Qualitative IVs & Dummy Variables; F-tests for IV subsets; ANOVA Versus Regression

This handout addresses 3 questions:

(1) How can the effects of qualitative independent variables (such as race) be included in a regression analysis? Our answer will include a discussion of *dummy variables*.

(2) How can you test whether a specific subset of the ß's are equal to zero (for example, how could you test H_0: $ß_3 = ß_4 = 0$)?

(3) What is the relationship between n-way ANOVA and multiple regression analysis?

In our income example, assume that the first ten cases are black and the last 10 are white. Let X3 = 1 if the respondent is Black, 0 if White. That is, X3 is a *dummy variable* for race. To answer the above questions, we will do the following:

1. Regress Y (income) on X3 (race)

2. Regress Y (income) on X1 (education) X2 (job experience) and X3 (race)

3. Do an *incremental F-test* of whether the difference between models 1 and 2 is statistically significant; that is, test the hypothesis

   H_0: $ß_1 = ß_2 = 0$
   H_A: $ß_1$ and/or $ß_2$ do not equal 0

4. Likewise, do an *incremental F-test* of the hypothesis

   H_0: $ß_3 = 0$
   H_A: $ß_3 <> 0$

5. Use *effect coding* to show the relationship between ANOVA and multiple regression analysis.

Solution. Except for the F-tests, there is very little in the way of new statistical technique here; rather, we are using old methods in new ways. Hence, we'll let the computer do most of the work.

Part 1: Regress Y (income) on X3 (race)

```
DATA LIST FREE / Educ JobExp Income Race.
BEGIN DATA.
    2      9      5.0    1
    4     18      9.7    1
    8     21     28.4    1
    8     12      8.8    1
    8     14     21.0    1
   10     16     26.6    1
   12     16     25.4    1
   12      9     23.1    1
   12     18     22.5    1
   12      5     19.5    1
   12      7     21.7    0
   13      9     24.8    0
   14     12     30.1    0
   14     17     24.8    0
   15     19     28.5    0
   15      6     26.0    0
   16     17     38.9    0
   16      1     22.1    0
   17     10     33.1    0
   21     17     48.3    0
END DATA.
VALUE LABELS      RACE 0 ‘White’ 1 ‘Black’.
REGRESSION
            /VARIABLES EDUC JOBEXP RACE INCOME
           /DESCRIPTIVES DEF /STATISTICS DEF CHANGE TOL
            /DEPENDENT INCOME
            /METHOD ENTER RACE
           /SCATTERPLOT (INCOME RACE).
```

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
EDUC	12.0500	4.47772	20
JOBEXP	12.6500	5.46062	20
RACE	.5000	.51299	20
INCOME	24.4150	9.78835	20

Correlations

		EDUC	JOBEXP	RACE	INCOME
Pearson Correlation	EDUC	1.000	-.107	-.745	.846
	JOBEXP	-.107	1.000	.216	.268
	RACE	-.745	.216	1.000	-.568
	INCOME	.846	.268	-.568	1.000

Model Summary[b]

					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	.568[a]	.322	.284	8.27976	.322	8.554	1	18	.009

[a] Predictors: (Constant), RACE

[b] Dependent Variable: INCOME

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	586.444	1	586.444	8.554	.009[a]
	Residual	1233.981	18	68.554		
	Total	1820.425	19			

a. Predictors: (Constant), RACE

b. Dependent Variable: INCOME

Coefficients[a]

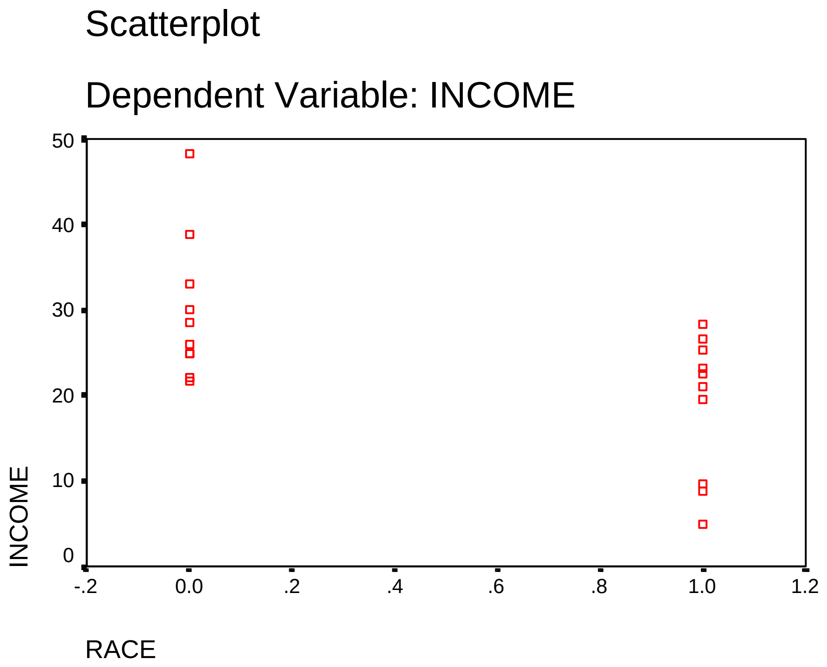
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	29.830	2.618		11.393	.000		
	RACE	-10.830	3.703	-.568	-2.925	.009	1.000	1.000

