Bivariate Data Cleaning

Bivariate Outliers In Simple Correlation/Regression Analyses

Imagine we are interested in the correlation between two variables. Being schooled about outliers we examine the distribution of each variable before beginning the analysis.

Statistics

		X1	Y1
N	Valid	50	50
	Missing	0	0
Mean		45.7200	46.4400
Std. Deviation		14.94349	8.13950
Skewness		-.277	-.384
Std. Error of Skewness		.337	.337
Minimum		17.00	29.00
Maximum		69.00	64.00

Percentiles

		Percentiles		
		25	50	75
Tukey's Hinges	X1	34.0000	47.5000	58.0000
	Y1	42.0000	47.5000	51.0000

For X 1.5*hinge spread = 36, with outlier boundaries of -2 & 94 → no outliers

For Y 1.5*hinge spread = 13.3, with outlier boundaries of 28.7 & 64.3 → no outliers (but close).

So, we proceed with getting the correlation between the variables.

Correlations

		X1	Y1
X1	Pearson Correlation	1	.191
	Sig. (2-tailed)	.	.184
	N	50	50
Y1	Pearson Correlation	.191	1
	Sig. (2-tailed)	.184	.
	N	50	50

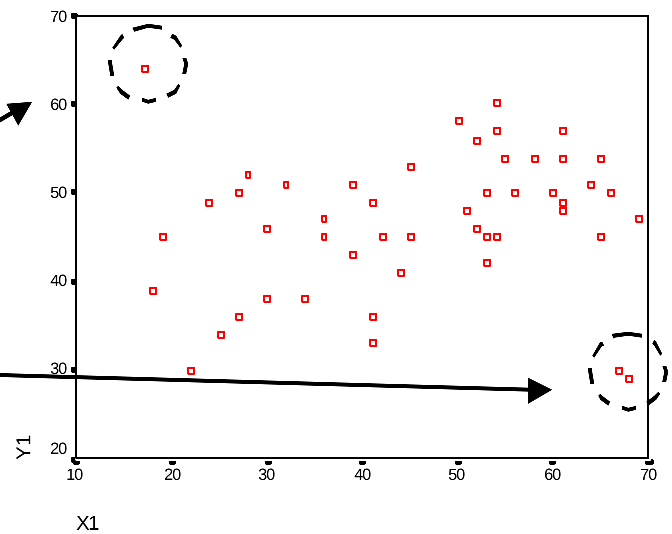
Clearly nonsignificant.

But we also knew enough to get the corresponding scatterplot.

Which looks more like a positive correlation, except for a few notable cases…

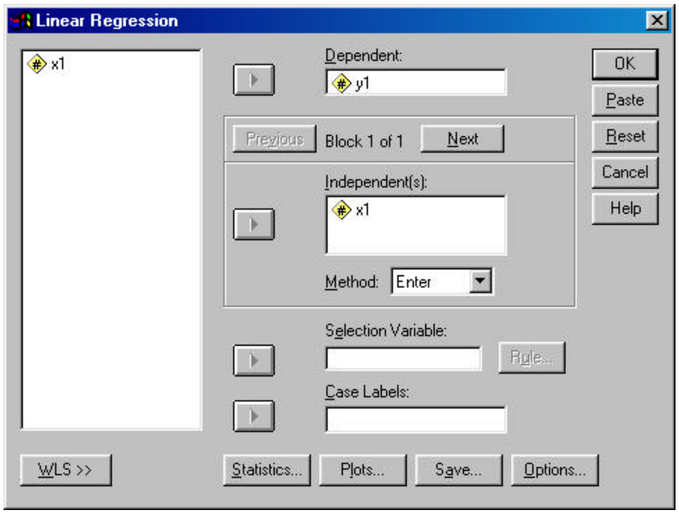
None of these cases are "outliers" on either X or Y, but they do seem to be bivariate outliers. That is, they are outliers relative to the central envelope of the scatterplot.

They might also be "influential cases" in a bivariate sense, because they might be pivoting the regression line and lowering the sample r -- notice that all are away from the means of X and Y!



One way to check if these are such "bivariate outliers" is to examine the residuals of the cases in the analysis. To do this, we obtain the bivariate regression formula, apply it back to each case obtaining the y', and then compute the residual as y-y'. Actually SPSS will do this for us within a regression run.

