Regressions to Remember: Increasing R^2 vs. Improving Individual Predictive Accuracy

Whenever a predictor is added to a multiple regression model the R^2 will increase numerically, and if that predictor makes a unique contribution to the model, the r2 will increase significantly. This increase in R2 means that, **on average**, individuals y' values are better predictors of their criterion values – **on average!**

Data from two groups were combined into this sample:
Group 1   N = 21   y' = 4*x1 + 3*x2 + **0***X3 + **-2***X4
Group 2   N = 218   y' = 4*x1 + 3*x2 + **3***X3 + **2***X4

Here are the results of hierarchical (nested) modeling with these data

Model Summary

Model	R	R Square	R Square Change	F Change	df1	df2	Sig. F Change
			Change Statistics				
1	.469[a]	.220	.220	33.284	2	236	.000
2	.488[b]	.239	.019	5.747	1	235	.017
3	.749[c]	.561	.323	172.063	1	234	.000

a. Predictors: (Constant), X2, X1

b. Predictors: (Constant), X2, X1, X3

c. Predictors: (Constant), X2, X1, X3, X4

d. Dependent Variable: Y

Notice that as predictors are added in steps 2 & 3, the overall fit of the modelR^2 increases.

Below is information about the residuals from each model overall, and for each of the groups.
➜ Smaller Std. Deviation means better prediction ⬅

GROUP		Y	Using X1 & X2	Using X1, X2 & X3	Using X1, X2, X3 & X4
1.00	Mean	464.45671232	-176.5355937	-181.9039301	-189.0495044
	N	21	21	21	21
	Std. Deviation	53.764469432	18.91977645	8.93296410	51.59396370
2.00	Mean	650.82106403	17.0057223	17.5228557	18.2111908
	N	218	218	218	218
	Std. Deviation	78.085906674	63.36845434	60.51015334	10.00374946
Total	Mean	634.44595363	.0000000	.0000000	.0000000
	N	239	239	239	239
	Std. Deviation				
	Std. Deviation	92.723242023	81.89051542	80.90724183	61.41839512

What are we looking at?

The "Y" column shows the standard deviation of Y scores for each group. The standard deviations in the other columns are standard deviations of the residuals (y-y') for that model with that group. As predictors are added to a model and R2 increases, the standard deviation of the residuals should decrease (better predictability = less variability in the residuals).

Look at the results from both groups taken together – Total:

Notice that, as would be expected from the increasing R^2 as each predictor is added to the model, the standard deviation of the residuals decreases with each added predictor. This means that as each predictor is added the variability in the residuals decreases – individual predictions are more accurate – on average.

Look at the results from the larger group – Group 2:

Notice that the variability in the residuals decreases as each predictor is added to the model, as expected. However, notice that the mean residual for this group actually increases from when X4 is added. The difference in the role of X4 for the two groups was enough that adding it actually decreased the accuracy of individual predictions in the majority group slightly!!

Look at the results from the smaller group – Group 1:

First, notice that the average residual for each model is **huge**! These folks are being predicted very poorly by this model, even when, in the first model, the predictors "work the same" the two groups. Notice also that the average residual increases a bit as each predictor is added to the model. Most importantly, please notice that when X4 is added into the model (the predictor with oppositely signed weights for the two groups) the standard deviation of the residuals increases dramatically, signaling how poorly this model predicts for this group.

Bottom Line??
- Adding predictors can never reduce a model's R-square!
  - If the predictor helps, it will have a significant regression weight and R-square will increase numerically and significantly.
  - If the predictor doesn't help, it will have a non-significant regression weight and R-square will increase numerically, but not significantly
- **HOWEVER!!**
  - Just because the R-square goes up when one or more predictors are added – correctly indicating better prediction on average…
  - **There may be individuals who are less accurately predicted**
  - **There may be groups that are less accurately predicted**