a. Dependent Variable: INCOME

According to these results, the average black makes $10,830 less than the average white. The average white makes $29,830 (do you see why?), the average black makes $19,000. This difference is significant at the .009 level.

Here is the scatterplot for race and income:

Charts



Since there are only 2 values for race, the points all fall along 2 straight lines. The regression line runs through the middle of the two lines you see here - it goes through the mean white income (29.83) and the mean black income (19). What this means is that, if we had to predict people's income, our best prediction would be the mean income for each person's racial group (i.e. if the respondent was black, we would predict their income was $19,000, if they were white we would predict $29,830). If race were unrelated to income, then our best prediction for each person would simply be the overall sample mean, i.e. we would predict $24,415 for everybody.

Part 2: Regress Y (income) on X1 (education) X2 (job experience) and X3 (race).

```
REGRESSION
           /VARIABLES EDUC JOBEXP RACE INCOME
          /DESCRIPTIVES DEF /STATISTICS DEF CHANGE TOL
           /DEPENDENT INCOME
           /METHOD ENTER RACE
           /ENTER EDUC JOBEXP.
```

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
EDUC	12.0500	4.47772	20
JOBEXP	12.6500	5.46062	20
RACE	.5000	.51299	20
INCOME	24.4150	9.78835	20

Correlations

		EDUC	JOBEXP	RACE	INCOME
Pearson Correlation	EDUC	1.000	-.107	-.745	.846
	JOBEXP	-.107	1.000	.216	.268
	RACE	-.745	.216	1.000	-.568
	INCOME	.846	.268	-.568	1.000

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	RACE[a]	.	Enter
2	JOBEXP, EDUC[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: INCOME

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	.568[a]	.322	.284	8.27976	.322	8.554	1	18	.009
2	.919[b]	.845	.816	4.19453	.523	27.068	2	16	.000

[a] Predictors: (Constant), RACE

[b] Predictors: (Constant), RACE, JOBEXP, EDUC

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	586.444	1	586.444	8.554	.009[a]
	Residual	1233.981	18	68.554		
	Total	1820.425	19			
2	Regression	1538.920	3	512.973	29.156	.000[b]
	Residual	281.505	16	17.594		
	Total	1820.425	19			

[a] Predictors: (Constant), RACE

[b] Predictors: (Constant), RACE, JOBEXP, EDUC

[c] Dependent Variable: INCOME

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	29.830	2.618		11.393	.000		
	RACE	-10.830	3.703	-.568	-2.925	.009	1.000	1.000
2	(Constant)	-7.864	5.369		-1.465	.162		
	RACE	.571	2.872	.030	.199	.845	.427	2.344
	EDUC	1.981	.323	.906	6.132	.000	.442	2.260
	JOBEXP	.642	.181	.358	3.545	.003	.947	1.056

[a] Dependent Variable: INCOME

Excluded Variables[b]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	EDUC	.950[a]	4.974	.000	.770	.445	2.245	.445
	JOBEXP	.409[a]	2.290	.035	.486	.953	1.049	.953

[a] Predictors in the Model: (Constant), RACE

[b] Dependent Variable: INCOME

(Note that, by using multiple ENTER lines on the same regression card, SPSS will give us the model results in sequence.) When RACE, EDUC and JOBEXP are all in the model, both job experience and education significantly affect income. However, the effect of race is NOT statistically significant. This suggests that much of the apparent effect of race in the previous model may be due to the association of race with education and job experience. Note the high correlation, -.745, between race and education; also note the low tolerance for race, which shows that it is highly correlated with education and job experience ($r_{321}^2 = 1 - Tol_3 = .573378$). (Education also has a low tolerance, but its effects are so strong that they remain highly significant even when race is included in the model.) This need not mean that race is irrelevant for determining income. Perhaps the effect of race on income is indirect. Race may affect education and job experience, which in turn affect income.

