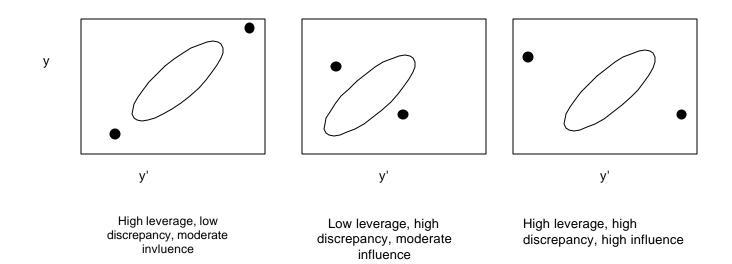
Multiple Regression Diagnostics

Multiple regression is probably the multivariate model that has benefited the most from systematic examinations and applications of data cleaning procedures -- and for good reason, since it is probably the most-used of all the models.

Influential Case Analysis

SPSS provides several diagnostic statistics that allow the case-by-case evaluation of the data for possible influential cases. We'll need some vocabulary...

- Leverage -- assesses outliers among the predictors. All the leverage stats are some variation on Mahalanobis distance (√ Σ(x μ)² where x is each predictor, in turn). Larger scores mean the case is further from the multivariate centroid of the sample. Cases with high leverage are "far out" -- but they might just be far along the regression line, so leverage isn't a sufficient criterion for exclusion.
- **Discrepancy** -- assesses the extent to which a case is in line with others (very similar to what we called "truly bivariate outliers" earlier
- **Influence** -- is the product of leverage and discrepancy -- and is the best single index of whether a case ought to be "hucked". Different combinations of leverage and discrepancy produce different influences ...
 - high leverage and low discrepancy → over-estimates of R², underestimates standard error of regression weights and increased Type I errors when testing H0: R² = 0 and H0: b=0
 - low leverage and high discrepancy → under-estimates of R², overestimates of standard error of regression weights and increased Type II errors when testing H0: R² = 0 and H0: b=0
 - high leverage and high discrepancy → "pivoting" the regression line, underestimates of the R², underestimates of regression weights, overestimates of their standard errors and incressed Type II errors when testing H0: R²=0 and H0: b=0



SPSS includes influence statistics that have a long history -- Cook's Distance, DfBeta and DfFit. When selected form the "Save" menu, these produce values for each case. For each of these, the usual "cutoff" is 1.0 -- cases with values larger than 1.0 are "suspected of being outliers". I found that same phrase in 5 different books and articles! It is all well and good for authors to tell us about suspicions, but we need to make and defend decisions. Usually there are few cases that have large values, and unless we have a really skimpy sample size, tossing them will be the best thing (more below).

Analyze \rightarrow Regression \rightarrow Linear \rightarrow Click the "Save" button

