ANOVA and Regression in SPSS: Pride in America 1996

Neil W. Henry 1998 - 2001

In 1996 the General Social Survey of adult Americans asked a series of questions about the respondent's pride in America. The items - with abbreviated phrasing - were:

1. PROUDDEM How proud are you of the way democracy works
2. PROUDPOL How proud are you of its political influence in the world
3. PROUDECO How proud are you of America's economic achievement
4. PROUDSSS How proud are you of its social security system
5. PROUDSCI How proud are you of its scientific and tech achievements
6. PROUDSPT How proud are you of its achievements in sports
7. PROUDART How proud are you of its achievements in the arts & literature
8. PROUDMIL How proud are you of America's armed forces
9. PROUDHIS How proud are you of its history
10. PROUDGRP How proud are you of its fair and equal treatment of minorities

There were four legal responses to the items:

1:Very Proud,
2:Somewhat Proud,
3:Not Very Proud and
4:Not Proud at All.

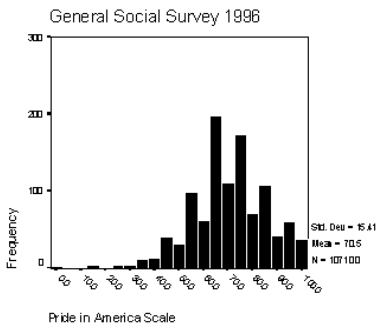
Approximately 1,300 people were asked this question and valid responses on each item ranged from 1,244 to 1,298. (As usual some people decided they "didn't know" or "couldn't say" how proud they were of some item or other.) The analysis that follows is based on the 1,071 persons who gave an interpretable response to all ten items.

I am not interested in an item-by-item analysis of this scale. Instead I constructed an additive scale of **"Pride in America"** by summing up the responses to the ten items. A "scalability analysis" of the items shows that each one is correlated at least 0.4 with the sum of the other nine; the **Cronbach alpha coefficient is 0.83**. This coefficient is a correlation-type measure of the internal consistency of the scale, a generalized split-half reliability coefficient. To make the scale more easily interpretable I adjusted the sum so that **the maximum value was 100 if the respondent said "Very Proud" to all 10 items, and its minimum was 0, if all ten responses were "Not at All Proud".** For those of you who like algebra, the Compute formula is:

PRIDE = 100*(40 - SUM)/30

Comment: In SPSS the compute function "SUM(X1,X2, . . . ,X10)" will compute the sum of all non-missing values on any of the variables. To include only cases where responses to all ten items were valid, I used the function "SUM.10(X1,X2,X3, . . . ,X10)".

The distribution of values of the scale **PRIDE** has mean 70 and standard deviation 15, with noticeable outliers at the low end. I will ignore the outliers for now, but it would undoubtedly be wise to examine them at some point.



Let's see how the responses to this scale are related to gender, race and political party identification. In its original form the latter variable **(partyid),** has seven categories, ranging from strong democrat to strong republican. As you can see from the following tables of subgroup means, the strong democrats and strong republicans both show the highest degree of pride in America, while the independents show the lowest.

Report

Pride in America Scale

RESPONDENTS SEX	Mean	N	Std. Deviation	Std. Error of Mean
MALE	71.80	493	15.07	.68
FEMALE	69.43	578	15.62	.65
Total	70.52	1071	15.41	.47

Report

Pride in America Scale

RACE OF RESPONDENT	Mean	N	Std. Deviation	Std. Error of Mean
WHITE	71.00	900	15.05	.50
BLACK	66.44	119	16.31	1.50
OTHER	71.60	52	18.04	2.50
Total	70.52	1071	15.41	.47

Report

Pride in America Scale

POLITICAL PARTY AFFILIATION	Mean	N	Std. Deviation	Std. Error of Mean
STRONG DEMOCRAT	72.11	150	17.45	1.43
NOT STR DEMOCRAT	69.41	215	14.83	1.01
IND,NEAR DEM	67.95	130	15.41	1.35
INDEPENDENT	68.01	136	16.55	1.42
IND,NEAR REP	68.65	101	14.89	1.48
NOT STR REPUBLICAN	72.23	191	14.38	1.04
STRONG REPUBLICAN	75.08	128	13.31	1.18
OTHER PARTY	68.17	20	14.24	3.19
Total	70.52	1071	15.41	.47

The **three one-way ANOVA tables** all show significance at the .02 level or less (the P-value for **sex** is .012). The **two-way ANOVA** on **race** and **partyid** gives the following results.