Part 3: Do an *Incremental F-test* of whether the difference between models 1 and 2 is statistically significant; that is, test the hypothesis

H_0: $ß_1 = ß_2 = 0$
H_A: $ß_1$ and/or $ß_2$ do not equal 0

Let SSE_u refer to the sum of squared errors for the “unconstrained model”, i.e. the model in which effects are estimated for X1 (education), X2 (job experience), and X3 (race). From part 2, we see that $SSE_u = 281.50528$. Now, let SSE_c refer to the sum of squared errors from the “constrained” model, i.e. the model in which race alone is included (we call it the “constrained” model because, by not including Education and Job experience in the model, we in effect impose the constraint that $ß_1 = ß_2 = 0$). From part 1, we see that $SSE_c = 1233.981$. The key question is, does adding X1 and X2 to the model significantly improve the fit - that is, are the residuals made significantly smaller by including X1 and X2? If so, SSE will decline. To test the hypothesis,

H_0: $ß_1 = ß_2 = 0$
H_A: $ß_1$ and/or $ß_2$ do not equal 0

the appropriate test statistic is

$$F_{J,N-K-1} = \frac{(SSE_c - SSE_u)/J}{SSE_u/(N-K-1)} = \frac{MSE_{c-u}}{MSE_u}$$

$$= \frac{(SSE_c - SSE_u)*(N-K-1)}{SSE_u * J}$$

$$= \frac{(R_u^2 - R_c^2)*(N-K-1)}{(1-R_u^2)*J}$$

where J = number of constraints, in this case, the number of parameters hypothesized to equal 0, and K is the number of variables in the *unconstrained* model. (Note that the last formula is obtained by dividing both the numerator and denominator of the immediately prior formula by

SST_y.) Put another way, J = the error d.f. for the constrained model (18) minus the error d.f. for the unconstrained model (16). In this problem, J = 2. Hence,

$$F_{J,N-K-1} = \frac{(SSE_c - SSE_u)/J}{SSE_u/(N-K-1)} = \frac{(1233.981 - 281.50528)/2}{281.50528/16} = \frac{476.24}{17.59} = 27.07$$

$$= \frac{(R_u^2 - R_c^2) * (N-K-1)}{(1 - R_u^2) * J} = \frac{(.84536 - .32215) * 16}{(1 - .84536) * 2} = \frac{8.37136}{.30928} = 27.07$$

$$= \frac{R_{change}^2 * (N-K-1)}{(1 - R_u^2) * J} = \frac{.52321 * 16}{.30928} = 27.07$$

(Compare this with the "F Change" and the "R Square Change" reported in the Spss printout.) Thus, we reject the null hypothesis.

NOTE: The above is referred to as an *incremental F test.* Contrast this with the *Global F Test*, where we test to see whether all the Betas in an equation equal 0.

When you can use incremental F. In order to use the incremental F test, it must be the case that

- One model is "nested" within the other; that is, one model must be a constrained, or special case, of the other. For example, if one model contains IVs X1-X3, and another model contains X1 only, the latter is a special case of the former, where the constraints imposed are $\beta 2 = \beta 3 = 0$. If, however, the second model included X1 and X6, it would not be nested within the first model and an incremental F test would not be appropriate.
- The sample is the same for each model estimated. This assumption might be violated if, say, missing data in variables used in the unconstrained model caused the unconstrained sample to be smaller than the constrained sample. You should be careful how missing data is getting handled in your statistical routines

Other comments

- Constrained and unconstrained are relative terms. An unconstrained model in one analysis can be the constrained model in another. In reality, every model is "constrained' in the sense that more variables could always be added to it.
- Other types of constraints can also be tested with an incremental F test. For example, we might want to test the hypothesis that $\beta 1 = \beta 2$, i.e. two variables have equal effects. We'll discuss such possibilities 2nd semester.
- There are alternatives to the incremental F test; in particular, Stata makes it easy to do what is known as a *Wald Test*, which does not require that you estimate both the constrained and unconstrained models. When software supports them, Wald tests are easy but not always optimal for hypothesis testing. We will talk about Wald tests and other alternative approaches in Stats II.