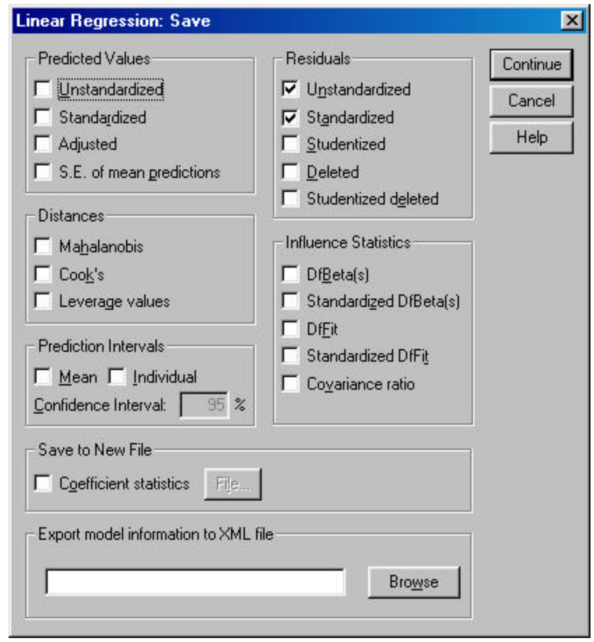
Analyze → Regression → Linear



Highlight and move the criterion and predictor variables.

Click "Save" and check the types of Residuals you'd like.

Then we "Examine" the residual values. To do this we can apply the same outlier formulas we applied to any input variable -- but applying this approach to the residual means we are looking for cases that are specifically "bivariate outliers".



Leaving out a few portions of the output, we get…

Statistics

Unstandardized Residual

N	Valid	50
	Missing	0
Minimum		-19.75953
Maximum		20.54999

Percentiles

		Percentiles		
		25	50	75
Tukey's Hinges	Unstandardized Residual	-4.05412	1.1034447	5.0527943

Applying the formulas, we get…

For X 1.5*hinge spread = 9.1, with outlier boundaries of -13.15 & 14.15 ➔ clearly there are bivariate outliers

If we trim these cases, we can take a look at the before and after regression, to examine the "influence" of these cases.

To trim them **Data ➔ Select Cases** with the formula (res_1 ge -13.15) and (res_1 le 14.15)

Here's the "Before-" and "After-Trimming" regression results.

Before

Model Summary[b]

Model	R	R Square
1	.191[a]	.037

a. Predictors: (Constant), X1

b. Dependent Variable: Y1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	118.596	1	118.596	1.820	.184[a]
	Residual	3127.724	48	65.161		
	Total	3246.320	49			

a. Predictors: (Constant), X1

b. Dependent Variable: Y1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	41.680	3.708		11.240	.000
	X1	.104	.077	.191	1.349	.184

a. Dependent Variable: Y1

Things to notice:

- The change in df error tells us we trimmed 4 cases (upper-left case on scatterplot is 2 cases)
- There was a substantial drop in MSerror, and in the Std Error of the regression coefficient
  - Both indicate the "error" in the model was reduced (since we trimmed cases with large residuals)
- The r and b values are substantially larger and statistically significant

After

Model Summary[b]

Model	R	R Square
1	.471[a]	.221

a. Predictors: (Constant), X1

b. Dependent Variable: Y1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	458.252	1	458.252	12.511	.001[a]
	Residual	1611.683	44	36.629		
	Total	2069.935	45			

a. Predictors: (Constant), X1

b. Dependent Variable: Y1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	36.556	3.126		11.695	.000
	X1	.231	.065	.471	3.537	.001

a. Dependent Variable: Y1

So -- check those scatterplots and examine residuals to look for bivariate outliers.

When doing multiple regression, examine each of the X-Y plots at a minimum. Examining all the inter-predictor plots can be important as well!

"Bivariate Outliers" In Group Comparison Analyses

This time we have a variable "Z" that tells which group each participant was in -- 1 = control and 2 = treatment. We know that there are no outliers on Y1, so we do the ANOVA (after taking out the data selection command).

We get …

Descriptives

Y1

	N	Mean	Std. Deviation
1.00	22	45.0455	9.76133
2.00	28	47.5357	6.74664
Total	50	46.4400	8.21437

We get no effect (darn).