ANOVA[a]

			Hierarchical Method				
			Sum of Squares	df	Mean Square	F	Sig.
Pride in America Scale	Main Effects	(Combined)	8198.762	9	910.974	3.967	.000
		RACE	2243.533	2	1121.767	4.885	.008
		POLITICAL PARTY	5955.229	7	850.747	3.705	.001
	2-Way Interactions	RACE * POLITICAL PARTY	5093.352	13	391.796	1.706	.054
	Model		13292.114	22	604.187	2.631	.000
	Residual		240651.884	1048	229.630		
	Total		253943.998	1070	237.331		

a. Pride in America Scale by RACE OF RESPONDENT, POLITICAL PARTY AFFILIATION

To begin with, note that the ANOVA table has several rows, some of which are nested within others. The rows for "model" "residual" and "total" are precisely what we would see if we had done a oneway ANOVA on the full 24 cell, grouping of the sample by **race** and **partyid**. The degrees of freedom for the model (between groups) are 22 rather than 23 because one of the 22 cells has zero observations, as the next table below shows. The model SS is the sum of the entries for **race, partyid** (the two main effects) and interaction. The ANOVA is described as ***hierarchical*** because each effect has been evaluated ***assuming that the previous effects have already been included in the model.*** Thus the table tells us that the 3 category race variable is, by itself, a significant predictor of pride (P = .008); party identification ***adds*** significantly to the predictability of pride (P = .001), and their interaction ***adds*** a little more predictability (P = .054). While the output of this SPSS procedure did not calculate the R-square values at each step, this can be done easily by taking appropriate ratios of the SS at each step to the total sum of squares.

To see what happens when the main effect for party identification is entered first into the model, I just change the order of the predictors in the SPSS command window. The resulting ANOVA table contains many of the same numbers. The one new piece of information in this table tells me that race is still an important predictor of pride in America, even after political affiliation has been controlled for.

The mean values and frequency counts in each of the 24 subgroups defined by the cross-tabulation of race with party identification are shown in the table entitles "Cell Means." Note that the Other Race, Other Party cell of the table is empty. The fact that White Strong Republicans have one of the highest mean levels of Pride in America, while Black Strong Republicans have the lowest mean level would

be an exciting and unexpected finding, if it were not for the fact that there are only two respondents in the Black, Strong Republican category!

ANOVA[a]

			Hierarchical Method				
			Sum of Squares	df	Mean Square	F	Sig.
Pride in America Scale	Main Effects	(Combined)	8198.762	9	910.974	3.967	.000
		POLITICAL PARTY	6042.891	7	863.270	3.759	.000
		RACE	2155.871	2	1077.936	4.694	.009
	2-Way Interactions	POLITICAL PARTY * RACE	5093.352	13	391.796	1.706	.054
	Model		13292.114	22	604.187	2.631	.000
	Residual		240651.884	1048	229.630		
	Total		253943.998	1070	237.331		

a. Pride in America Scale by POLITICAL PARTY AFFILIATION, RACE OF RESPONDENT

Cell Means[b]

POLITICAL PARTY AFFILIATION	Pride in America Scale							
	Mean				N			
	RACE OF RESPONDENT				RACE OF RESPONDENT			
	WHITE	BLACK	OTHER	Total	WHITE	BLACK	OTHER	Total
STRONG DEMOCRAT	73.23	67.21	83.33	72.11	98	43	9	150
NOT STRONG DEMOCRAT	70.30	65.96	68.95	69.41	158	38	19	215
IND,NEAR DEM	67.79	70.22	64.67	67.95	110	15	5	130
INDEPENDENT	68.62	60.91	69.05	68.01	118	11	7	136
IND,NEAR REP	69.01	78.89	52.50	68.65	94	3	4	101
NOT STRONG REPUBLICAN	72.28	67.33	75.33	72.23	181	5	5	191
STRONG REPUBLICAN	75.07	53.33	90.00	75.08	123	2	3	128
OTHER PARTY	69.81	53.33		68.17	18	2	0	20
Total	71.00	66.44	71.60	70.52[a]	900	119	52	1071

a. Grand Mean

b. Pride in America Scale by RACE OF RESPONDENT, POLITICAL PARTY AFFILIATION

To carry out the regression analysis that is equivalent to this analysis of variance **I need to create many indicator variables:** at least as many as the number of degrees of freedom in the model (22). To begin with I would define **Black** and **Other Race** indicators, figuring that my best story would come from comparisons of these groups to the dominant "white" group. Next, I would **define indicators for each of the party identification categories except Independents.** That way regression coefficients would allow for contrasts of particular groups to the (presumed) center of the political spectrum. Defining the 13 or 14 indicators needed for the interaction terms is trickier, if I want the resulting regression coefficients to be easily interpretable. I can, however, just take the 14 product variables defined by multiplying each of the 2 race indicators by each of the 7 party indicators I have already created.

