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.

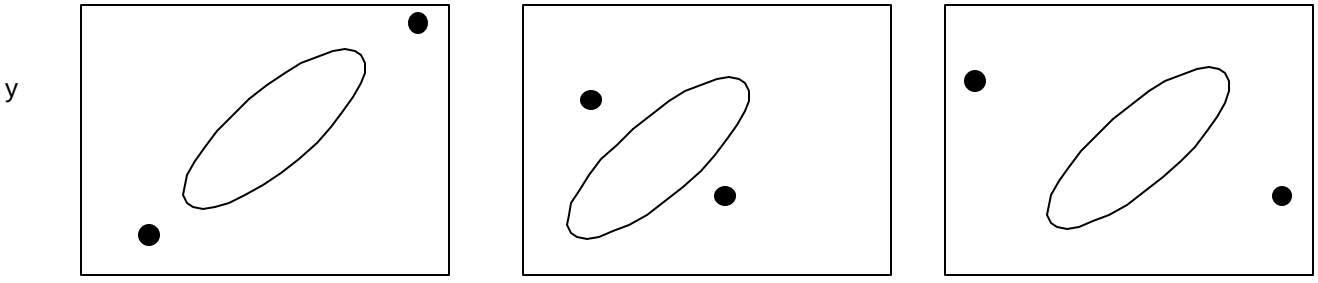
Influential Case Analysis

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 ($\sqrt{\Sigma(x - \mu)^2}$ 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 R^2, underestimates standard error of regression weights and increased Type I errors when testing H0: $R^2 = 0$ and H0: b=0
  - low leverage and high discrepancy → under-estimates of R^2, overestimates of standard error of regression weights and increased Type II errors when testing H0: $R^2 = 0$ and H0: b=0
  - high leverage and high discrepancy → "pivoting" the regression line, underestimates of the R^2, underestimates of regression weights, overestimates of their standard errors and incresed Type II errors when testing H0: $R^2=0$ and H0: b=0



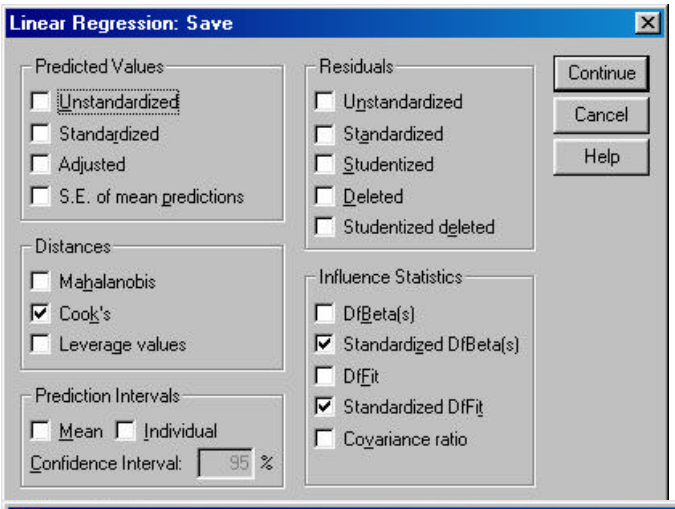
High leverage, low discrepancy, moderate invluence

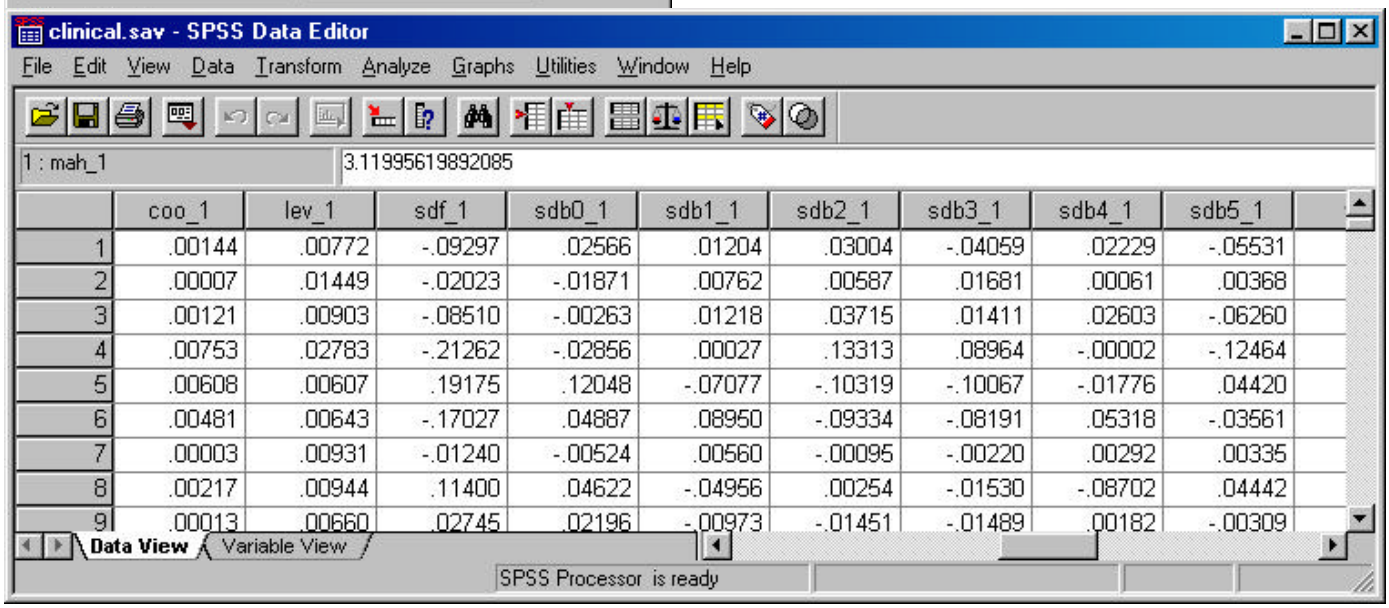
Low leverage, high discrepancy, moderate influence

