Dale Berger

Equivalence of ANOVA and Regression

Source: htpp://wise.cgu.edu   Guides and Downloads

This handout illustrates the equivalence of ANOVA and regression analyses for a one-way CR-3 design and a two-way CRF 2,4 design. We conduct an ANOVA analysis and then a regression analysis on the same data, using dummy coding for categorical independent variables.

We would like to compare reading readiness for students in three preschools (hypothetical data).

	Head Start	Montessori	Home School
	102	100	101
	90	108	103
	97	104	110
	94	111	106
	98	105	106
	101	102	98
Mean	**97.0**	**105.0**	**104.0**
SD	**4.472**	**4.000**	**4.243**

	school	score
1	1	102
2	1	90
3	1	97
4	1	94
5	1	98
6	1	101
7	2	100
8	2	108

For illustration, we will ask SPSS to compute simple LSD comparisons. We have 18 lines of data with two columns, **school** and **score**. **school** = 1 for the first six rows, 2 for the next six, etc. Click Analysis, Compare means, One-Way ANOVA…, select **score** as the Dependent List and **school** as the Factor. Click Post Hoc, select LSD, click Continue, click Options, select Descriptives, click Continue, click Paste to generate the syntax shown below. Run this syntax.

```
ONEWAY
  score BY school
  /STATISTICS DESCRIPTIVES
  /MISSING ANALYSIS
  /POSTHOC = LSD ALPHA(.05).
```

Descriptives

score

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum
1 Head Start	6	97.00	4.472	1.826	92.31	101.69	90	102
2 Montessori	6	105.00	4.000	1.633	100.80	109.20	100	111
3 Home School	6	104.00	4.243	1.732	99.55	108.45	98	110
Total	18	102.00	5.412	1.276	99.31	104.69	90	111

ANOVA

score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	228.000	2	114.000	6.333	.010
Within Groups	270.000	15	18.000		
Total	498.000	17			

Multiple Comparisons

score
LSD

(I) Pre-School	(J) Pre-School	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1 Head Start	2 Montessori	-8.000*	2.449	.005	-13.22	-2.78
	3 Home School	-7.000*	2.449	.012	-12.22	-1.78
2 Montessori	1 Head Start	8.000*	2.449	.005	2.78	13.22
	3 Home School	1.000	2.449	.689	-4.22	6.22
3 Home School	1 Head Start	7.000*	2.449	.012	1.78	12.22
	2 Montessori	-1.000	2.449	.689	-6.22	4.22

*. The mean difference is significant at the .05 level.

Now we will analyze the data with regression. It would be a BIG mistake to use school as a predictor variable in the current form. The numbers 1,2,3 are simply labels that do not indicate the amount of 'school.' There are two degrees of freedom, so we need two 'indicator' or dummy variables to capture the school variable for regression. Only two dummy variables are needed, but we will show what happens when we use all three. It is easy to create these variables using the syntax window. We can enter the first recode, copy it twice, and edit the two copies as shown.

```
RECODE  school  (1=1)(2,3=0)(ELSE=sysmis)  INTO  dum1 .
RECODE  school  (2=1)(1,3=0)(ELSE=sysmis)  INTO  dum2 .
RECODE  school  (3=1)(1,2=0)(ELSE=sysmis)  INTO  dum3 .
EXECUTE .
```

We can also do this with point-and-click, but it is more work. Click Transform, Recode, Into Different Variables…, select **school** as the *numeric variable*, click Old and New Values, enter **1** as the *Old value*, enter **1** as the *New Value*, click Add, click All other values, enter **0** as the *New Value*, click Continue, under *Output Variable* enter the Name as **dum1**, click Change, and click Paste. This gives you the first line of recodes shown above. You can repeat this process for the other two dummy variables, or you could do the first one by point-and-click, and then Paste the command, copy the line twice, and edit the copies for the other two dummy variables.

To run the regression, click Analyze, Regression, Linear…, select **score** as the *Dependent*, highlight all three dummy variables and click the arrow to make them all *Independents*. Click Statistics and select Estimates, Model fit, R squared change, and Descriptives. Click Continue and Paste to save the syntax.