Part 4: Likewise, do an F-test of the hypothesis

H_0: ß$_3$ = 0
H_A: ß$_3$ <> 0

The procedure is the same as in part three - except that this time the constrained model is the model in which X1 and X2 are included, and X3 is excluded. We estimated this model in the Multiple Regression - Introduction handout. The exact value for SSE_c is 282.20025. Hence, the F-test is

$$F_{1,16} = \frac{(282.20025 - 281.50528)/1}{281.50528/16} = \frac{0.695}{17.594} = .0395$$

Incidentally, note that the T-value for Race reported in Part 2 was .199, and that $.199^2 = .0395$, the same as the F-value we just got. Hence, both the T-test and the F-test of the hypothesis

H_0: ß$_3$ = 0
H_A: ß$_3$ <> 0

yield the same results. Likewise, an F test of

H_0: ß$_1$ = 0
H_A: ß$_1$ <> 0

would equal t^2, i.e. 6.132^2.

Incidentally, SPSS can provide you with all possible F-tests of interest, by using the TEST parameter on the Regression card. As the SPSS manual explains,

TEST (varlist) (varlist) *R^2 change and its significance for sets of independent variables.* This method first adds all variables specified on TEST to the current equation. It then removes in turn each subset from the equation and displays requested statistics. Specify test subsets in parentheses. A variable can be used in more than one subset, and each subset can include any number of variables. Variables named on TEST remain in the equation when the method is completed.

In this case, the F-tests are:

```
REGRESSION
           /VARIABLES EDUC JOBEXP RACE INCOME
          /DESCRIPTIVES DEF /STATISTICS DEF CHANGE TOL
           /DEPENDENT INCOME
            /Method=TEST (EDUC) (JOBEXP) (RACE)
                 (EDUC JOBEXP) (EDUC RACE) (JOBEXP RACE).
```

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	RACE, JOBEXP, EDUC	.	Test

[a] Dependent Variable: INCOME

ANOVA[c]

Model			Sum of Squares	df	Mean Square	F	Sig.	R Square Change
1	Subset Tests	EDUC	661.469	1	661.469	37.596	.000[a]	.363
		JOBEXP	221.054	1	221.054	12.564	.003[a]	.121
		RACE	.695	1	.695	.040	.845[a]	.000
		EDUC, JOBEXP	952.476	2	476.238	27.068	.000[a]	.523
		EDUC, RACE	1408.425	2	704.212	40.026	.000[a]	.774
		JOBEXP, RACE	236.867	2	118.433	6.731	.008[a]	.130
	Regression		1538.920	3	512.973	29.156	.000[b]	
	Residual		281.505	16	17.594			
	Total		1820.425	19				

[a] Tested against the full model.

[b] Predictors in the Full Model: (Constant), RACE, JOBEXP, EDUC.

[c] Dependent Variable: INCOME

Additional Comments on *Dummy Variables* & *Incremental F Tests*

(1) We frequently want to examine the effects of both quantitative and qualitative independent variables on quantitative dependent variables. *Dummy variables* provide a means by which qualitative variables can be included in regression analysis. The procedure for computing dummy variables is as follows:

(a) Suppose there are L groups. You will compute L-1 dummy variables. In the present example, L = 2, since you have two groups, whites and blacks. Hence, one dummy variable was computed. If we had 3 groups (for example, white, black, and other) we would construct 2 dummy variables.

(b) The Lth group is coded 0 on every dummy variable. We refer to this as the “excluded category.” In this case, white was the excluded category, and whites were coded 0 on the 1 dummy variable.

(c) The first group is coded 1 on the first dummy variable. The other L-2 groups (i.e. groups other than the first and the Lth) are coded 0. On the second dummy variable (if there is one), the second group is coded 1, and the other L-2 groups are coded zero. Repeat this procedure for each of the L-1 dummy variables.

For example, suppose our categories were white, black, and other, and we wanted white to be the excluded category. Then,

Dummy1 = 1 if black, 0 if other, 0 if white
Dummy2 = 0 if black, 1 if other, 0 if white

Incidentally, note that if we wanted to compute it, Dummy3 = 1 - Dummy1 - Dummy2. We do not include Dummy3 in our regression models, because if we did, we would run into a situation of perfect collinearity.

(2) An alternative to dummy variable coding is *effect coding*. Computational procedures are the same, except that, for step b, the excluded category is coded -1 on every effect variable. Hence, if our categories were white, black, and other, the effect variables would be coded as

Effect1 = 1 if black, 0 if other, -1 if white
Effect2 = 0 if black, 1 if other, -1 if white

(3) Dummy variable coding and effect coding yield algebraically equivalent results; that is, you get the same R^2, same F values, etc. The estimates of the ß's differ, but you can easily convert parameters obtained using dummy variable coding to parameters obtained using effect coding.

