Chapter 15
Regression with Categorical Predictor Variables

Regression with Categorical Predictor Variables

1. Overview of regression with categorical predictors

- Thus far, we have considered the OLS regression model with continuous predictor and continuous outcome variables. In the regression model, there are no distributional assumptions regarding the shape of X; Thus, it is not *necessary* for X to be a continuous variable.
- In this section we will consider regression models with a single categorical predictor and a continuous outcome variable.
    - These analyses could also be conducted in an ANOVA framework. We will explore the relationship between ANOVA and regression.

- The big issue regarding categorical predictor variables is how to represent a categorical predictor in a regression equation. Consider an example of the relationship between religion and attitudes toward abortion. In your dataset, you have religion coded categorically. A couple of problems immediately arise:
    - Because religion is not quantitative, there is not a unique coding scheme. Coding scheme A and coding scheme B are both valid ways to code religion – we need to make sure that our results are not dependent on how we have coded the categorical predictor variable.

Coding A

Religion	Code
Catholic	1
Protestant	2
Jewish	3
Other	4

Coding B

Religion	Code
Protestant	1
Jewish	2
Catholic	3
Other	4

    - Even if we solve the coding problem (say we could get all researchers to agree on coding scheme A), the regression model estimates a linear relationship between the predictor variable and the outcome variable.

$$Y = b_0 + b_1 X$$

$$AttitudesTowardAbortion = b_0 + b_1(\text{Re}ligion)$$

- Consider the interpretation of b_1: A one-unit increase in religion is associated with a b_1 using increase in attitudes toward abortion.
- But what is a one-unit increase in religion!?!
- We need to consider alternative methods of coding for categorical predictor variables

- We will consider three ways to code categorical predictor variables for regression:
  - Dummy coding
  - Effects coding
  - Contrast coding

  What all these methods have in common is that for a categorical predictor variable with *a* levels, we code it into *a*-1 different indicator variables. All *a*-1 indicator variables that we create must be entered into the regression equation.

2. Dummy coding
   - For dummy coding, one group is specified to be the reference group and is given a value of 0 for each of the (*a*-1) indicator variables.

Dummy Coding of Gender (a = 2)

Gender	D_1
Male	1
Female	0

Dummy Coding of Treatment Groups (a = 3)

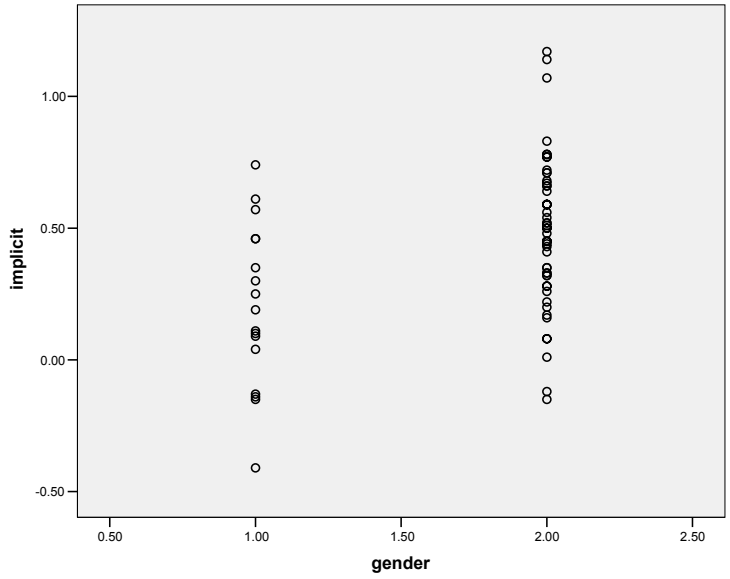
Group	D_1	D_2
Treatment 1	0	1
Treatment 2	1	0
Control	0	0

Dummy Coding of Religion (a = 4)

Religion	D_1	D_2	D_3
Protestant	0	0	0
Catholic	1	0	0
Jewish	0	1	0
Other	0	0	1

- The choice of the reference group is statistically arbitrary, but it affects how you interpret the resulting regression parameters. Here are some considerations that should guide your choice of reference group (Hardy, 1993):
    - The reference group should serve as a useful comparison (e.g., a control group; a standard treatment; or the group expected to have the highest/lowest score).
    - The reference group should be a meaningful category (e.g., not an *other* category).
    - If the sample sizes in each group are unequal, it is best if the reference group not have a small sample size relative to the other groups.

- Dummy coding a dichotomous variable
    - We wish to examine whether gender predicts level of implicit self-esteem (as measured by a Single Category Implicit Association Test). Implicit self-esteem data are obtained from a sample of women (n = 56) and men (n = 17).



    - In the data gender is coded with male = 1 and female = 2.

    - For a dummy coded indicator variable, we need to recode the variable, Let's use women as the reference group (imagine we live in a gynocentric world).