High leverage, high discrepancy, high influence

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 → Regression → Linear → Click the "Save" button





	coo_1	lev_1	sdf_1	sdb0_1	sdb1_1	sdb2_1	sdb3_1	sdb4_1	sdb5_1
1	.00144	.00772	-.09297	.02566	.01204	.03004	-.04059	.02229	-.05531
2	.00007	.01449	-.02023	-.01871	.00762	.00587	.01681	.00061	.00368
3	.00121	.00903	-.08510	-.00263	.01218	.03715	.01411	.02603	-.06260
4	.00753	.02783	-.21262	-.02856	.00027	.13313	.08964	-.00002	-.12464
5	.00608	.00607	.19175	.12048	-.07077	-.10319	-.10067	-.01776	.04420
6	.00481	.00643	-.17027	.04887	.08950	-.09334	-.08191	.05318	-.03561
7	.00003	.00931	-.01240	-.00524	.00560	-.00095	-.00220	.00292	.00335
8	.00217	.00944	.11400	.04622	-.04956	.00254	-.01530	-.08702	.04442
9	.00013	.00660	.02745	.02196	-.00973	-.01451	-.01489	.00182	-.00309

- **coo_1 → 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.

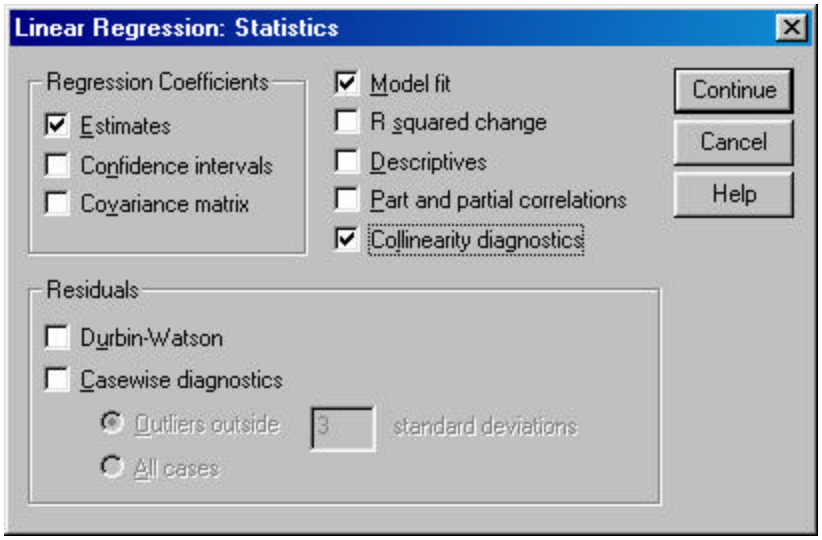
Collinearity Analysis

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} 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.

Analyze → Regression → Linear → Click the "Statistics" button

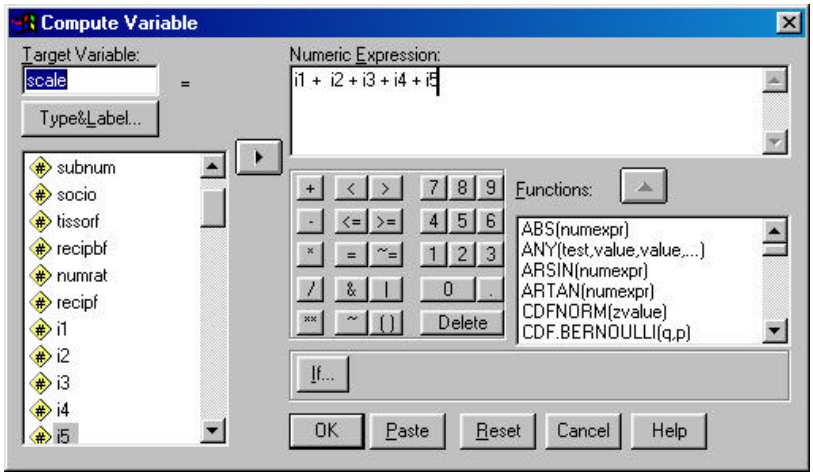


Here's an example of an interesting result from a common mistake. We're trying to predict the number of friends someone reports having from self-reports of the frequency with which they engage in five behaviors.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations	Collinearity Statistics
		B	Std. Error	Beta			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

a. Dependent Variable: how many friends sub listed



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!

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics
		B	Std. Error	Beta			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

a. Dependent Variable: how many friends sub listed

Excluded Variables[b]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum 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".