(4) Dummy variable coding is probably most commonly used. However, as you hopefully discovered in your homework, effect coding provides a means by which 1-way analysis of variance problems can be addressed using multiple regression. It can be shown that n-way analysis of variance is merely a special case of multiple regression analysis, and both fall under the heading of the "general linear model".

(5) The incremental F-test used here is useful when you want to examine the effects of 2 or more dummy variables. However, as I have illustrated here, you can use this F test in other situations where constraints are being imposed on some of the parameters. Besides testing whether one or more parameters equal zero, it is also fairly common to test whether two or more parameters are equal, or whether one or more parameters equal some specific value or values. The procedures by which you impose such constraints will be discussed in Stats II.

Part 5. Use *effect coding* to show the relationship between ANOVA and multiple regression analysis.

Let us rework Part 1, only this time X3 is coded 1 if black, -1 if white. The results are:

```
* Now use FFECT Coding.
RECODE  RACE (1 = 1)(0 = -1).
VALUE LABELS      RACE -1 ‘White’ 1 ‘Black’.
REGRESSION
           /VARIABLES EDUC JOBEXP RACE INCOME
           /DESCRIPTIVES DEF /STATISTICS DEF CHANGE TOL
           /DEPENDENT INCOME
           /METHOD ENTER RACE.
```

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
EDUC	12.0500	4.47772	20
JOBEXP	12.6500	5.46062	20
RACE	.0000	1.02598	20
INCOME	24.4150	9.78835	20

Correlations

		EDUC	JOBEXP	RACE	INCOME
Pearson Correlation	EDUC	1.000	-.107	-.745	.846
	JOBEXP	-.107	1.000	.216	.268
	RACE	-.745	.216	1.000	-.568
	INCOME	.846	.268	-.568	1.000

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	.568[a]	.322	.284	8.27976	.322	8.554	1	18	.009

a. Predictors: (Constant), RACE

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	586.444	1	586.444	8.554	.009[a]
	Residual	1233.981	18	68.554		
	Total	1820.425	19			

a. Predictors: (Constant), RACE

b. Dependent Variable: INCOME

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	24.415	1.851		13.187	.000		
	RACE	-5.415	1.851	-.568	-2.925	.009	1.000	1.000

a. Dependent Variable: INCOME

Notice that:

(1) R^2 is the same, F is the same, the value of the T-test for race is the same regardless of whether effect coding or dummy variable coding is used. In fact, everything is the same, except for the columns labeled B and SE B, and the T value for the constant. (SIG T would also differ for the constant if we carried it out enough decimal places.)

(2) The constant is equal to the mean of income. (This is only true when there are equal numbers in each group however.)

(3) The coefficient of -5.415 for race implies that blacks (who are coded 1) have an average income that is $5,415 below the mean (i.e., $19,000). Whites (who are coded -1) have an average income that is $5,415 above the mean (i.e. $29,830). Hence, the difference between the average income of blacks and whites is 2 * $5,415 = $10,830, which is the same conclusion we reached before.

We will now approach this problem using 1-way analysis of variance. The results from the SPSS program ANOVA are as follows:

```
RECODE  RACE (-1 = 1)(1 = 2).
VALUE LABELS      RACE 1 ‘White’ 2 ‘Black’.
ANOVA   INCOME BY RACE (1,2)/Method = Experimental/ STAT MCA.
```

ANOVA

ANOVA[a]

			Experimental Method				
			Sum of Squares	df	Mean Square	F	Sig.
INCOME	Main Effects	RACE	586.445	1	586.445	8.554	.009
	Model		586.445	1	586.445	8.554	.009
	Residual		1233.981	18	68.554		
	Total		1820.425	19	95.812		

a. INCOME by RACE

MCA[a]

				Predicted Mean		Deviation	
			N	Unadjusted	Adjusted for Factors	Unadjusted	Adjusted for Factors
INCOME	RACE	1.00 White	10	29.8300	29.8300	5.4150	5.4150
		2.00 Black	10	19.0000	19.0000	-5.4150	-5.4150

a. INCOME by RACE

Model Goodness of Fit

	R	R Squared
INCOME by RACE	.568	.322

Note that everything here is basically the same as in the regression analysis.