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA CHANGE
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT score
  /METHOD=ENTER dum1 dum2 dum3  .
```

Much of this syntax is default. If we used SPSS stepwise (usually a bad idea), a variable not in the model would be entered if its probability was less than .05 and a variable in the model would be removed if its probability was greater than .10. Thus, PIN(.05) and POUT(.10).
We do not require the regression line to pass through the origin, hence NOORIGIN.

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
score	102.00	5.412	18
dum1	.33	.485	18
dum2	.33	.485	18
dum3	.33	.485	18

School	dum1	dum2	dum3
1	1	0	0
2	0	1	0
3	0	0	1

Correlations

		score	dum1	dum2	dum3
Pearson Correlation	score	1.000	-.672	.403	.269
	dum1	-.672	1.000	-.500	-.500
	dum2	.403	-.500	1.000	-.500
	dum3	.269	-.500	-.500	1.000
Sig. (1-tailed)	score	.	.001	.048	.140
	dum1	.001	.	.017	.017
	dum2	.048	.017	.	.017
	dum3	.140	.017	.017	.
N	score	18	18	18	18
	dum1	18	18	18	18
	dum2	18	18	18	18
	dum3	18	18	18	18

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	dum3, dum2[a]	.	Enter

a. Tolerance = .000 limits reached.

b. Dependent Variable: score

Note that SPSS used only two of the three dummy variables. For the test of the overall school effect, it doesn't matter which two are used. $F(2, 15) = 6.333$, $p = .010$.

Model Summary

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.677[a]	.458	.386	4.243	.458	6.333	2	15	.010

a. Predictors: (Constant), dum3, dum2

Compare this *F* value to the *F* from the one-way ANOVA test. The result is identical. The null hypothesis also is identical. (Ho: all three school means are equal.) Furthermore, the assumptions are identical – random independent sampling, normal distributions of error, equal variances.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	228.000	2	114.000	6.333	.010[a]
	Residual	270.000	15	18.000		
	Total	498.000	17			

a. Predictors: (Constant), dum3, dum2

b. Dependent Variable: score

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	97.000	1.732		56.003	.000
	dum2	8.000	2.449	.717	3.266	.005
	dum3	7.000	2.449	.627	2.858	.012

a. Dependent Variable: score

Excluded Variables[b]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	dum1	.[a]	.	.	.	.000

a. Predictors in the Model: (Constant), dum3, dum2

b. Dependent Variable: score

dum1 was not included in the analysis because after **dum2** and **dum3** were entered into the model, there is no unique information about school left to be contributed by **dum1** so tolerance=0. When we know the values of **dum2** and **dum3**, we can determine the value for **dum1**. The group that is omitted is called the 'reference' group.

Predicted **score** = 97.000 + 8.000***dum2** + 7.000***dum3**.

What is the predicted score for someone in the reference group? In this case, **dum2** and **dum3** both equal zero, so the predicted score is 97, the constant. Note that the constant is the mean for the reference group. The null hypothesis for the test of the constant is that the value of the constant is zero in the population (e.g., that the population mean for the reference group is zero).

What is the predicted score for someone in Group 2? For Group 2, **dum2**=1 and **dum3**=0. This gives a predicted value of 97 + 8 = 105, which is the mean for Group 2. The B weight for **dum2** is 8.000, which is the difference between the mean for Group 2 and the mean for the reference group.

The null hypothesis for the test of B for **dum2** is that the population value is zero for B, which would be true if the population means were equal for Group 2 and the reference group. We find this difference to be statistically significant, with t=3.266 and p=.005. Compare this to the LSD test in the ANOVA comparing the first two groups (-8.000 / 2.449 = 3.266).
Important note: The test of the B coefficient for **dum2** in the regression analysis is a test of the difference in the means between Group 2 and the reference group. In contrast, the test of the simple correlation of **dum2** with **score** provides a test of the difference between the mean for Group 2 and the mean for all other groups combined. This distinction between the test of B and the test of r is critical because otherwise the test of B in regression is likely to be misinterpreted. Important practice: Explain the difference between these two tests of **dum2** to Bumble.

For illustration, we will fit a model with only **dum1** and **dum2** as predictors. We will do this hierarchically, with **dum2** entered first. Given what you know about the group means, can you figure out what the constant will be and what the two B coefficients will be in the final model?

Here is the syntax for the new model:

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA CHANGE
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT score
  /METHOD=ENTER dum2 /METHOD=ENTER dum1 .
```