Predicted \	/alues	Residua	als	Continue	1					
☐ Standag	dized		dardized	Cancel	4					
T Adjuste	d	<u> </u>	entized	Help						
S.E. of	mean predictions	□ <u>D</u> ele	ted							
D ' 1		🗖 🗖 Stud	entized d <u>e</u> leted							
Distances		- Influenc	e Statistics							
✓ Mahala ✓ Cook's	nobis	No. Contraction								
IV Loo <u>k</u> 's ∏ Leverag		Df <u>B</u> e	eta(s) dardized DfBeta(a						
, Levela <u>t</u>	le values	Urriter □ Urriter Urriter_		sj						
Prediction	Intervals		dardized DfFit							
<u> </u>	🗖 Individual		ariance ratio							
Confidence	Interval: 95	%								
ALCONOMIC STATEMENT	.sav - SPSS ⊻iew <u>D</u> ata	Data Editor Iransform <u>A</u> r	nalyze <u>G</u> raphs	s <u>U</u> tilities <u>W</u> i	ndow <u>H</u> elp				<u> </u>	
ile <u>E</u> dit		Iransform Ar	1 1 1	* (č 🔳	ndow <u>H</u> elp	0			<u> </u>	
[ile <u>E</u> dit 222 □2 €	⊻iew <u>D</u> ata	Iransform Ar	- 12 #	* (č 🔳		sdb2_1	sdb3_1	sdb4_1	sdb5_1	
Eile <u>E</u> dit	View Data	Iransform Ar	= [?] /4	<u>* 1 </u>	 		sdb3_1 04059	sdb4_1 .02229		
[ile <u>E</u> dit 22 12 (e	View Data	Iransform Ar	# ? 44 995619892085 sdf_1	* 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	sdb1_1	sdb2_1			sdb5_1	
ile <u>E</u> dit 2 : mah_1 1	<u>V</u> iew <u>D</u> ata 	Iransform Ar 3.11 Iev_1 .00772	≖ ि? A4 995619892085 sdf_1 09297	sdb0_1 .02566	sdb1_1 .01204	sdb2_1 .03004	04059	.02229	sdb5_1 05531	
ile <u>E</u> dit 2 1 1 1 2	View Data	Iransform Ar		sdb0_1 .02566 01871	sdb1_1 .01204 .00762	sdb2_1 .03004 .00587	04059 .01681	.02229 .00061	sdb5_1 05531 .00368	
ile <u>E</u> dit	View Data	Iransform Ar 3.11 Iev_1 .00772 .01449 .00903	Image: block of the second s	sdb0_1 .02566 01871 00263	sdb1_1 .01204 .00762 .01218	sdb2_1 .03004 .00587 .03715	04059 .01681 .01411	.02229 .00061 .02603	sdb5_1 05531 .00368 06260	
ile <u>E</u> dit 2 [] (: mah_1 1 2 3 4	View Data	Iransform ▲r Image: Constraint of the section of	Image: boot state Image: boot state 995619892085 sdf_1 09297 .02023 08510 .21262	sdb0_1 .02566 01871 00263 02856	sdb1_1 .01204 .00762 .01218 .00027	sdb2_1 .03004 .00587 .03715 .13313	04059 .01681 .01411 .08964	.02229 .00061 .02603 00002	sdb5_1 05531 .00368 06260 12464	
ile <u>E</u> dit → Imate : mate 1 2 3 4 5	View Data Coo_1 .00144 .00007 .00121 .00753 .00608	Iransform ▲r □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □	Image: boot state Image: boot state 995619892085 sdf_1 09297 .02023 08510 .08510 21262 .19175	sdb0_1 .02566 01871 00263 02856 .12048	sdb1_1 .01204 .00762 .01218 .00027 07077	sdb2_1 .03004 .00587 .03715 .13313 .10319	04059 .01681 .01411 .08964 10067	.02229 .00061 .02603 00002 01776	sdb5_1 05531 .00368 06260 12464 .04420	
le <u>E</u> dit	View Data Coo_1 .00144 .00007 .00121 .00753 .00608 .00481	Iransform Ar □ □ □ 1 □ □ 1 □ □ 1 □ □ 1 □ □ 0 □ □ 0 □ □ 0 □ □ 0 □ □ 0 □ □ 0 □ □ 0 □ □ 0 □ □ 0 □ □ 0 □ □ 0 □ □ 0 □ □ 0 □ □ 0 □ □ 0 □ □ 0 □ □ 0 □ □	Image: boot state Image: boot state 995619892085 sdf_1 09297 .09207 02023 .08510 21262 .19175 17027 .	sdb0_1 .02566 01871 00263 02856 .12048 .04887	sdb1_1 .01204 .00762 .01218 .00027 .07077 .08950	sdb2_1 .03004 .00587 .03715 .13313 .10319 09334	04059 .01681 .01411 .08964 10067 08191	.02229 .00061 .02603 00002 01776 .05318	sdb5_1 05531 .00368 06260 12464 .04420 03561	
ile <u>E</u> dit mah_1 (1 2 3 4 5 6 7 8 9	View Data 2001 2001 2001 2001 2001 2000 2001 2000 200 2000 2	Image: Image	Image: Non-State Image: Non-State 995619892085 sdf_1 09297 .02023 08510 .21262 .19175 .17027 01240 .01240	sdb0_1 .02566 01871 00263 .02856 .12048 .04887 00524	sdb1_1 .01204 .00762 .01218 .00027 .07077 .08950 .00560	sdb2_1 .03004 .00587 .03715 .13313 .10319 .09334 .00095	04059 .01681 .01411 .08964 10067 08191 00220	.02229 .00061 .02603 00002 01776 .05318 .00292	sdb5_1 05531 .00368 06260 12464 .04420 03561 .00335	

- **coo_1** \rightarrow **Cook's Distance** you get one for each regression model (_1, _2, etc.)
- sdf_1 → Standardized DfFit -- you get one each regression model (_1, _2, etc.)
- sdb0_1 -> Standardized DfBeta -- you get one for each predictor, for each model
 - 0_1 is for constant from first model
 - 1_1 is for first predictor in first model
 - 1_2 is for first predictor in second model

Use Cook's and DfFit to make a "keep - dump" decision for each case. Use the DfBetas to identify specific predictors that might be leading to this being an influential case. For example, sometimes they identify a variable for a specific case that, when Windsorized, reduces the influence of that case and allows you to keep it in the analysis.

Important point: What model to consider for influential cases? The "old advice" was to focus on the full model. However, a case might be influential for a reduced model without being influential for the full model (easier to hide amongst larger, mor collinear predictor set.

Collinearity Analysis

Why does collinearity cause "problems"? The higher the collinearity, the greater the discrepancy between bivariate and multivariate contributions of variables. This is "reality" because predictors are correlated with each other, and so combinations of predictors will bring that collinearity with them. However, when we start piling up the predictors, then that very real collinearity can produce apparently uninteresting and possible confusing results (remember the crocodiles!). The best way to handle this very real and representative kind of collinearity is to do what you already know is important -- compare the results from bivariate correlations and different nested and non-nested models to get a complete story about how specific predictors relate to the criterion.