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2243.533	2	1121.767	4.760	.009[a]
	Residual	251700.465	1068	235.675		
	Total	253943.998	1070			
2	Regression	8198.762	9	910.974	3.933	.000[b]
	Residual	245745.236	1061	231.617		
	Total	253943.998	1070			

a. Predictors: (Constant), Black, Other Race

b. Predictors: (Constant), Black, Other race, dem leaning independent, other party, rep leaning independent, strong republican, strong democrat, republican, democrat

c. Dependent Variable: Pride in America Scale

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.094[a]	.009	.007	15.35	.009	4.760	2	1068	.009
2	.180[b]	.032	.024	15.22	.023	3.673	7	1061	.001

a. Predictors: (Constant), black, other race indicators

b. Predictors: (Constant),black, other race indicators, dem leaning independent, other party, rep leaning independent, strong republican, strong democrat, republican, democrat

The two tables above show what happens when the race indicators and then the party identification indicators are entered into the multiple regression procedure of SPSS. The entries in the ANOVA table are the same as those in the first ANOVA table I reported, though they are arranged differently. The R-squared table is useful. The regression routine, of course, gives us more detailed information about the individual categories of race and party identification. The individual regression coefficients provide us with measures of the effect of being in one group or another, compared with other groups. The significance tests on the coefficients are t-tests of ***contrasts,*** in the jargon of analysis of variance.

Among other things, I can see that the **Other Race group has about the same level of pride as whites do**, whether party is controlled or not (P-values .68 and .78, respectively). That indicator could be dropped from the model. Likewise, **when race is controlled for, the republican-leaning independents and democrat-leaning independents cannot be distinguished from the centrist independents (P = .84 and .95 respectively)**. In effect, we may simply collapse these response categories into a single "independent" category, as Agresti and Finlay did in their example in Chapter 12.

Having a strong commitment to one party or the other does make a difference, however, controlling for race. These conclusions could, in a general way, be made by looking at the table of means, but the regression analysis allows us to use the significance test logic more easily.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	70.996	.512		138.740	.000
	Black	-4.554	1.497	-.093	-3.041	.002
	Other Race	.606	2.190	.008	.277	.782
2	(Constant)	68.340	1.317		51.889	.000
	Black	-4.603	1.555	-.094	-2.960	.003
	Other Race	.915	2.190	.013	.418	.676
	strong democrat	5.036	1.831	.113	2.751	.006
	democrat	1.804	1.677	.047	1.076	.282
	dem-indep	.105	1.868	.002	.056	.955
	rep-indep	.407	2.001	.008	.204	.839
	republican	3.990	1.711	.099	2.332	.020
	strong republican	6.789	1.878	.143	3.614	.000
	other party	.287	3.646	.003	.079	.937

a. Dependent Variable: Pride in America Scale

In order to see what the no-interaction model looks like when non-significant indicators are removed, **I next entered all nine predictors and had SPSS perform a "backwards" analysis, dropping the least significant term at each step.** The table that follows shows two of these reduced models. While it does contain a term with P > .05, I prefer Model 5 because the symmetric use of the party labels makes it a little easier to discuss. All these models have virtually the same value of R-square.

Finally, I used the General Linear Model, Univariate (GLM) procedure within SPSS, which produces output similar to what Agresti and Finlay show in Chapter 12. This output combines aspects of the regression and ANOVA approaches, by ***arbitrarily*** selecting one category of each discrete predictor variable (factor) to omit from the regression equation. I don't have to create any indicator variables, but still get the regression coefficients that would correspond to the indicator variables. I specified the "***Type I Sum of Squares***" option, which is the same as the ***hierarchical*** method I used earlier ***(and which I recommend using whenever possible)***. Race was entered first, then party identification, and then the interactions, consistent with the previous multiple regression analysis. As you can see, adding the interaction

terms to the main effect terms increased the R-square from .032 to .052, and the adjusted R-square from .024 to .032. Removing the many non-significant terms from the model would decrease the model's degrees of freedom. It would make little difference in the R-square, but would increase the adjusted R-square. **(Adjusted R-square will increase whenever a variable with a t-coefficient less than 1 in magnitude is removed from a regression model.)**