Model Summary

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.403[a]	.163	.110	5.105	.163	3.108	1	16	.097
2	.677[b]	.458	.386	4.243	.295	8.167	1	15	.012

a. Predictors: (Constant), dum2

b. Predictors: (Constant), dum2, dum1

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	81.000	1	81.000	3.108	.097[a]
	Residual	417.000	16	26.063		
	Total	498.000	17			
2	Regression	228.000	2	114.000	6.333	.010[b]
	Residual	270.000	15	18.000		
	Total	498.000	17			

a. Predictors: (Constant), dum2

b. Predictors: (Constant), dum2, dum1

c. Dependent Variable: score

Compare Model 2 to the previous analysis. We find exactly the same ability to predict **score** using these two dummy variables. **School** has only two *df*, and two dummy variables are adequate to capture those two pieces of information. $F = 6.333$, $p = .010$. We can do it in many different ways.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	100.500	1.474		68.194	.000
	dum2	4.500	2.553	.403	1.763	.097
2	(Constant)	104.000	1.732		60.044	.000
	dum2	1.000	2.449	.090	.408	.689
	dum1	-7.000	2.449	-.627	-2.858	.012

a. Dependent Variable: score

The final regression model now is Predicted **score** = 104.000 -7.000***dum1** + 1.000***dum2**. The test of statistical significance for **dum2** is no longer statistically significant. Why?

In this model, Group 3 is the reference group. The B coefficient of 1.000 for **dum2** is the difference between the mean for the reference group (104) and the mean for Group 2 (105). This difference is not statistically significant.

In Model 1 the coefficient for **dum2** is 4.500. How can we interpret this value?
Hint: What does **dum2** test when nothing else is in the model? Which means are compared?

In Model 1, B for **dum2** = 4.500 is the difference between the mean for Group 2 (105) versus Group 1 (97) and Group 3 (104) combined (average 100.5). 4.500 = 105 – 100.5. In Model 2, where **dum1** is also in the model, the only unique information **dum2** contributes is the distinction between Group 2 and Group 3.

Important lesson: The interpretation of a test of a variable depends critically upon what else is in the model. The test of a variable in a model is a test of the unique contribution of the predictor variable beyond variance explained by all of the other predictor variables in the model. Thus, the test depends on what else is in the model and how much those variables overlap the predictor of interest.

Study questions:

Interpret the test of statistical significance for **dum1**. Interpret the B coefficient for **dum1**.

The t = 60.044 for the constant is quite impressive. What is the null hypothesis of this test?

How many degrees of freedom do we have on the t tests? [Hint: df = 15 for Model 2. Why?]

What would the regression model be if we used both **dum1** and **dum3** as the predictors?

Bumble says that he is especially interested in how Head Start compares to other preschool programs. He says that **dum1** provides a perfect way to compare Head Start to the other two programs because it takes on the value of 1 for Head Start and 0 for all other groups. He concluded from the table of coefficients that the Head Start program is significantly different from the other two programs, p=.012. Is this a correct conclusion? Explain to Bumble.

Now we show the equivalence of two-factor ANOVA and multiple regression analyses. The example is a 2x4 design with two observations in each of eight cells (CRF2,4; $n_{ij} = 2$). The data are from a hypothetical study of stress control. Factor A is Training (1=Training; 2=No Training), Factor B is Stress (1=None; 2=Low; 3=Medium; 4=High), and the dependent measure Y is the number of errors made in a data transcription task. Here are the entire syntax file and data set.

```
Title  "Demonstration of equivalence of ANOVA and MR/C".
Data list
 /A 1  B 3  Y 5-6.
Begin data
1 1  3
1 1  5
1 2  4
1 2  6
1 3  5
1 3  9
1 4  4
1 4  6
2 1  4
2 1  4
2 2  5
2 2  3
2 3  5
2 3  5
2 4 10
2 4 10
end data.
Variable labels
  A 'Training'
 /B 'Stress'
 /Y 'Errors'.
Value labels
  A  1 'Trained'  2 'No Train'
 /B  1 'None'  2 'Low'  3 'Medium'  4 'High'.
Recode A (1=1)(2=0) into ADUM.
Recode B (1=1)(2,3,4=0)(else=sysmis) into BDUM1.
Recode B (2=1)(1,3,4=0)(else=sysmis) into BDUM2.
Recode B (3=1)(1,2,4=0)(else=sysmis) into BDUM3.
Compute AB11 = ADUM*BDUM1.
Compute AB12 = ADUM*BDUM2.
Compute AB13 = ADUM*BDUM3.

ANOVA  y BY a(1 2) b(1 4).

REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS defaults change
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT y
  /METHOD=ENTER adum
  /METHOD=ENTER bdum1 bdum2 bdum3
  /METHOD=ENTER ab11 ab12 ab13  .
```

Stress Level

	None B_1	Low B_2	Medium B_3	High B_4	means
1: Training A_1	3, 5	4, 6	5, 9	4, 6	5.25
0: No training A_2	4, 4	5, 3	5, 5	10, 10	5.75
means	4.00	4.50	6.00	7.50	5.50

Note the recode of A, so training is coded 1 and no training is coded 0 for ADUM

Create three dummy variables for the four levels of B.