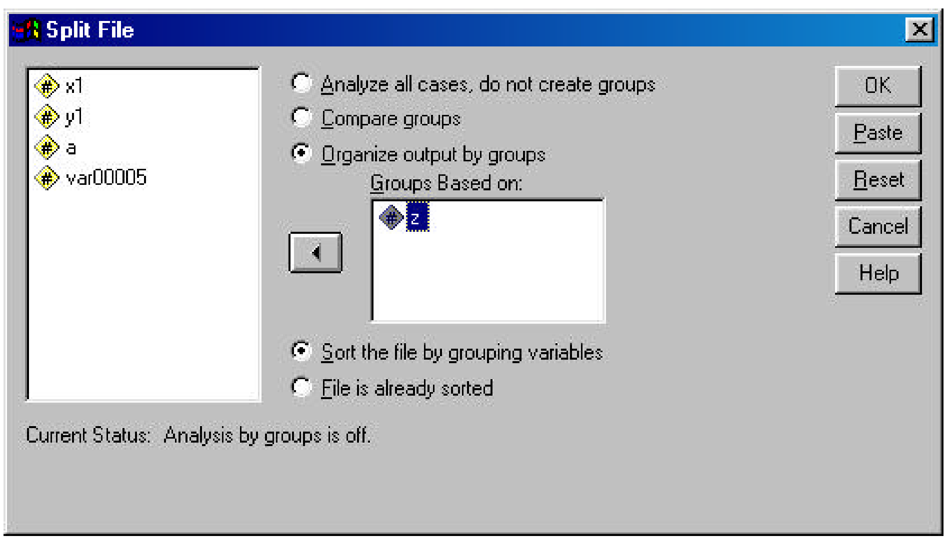
Then we consider, the purpose of **each** group is to represent their respective population. So, maybe our preliminary outlier analyses should be conducted separately for each group!

ANOVA

Y1

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	76.401	1	76.401	1.135	.292
Within Groups	3229.919	48	67.290		
Total	3306.320	49			

Data ➔ Split Files



All subsequent analyses will be done for each group defined by this variable.

Analyze ➔ **Descriptive Statistics** ➔ **Explore** and get the "usual stuff" (only partial results shown below)

Group Z = 1

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Y1	22	30.00	59.00	45.0455	9.76133
Valid N (listwise)	22				

a. Z = 1.00

Percentiles[a]

		Percentiles		
		25	50	75
Tukey's Hinges	Y1	36.0000	47.5000	54.0000

a. Z = 1.00

Outlier boundaries would be…

9 and 81 ➔ no outliers

Group Z = 2

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Y1	28	28.00	56.00	47.5357	6.74664
Valid N (listwise)	28				

a. Z = 2.00

Percentiles[a]

		Percentiles		
		25	50	75
Tukey's Hinges	Y1	45.0000	47.5000	51.0000

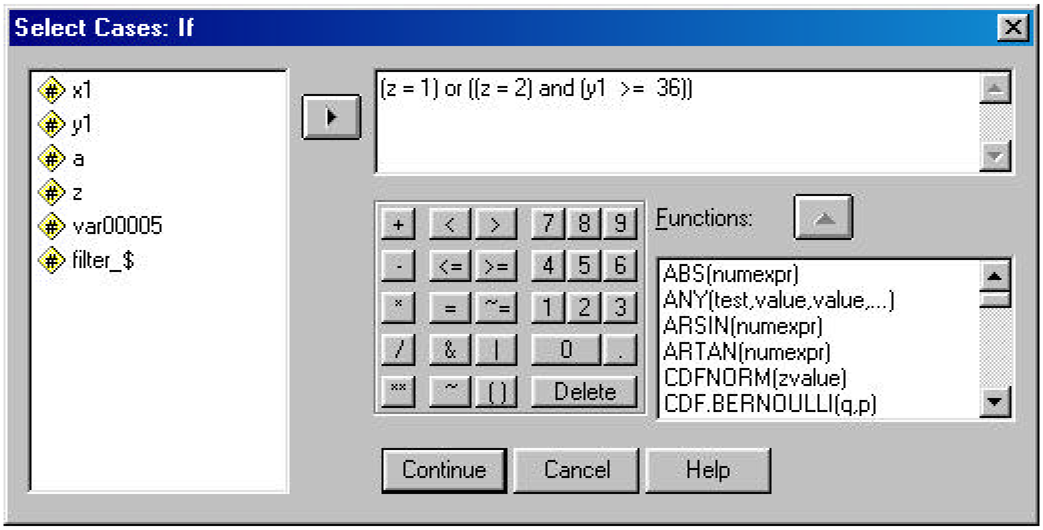
a. Z = 2.00

Outlier boundaries would be …

36 and 60 ➔ at least one "too small" outlier

We can trim outlying values in the Z=2 group, using the following in the Select Cases command…

