066	.287	.892	1.121
	Sensitivity to Reward Scale	-.116	.067	-.090	-1.744	.082	.867	1.154
	FES Cohesion	-.250	.097	-.131	-2.574	.010	.890	1.123

a. Dependent Variable: AUDIT total score

Normality of residuals



Should be square-ish

Cf T&F

Deciding which output to report

Need to run multiple analysis:

- Non-transformed, outliers in
- Non-transformed, outliers out
- Transformed, outliers in
- Transformed, outliers out

May not need if transformation "fixes" the outliers

Summary table

Use this table to make the decision of which analysis to report

		IVs	**Transformed**	
			N	Y
Outliers	IN			
	OUT			

Rules: 1) If significance doesn't change (to or from non- significance), report NON-TRANSFORMED, if changes report TRANSFORMED

2) Univariate and multivariate outlier may need to be deleted if influencing results and/or don't seem to be part of the pop. of interest. Try to reduce the influence by transformation or censoring. If in doubt consult with supervisor.