Create three interaction components for AxB ($df = 3$)

Hierarchical regression analysis: Enter the main effects before the interaction terms.

ANOVA[a,b]

			Unique Method				
			Sum of Squares	df	Mean Square	F	Sig.
Errors	Main Effects	(Combined)	31.000	4	7.750	3.875	.049
		Training	1.000	1	1.000	.500	.500
		Stress	30.000	3	10.000	5.000	.031
	2-Way Interactions	Training * Stress	29.000	3	9.667	4.833	.033
	Model		60.000	7	8.571	4.286	.029
	Residual		16.000	8	2.000		
	Total		76.000	15	5.067		

a. Errors by Training, Stress

b. All effects entered simultaneously

A (Training); B (Stress); AB (Training * Stress); All (Model)

SS total (Total Sum of Squares)

MSres = MS within cells
= error term from final regression model

This is a standard 2x4 between subjects ANOVA. Note the SS, df, and F test results. We will compare these results to those found with regression.

Correlations

		Errors	ADUM	BDUM1	BDUM2	BDUM3	AB11	AB12	AB13
Pearson Correlation	Errors	1.000	-.115	-.397	-.265	.132	-.260	-.087	.260
	ADUM	-.115	1.000	.000	.000	.000	.378	.378	.378
	BDUM1	-.397	.000	1.000	-.333	-.333	.655	-.218	-.218
	BDUM2	-.265	.000	-.333	1.000	-.333	-.218	.655	-.218
	BDUM3	.132	.000	-.333	-.333	1.000	-.218	-.218	.655
	AB11	-.260	.378	.655	-.218	-.218	1.000	-.143	-.143
	AB12	-.087	.378	-.218	.655	-.218	-.143	1.000	-.143
	AB13	.260	.378	-.218	-.218	.655	-.143	-.143	1.000
Sig. (1-tailed)	Errors	.	.336	.064	.161	.312	.165	.375	.165
	ADUM	.336	.	.500	.500	.500	.074	.074	.074
	BDUM1	.064	.500	.	.104	.104	.003	.208	.208
	BDUM2	.161	.500	.104	.	.104	.208	.003	.208
	BDUM3	.312	.500	.104	.104	.	.208	.208	.003
	AB11	.165	.074	.003	.208	.208	.	.299	.299
	AB12	.375	.074	.208	.003	.208	.299	.	.299
	AB13	.165	.074	.208	.208	.003	.299	.299	.
N	Errors	16	16	16	16	16	16	16	16
	ADUM	16	16	16	16	16	16	16	16
	BDUM1	16	16	16	16	16	16	16	16
	BDUM2	16	16	16	16	16	16	16	16
	BDUM3	16	16	16	16	16	16	16	16
	AB11	16	16	16	16	16	16	16	16
	AB12	16	16	16	16	16	16	16	16
	AB13	16	16	16	16	16	16	16	16

Because of equal N in each cell, the Training condition (ADUM) is orthogonal to each Stress condition, BDUM1, BDUM2, and BDUM3 (note that those correlations are zero).

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	.115[a]	.013	-.057	2.315	.013	.187	1	14	.672
2	.639[b]	.408	.193	2.023	.395	2.444	3	11	.119
3	.889[c]	.789	.605	1.414	.382	4.833	3	8	.033

a. Predictors: (Constant), Training

b. Predictors: (Constant), Training, BDUM3, BDUM2, BDUM1

c. Predictors: (Constant), Training, BDUM3, BDUM2, BDUM1, AB11, AB13, AB12

In default regression, each model uses a different error term. Only the final model uses the error term that is used by default in ANOVA. The last set of variables added are the interaction terms. The Standard Error of the Estimate is the square root of the Mean Square for the model.