What we're doing is: 1) selecting everybody in Z=1 (that group has no outliers)
2) selecting only those cases from Z=2 with values that are not outliers



When we do this the ANOVA results are …

Descriptives

Y1

	N	Mean	Std. Deviation
1.00	22	45.0455	9.76133
2.00	27	49.3343	5.65457
Total	50	46.4400	8.21437

ANOVA

Y1

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	298.869	1	298.869	4.867	.032
Within Groups	2947.451	46	61.405		
Total	3246.320	47			

Notice what changes and what doesn't…

Nothing changes for the Z=1 group

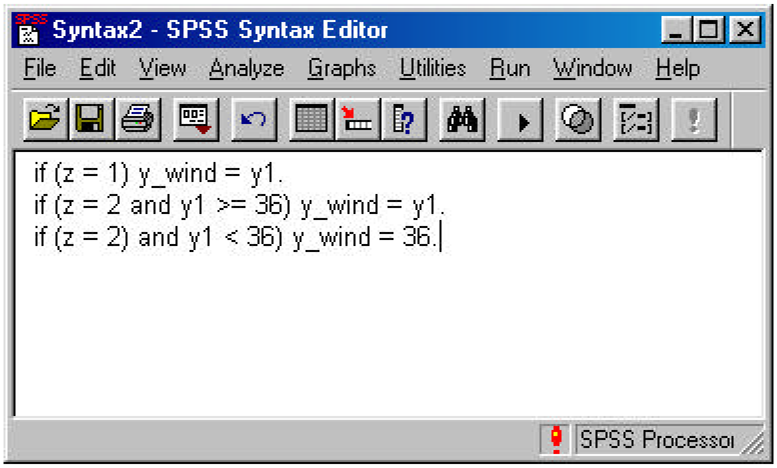
For the Z=2 group
- the sample size drops by 1
- the mean increases (since all the outliers were "too small" outliers)
- the std decreases (because extreme values were trimmed)

The combined results is a significant mean difference -- the previous results with the "full data set" were misleading because of single extreme case!!!

We can Windsorize outlying values in Z=2 using the following commands -- these build a new variable "y-wind" that

- Has the value of y1 for those in the Z=1 group -- no cases need to be Windsorized
- "too small" outliers are changed to a value of 36 for the Z=2 group -- Windsorizing

File ➔ New ➔ Syntax this will open a syntax window into which we can type useful commands…



- The first line assigns the y1 score of each person in group Z=1 as their y_wind score
- The second line assigns the y1 score of "non-outliers" in group Z=2 as their y_wind score
- The third line assigns a y_wind score of "36" (the smallest non-outlying score) to "outliers" in the group Z=2

Then click on the right-pointing arrow to perform the assignments

The results of the Windsorized ANOVA are …

Descriptives

Y1

	N	Mean	Std. Deviation
1.00	22	43.6818	9.81440
2.00	28	48.6071	5.85212
Total	50	46.4400	8.13950

The mean and std of the Z=2 group change in the same direction, but not as much, as the trimmed data.

Similarly, the MSE doesn't drop as much, but the result still changes to a significant mean difference.

ANOVA

Y1

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	298.869	1	295.870	4.732	.042
Within Groups	2947.451	47	62.539		
Total	3246.320	48			

Applying group-specific outlier analysis

- Do outlier analyses and trimming/Windsorizing separately for each group
  - This gets lengthy when working with factorial designs -- remember the purpose of each condition!!!
- Some suggest taking an analogous approach with doing regression analyses that involve a mix of continuous and categorical predictors --- examining each group defined by each categorical variable for outliers on each quantitative variable (and then following up for bivariate outliers among pairs of quantitative variables).