Another issue is when the collinearity is sufficient to perturb the mathematics of regression analysis. In order to compute the multiple regression weights, we have invert the correlation matrix $(X^1 \text{ where } X^*X^1 = I)$. If there is sufficient collinearity, the computation of this inversion will be perturbed, and the resulting regression weights will be wrong. Perhaps the clearest indication that something like this has gone wrong is if any of the predictors have a standardized regression weight (β) that is > 1.0 or < -1.0. When this happens one or more variables will have to be deleted or combined to reduce the collinearity.

The most common summary statistic for evaluating collinearity is **tolerance**. The tolerance value for a particular predictor in a particular model is $1 - R^2$, where the R^2 is obtained using that predictor as a criterion and all others as predictors. SPSS automatically does a tolerance analysis and won't enter the regression model any variable with tolerance < .001 -- that's a variable that shares more than 99.9% of its variance with the rest of the predictor set. While this is a common "cutoff", lots of texts and articles also suggest taking a look at what happens when you delete from the model variables that have "relatively small" tolerances.

Anal	/78 →	Regression →	Linear -	•	Click the	"Statistics"	button
Anary	125 /	Regression 7	Lilleai	/	CHCK THE	Statistics	Dutton

Regression Coefficients <u>E</u> stimates Co <u>n</u> fidence intervals Co <u>v</u> ariance matrix	 Model fit R squared change Descriptives Part and partial correlations Collinearity diagnostics 	Continue Cancel Help
Residuals Durbin-Watson Casewise diagnostics Dutliers outside C All cases	3 standard deviations	

Here's an example of an interesting result from a common mistake. We're trying to predict the number of friends someone reports having from self-reports of the frequency with which they engage in five behaviors.

		Unstandardized Coefficients		Standardi zed Coefficien ts			Correlatio ns	Collinearit y Statistics	
Model		В	Std. Error	Beta	t	Sig.	Zero-order	Tolerance	
1	(Constant)	18.529	3.402		5.447	.000			
	tell jokes	1.392	.556	.081	2.506	.012	.261	.841	
	get others to do things my way	732	.604	038	-1.211	.226	123	.878	
	stick up for others	424	.630	021	673	.501	107	.919	
	forget to return items	571	.603	029	947	.344	053	.916	
	make jokes when others clumsy	9.908E-03	.583	.001	.017	.986	105	.869	

Coefficients^a

a. Dependent Variable: how many friends sub listed

arget Variable:		Numeric Expression:	
scale =		i1 + i2 + i3 + i4 + i5	A
Type&Label		×-	
		J	Y
🗰 subnum 🗰 socio		+ < > 7 8 9 Functions:	
 tissorf 			
🛞 recipbf		→ <= >= 4 5 6 ABS(numexpr) × = ~= 1 2 3 ANY(test,value,value,)	
🛞 numrat		ARSIN(numexpr)	
🚸 recipf			
🗰 i1		CDF.BERNOULLI(q,p)	-
🛞 i2		Jf	15
🛞 i3			
🗰 i4			
🛞 i5	<u>•</u>	OK <u>P</u> aste <u>R</u> eset Cancel Help	

Notice that the scale is composed of the same variables that were the predictor in the first regression model.

The common explanation is that if we include the scale along with the items as predictors is that the scale will be "dumped" because it is perfectly predictable from the items.

But that isn't always what happens -- see below!

		Unstandardized Coefficients		Standardi zed Coefficien ts			Collinearit y Statistics
Model		В	Std. Error	Beta	t	Sig.	Tolerance
1	(Constant)	18.529	3.402		5.447	.000	
	get others to do things my way	2.124	.905	111	-2.346	.019	.392
	stick up for others	1.199	.932	089	-1.949	.052	.420
	forget to return items	1.963	.840	101	-2.336	.020	.472
	make jokes when others clumsy	1.382	.876	075	-1.578	.115	.385
	SCALE	1.392	.556	.212	2.506	.012	.122

Coefficients

a. Dependent Variable: how many friends sub listed

Excluded Variables^b

						Coll	inearity Stati	stics
					Partial			Minimum
Model		Beta In	t	Sig.	Correlation	Tolerance	VIF	Tolerance
1	tell jokes	.a				.000		.000

a. Predictors in the Model: (Constant), SCALE, stick up for others, forget to return items, get others to do things my way, make jokes when others clumsy

b. Dependent Variable: how many friends sub listed

What got dumped because of especially low tolerance (.000 which means $R^2 = 1.00$) was not the scale, but one of the items -- "tell jokes" -- the only one that contributed in the item model. With scale and the other 4 items in the model, some interesting things happen (notice -- we'd not know them to be interesting if we'd not run the item model). Specifically, three of the predictors look to be suppressor variables -- we could work really hard to tell an interesting story about these "suppressors", but it is really just a poor set of estimates produced by the collinearity of including 4 of the 5 items composing "scale".