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
5	(Constant)	68.541	.782		87.644	.000
	Black	-4.688	1.540	-.096	-3.044	.002
	strong democrat	4.914	1.495	.111	3.286	.001
	democrat	1.699	1.301	.044	1.306	.192
	republican	3.816	1.346	.095	2.835	.005
	strong republican	6.610	1.552	.139	4.259	.000
6	(Constant)	69.121	.644		107.374	.000
	Black	-4.458	1.530	-.091	-2.913	.004
	strong democrat	4.268	1.412	.096	3.023	.003
	republican	3.229	1.269	.080	2.544	.011
	strong republican	6.027	1.487	.127	4.053	.000

a. Dependent Variable: Pride in America Scale

Tests of Between-Subjects Effects

Dependent Variable: Pride in America Scale

Source	Type I Sum of Squares	df	Mean Square	F	Sig.	Eta Squared	Noncent. Parameter	Observed Power[a]
Corrected Model	13292.114[b]	22	604.187	2.631	.000	.052	57.885	1.000
RACE	2243.533	2	1121.767	4.885	.008	.009	9.770	.805
PARTYID	5955.229	7	850.747	3.705	.001	.024	25.934	.978
RACE * PARTYID	5093.352	13	391.796	1.706	.054	.021	22.181	.892
Error	240651.884	1048	229.630					
Total	253943.998	1070						

a. Computed using alpha = .05

b. R Squared = .052 (Adjusted R Squared = .032)

The regression coefficients in the model that includes all the possible interaction terms are quite different than the ones in the model that only contains main effect coefficients for race and party identification. One message that I hope it conveys is the impossibility of speaking of "the effect" of a predictor variable in a complex model that involves interaction terms, nonlinear terms, and/or variables whose causal connection to the predictor is unknown.

Parameter Estimates

Dependent Variable	Parameter	B	Std. Error	t	Sig.	95% Confidence Interval		Eta Squared
						Lower Bound	Upper Bound	
Pride in America Scale	Intercept	90.000	17.498	5.144	.000	55.665	124.335	.025
	[RACE=1]	-20.185	17.859	-1.130	.259	-55.228	14.858	.001
	[RACE=2]	-36.667	13.833	-2.651	.008	-63.811	-9.523	.007
	[RACE=3]	0	.	.	.	.	.	.
	[PARTYID=0]	-6.667	18.212	-.366	.714	-42.403	29.070	.000
	[PARTYID=1]	-21.053	17.840	-1.180	.238	-56.058	13.953	.001
	[PARTYID=2]	-25.333	18.764	-1.350	.177	-62.153	11.487	.002
	[PARTYID=3]	-20.952	18.411	-1.138	.255	-57.080	15.175	.001
	[PARTYID=4]	-37.500	19.068	-1.967	.049	-74.915	-.085	.004
	[PARTYID=5]	-14.667	18.764	-.782	.435	-51.487	22.153	.001
	[PARTYID=6]	.000	15.154	.000	1.000	-29.735	29.735	.000
	[PARTYID=7]	0	.	.	.	.	.	.
	[RACE=1] * [PARTYID=0]	10.083	18.622	.541	.588	-26.458	46.624	.000
	[RACE=1] * [PARTYID=1]	21.533	18.234	1.181	.238	-14.246	57.312	.001
	[RACE=1] * [PARTYID=2]	23.306	19.156	1.217	.224	-14.282	60.894	.001
	[RACE=1] * [PARTYID=3]	19.753	18.806	1.050	.294	-17.149	56.656	.001
	[RACE=1] * [PARTYID=4]	36.692	19.462	1.885	.060	-1.497	74.882	.003
	[RACE=1] * [PARTYID=5]	17.135	19.134	.896	.371	-20.411	54.682	.001
	[RACE=1] * [PARTYID=6]	5.253	15.629	.336	.737	-25.414	35.920	.000
	[RACE=2] * [PARTYID=0]	20.543	14.907	1.378	.168	-8.708	49.793	.002
	[RACE=2] * [PARTYID=1]	33.684	14.474	2.327	.020	5.284	62.085	.005
	[RACE=2] * [PARTYID=2]	42.222	15.893	2.657	.008	11.036	73.408	.007
	[RACE=2] * [PARTYID=3]	28.528	15.654	1.822	.069	-2.188	59.244	.003
	[RACE=2] * [PARTYID=4]	63.056	18.036	3.496	.000	27.664	98.447	.012
	[RACE=2] * [PARTYID=5]	28.667	16.829	1.703	.089	-4.355	61.689	.003

a. Computed using alpha = .05