ANOVA[d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.000	1	1.000	.187	.672[a]
	Residual	75.000	14	5.357		
	Total	76.000	15			
2	Regression	31.000	4	7.750	1.894	.182[b]
	Residual	45.000	11	4.091		
	Total	76.000	15			
3	Regression	60.000	7	8.571	4.286	.029[c]
	Residual	16.000	8	2.000		
	Total	76.000	15			

1: SSres = SS_B + SS_{AB} + SSw/c

Model 2 residual includes AB plus within cell error.

a. Predictors: (Constant), Training

b. Predictors: (Constant), Training, BDUM3, BDUM2, BDUM1

c. Predictors: (Constant), Training, BDUM3, BDUM2, BDUM1, AB11, AB13, AB12

d. Dependent Variable: Errors

The final model (Model 3) uses the final error term where MSerr=2.000 with df=8. The test of Model 3 is identical to the test of the overall model from the 2x4 ANOVA analysis.

In Model 1 we see that the Sum of Squares for the Training effect (A) is 1.000 with df=1. That is what we found in the 2x4 ANOVA as well. However, the error term in regression by default includes everything else, including the B effect (df=3) and the interaction (df=3) as well as the within cell error (df=8) for a total of df=14. We could run the regression analysis but compute the standard F test for ANOVA by using the final error term from Model 3 to test each of the increments in SS Regression. Thus, the increment at Step 1 = 1.000 which is $SS_{A.}$ The increment at Step 2 is SS_B = 31.000-1.000 = 30.000 which is SS_B. We can divide by df_B = 3 to get MS_B = 10 and test this with an F ratio against the error from the final step MSwc=2.000 with df=8. Similarly, the increment for the interaction = SS_{AxB} = 60.000-31.000 = 29.000 with df=7-4=3.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	5.750	.818		7.027	.000
	ADUM	-.500	1.157	-.115	-.432	.672
2	(Constant)	7.750	1.131		6.854	.000
	ADUM	-.500	1.011	-.115	-.494	.631
	BDUM1	-3.500	1.430	-.695	-2.447	.032
	BDUM2	-3.000	1.430	-.596	-2.098	.060
	BDUM3	-1.500	1.430	-.298	-1.049	.317
3	(Constant)	10.000	1.000		10.000	.000
	ADUM	-5.000	1.414	-1.147	-3.536	.008
	BDUM1	-6.000	1.414	-1.192	-4.243	.003
	BDUM2	-6.000	1.414	-1.192	-4.243	.003
	BDUM3	-5.000	1.414	-.993	-3.536	.008
	AB11	5.000	2.000	.759	2.500	.037
	AB12	6.000	2.000	.910	3.000	.017
	AB13	7.000	2.000	1.062	3.500	.008

a. Dependent Variable: Errors

Let's interpret the B coefficients. For ADUM the dummy variable compares training to no training (Factor A). The mean for the training group A_1=5.25 and the mean for the control group A_2 = 5.75. The B coefficient for A in Model 1 = -.500, the difference between the means. The test of statistical significance is a test of the equality of the two means in the population. (But be careful when selecting the error term for this test. Regular ANOVA uses the within cell error, which is comparable to the final error term from Model 3 in regression, after all factors are in the model.)

The reference group for the B factor is Group 4. In Model 2 we see that the coefficient for BDUM1 = -3.500. Checking the data, we see that the mean for Group 4 = 7.500 while the mean for Group 1 = 4.000. The difference is that Group 1 is 3.500 points lower than the reference group. The test of BDUM1 is a test of the difference between the population means for groups 1 and 4. Again, we should be thoughtful about choosing the appropriate error term.

Because of the way we calculated the interaction components, they overlap with the main effects substantially. The contribution of the interaction terms beyond the main effects are of interest, but the contribution of the main effects beyond the interactions is not generally useful. Thus, we should NOT interpret the tests of the main effects in the final model. Rather, we should interpret them at the step where they were entered into the model, which must be prior to the interactions.

The constant of 7.75 in Model 2 is the predicted value for a case that is equal to zero on all predictors, which is a case in cell A2B4 (no training, high stress), assuming no AB interaction. The mean for A2 is 5.75, which is +.25 above the grand mean of 5.50. The mean for B4 is 7.50, which is +2.00 above the grand mean. The predicted value for cell A2B4, assuming no interaction, is the grand mean plus the row effect plus the column effect: 5.50 + .25 + 2.00 = 7.75.