Dummy Coding of Gender (a = 2)

Gender	D_1
Male	1
Female	0

IF (gender = 2) dummy = 0.
IF (gender = 1) dummy = 1.

- Now, we can predict implicit self esteem from the dummy-coded gender variable in an OLS regression.

$$Implicit\ \ Self-Esteem = b_0 + b_1 * Dummy$$

- Using this equation, we can obtain separate regression lines for women and men by substituting appropriate values for the dummy variable.

For women: Dummy = 0

$$Implicit\ \ Self-Esteem = b_0 + b_1 * 0$$
$$= b_0$$

For men: Dummy = 1

$$Implicit\ \ Self-Esteem = b_0 + b_1 * 1$$
$$= b_0 + b_1$$

- Interpreting the parameters:

- b_0 = The average self-esteem of women (the reference group)

  The test of b_0 tells us whether the mean score on the outcome variable of the reference group differs from zero.

- b_1 = The difference in self-esteem between women and men

  The test of b_1 tells us whether the mean score on the outcome variable differs between the reference group and the alternative group.

- If we wanted a test of whether the average self-esteem of men differed from zero, we could re-run the analysis with men as the reference group.

- Interpreting other regression output
  - The Pearson correlation between D_1 and Y, r_{D_1Y}, is the point biserial correation between gender (male vs. female) and Y.
  - $R^2 = r^2_{D_1Y}$ is the percentage of variance (of the outcome variable) that can be accounted for by the female/male dichotomy.

- Running the analysis in SPSS

```
REGRESSION
  /STATISTICS COEFF OUTS R ANOVA ZPP
  /DEPENDENT implicit
  /METHOD=ENTER dummy.
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.404[a]	.163	.151	.28264

a. Predictors: (Constant), dummy

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	.493	.038		13.059	.000			
	dummy	-.291	.078	-.404	-3.716	.000	-.404	-.404	-.404

a. Dependent Variable: implicit

$$Implicit\ \ Self-Esteem = .493 + (-.291) * Dummy$$

- The test of b_0 indicates that women have more positive than negative associations with the self (high self-esteem), $b = .49, t(71) = 13.06, p < .01$. Also, we know $\overline{Y}_{Women} = 0.49$.
- The test of b_1 indicates that men's self esteem differs from women's self-esteem by -0.29 units (that is, they have lower self-esteem), $b = -.29, t(71) = 3.72, p < .01$.
- By inference, we know that the average self-esteem of men is $b_0 + b_1 = .493 - .291 = .202$. However, with this dummy coding, we do not obtain a test of whether or not the mean score for men differs from zero.

- The female/male distinction accounts for 16% of the variance in implicit self-esteem scores, $R^2 = .16$.

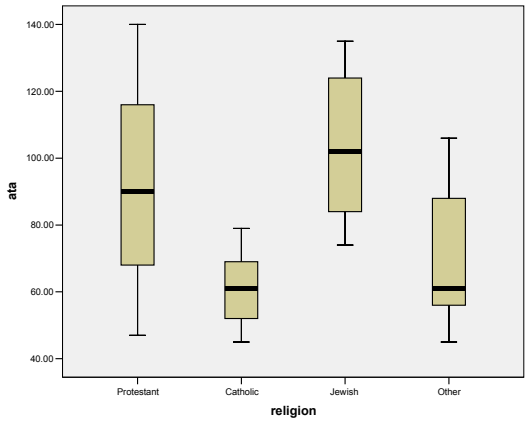
- Confirming the results in SPSS

```
EXAMINE VARIABLES=implicit BY gender.
```

Descriptives

	gender		Statistic	Std. Error
implicit	Male	Mean	.2024	.07529
	Female	Mean	.4932	.03662

- Dummy coding a categorical variable with more than 2 levels
  - Let's return to our example of the relationship between religion and attitudes toward abortion. We obtain data from 36 individuals (Protestant, $n = 13$; Catholic, $n = 9$; Jewish, $n = 6$; Other, $n = 8$).



  - Because religion has four levels, we need to create 3 dummy variables. We have a choice of four possible reference groups:

Reference Group = Protestant

Religion	D_1	D_2	D_3
Protestant	0	0	0
Catholic	1	0	0
Jewish	0	1	0
Other	0	0	1

Reference Group = Jewish

Religion	D_1	D_2	D_3
Protestant	1	0	0
Catholic	0	1	0
Jewish	0	0	0
Other	0	0	1

Reference Group = Catholic

Religion	D_1	D_2	D_3
Protestant	1	0	0
Catholic	0	0	0
Jewish	0	1	0
Other	0	0	1

Reference Group = Other

Religion	D_1	D_2	D_3
Protestant	1	0	0
Catholic	0	1	0
Jewish	0	0	1
Other	0	0	0

- For this example, we will use Protestant as the reference group.

```
IF (religion = 2) dummy1 = 1.
IF (religion ne 2) dummy1 = 0.

IF (religion = 3) dummy2 = 1.
IF (religion ne 3) dummy2 = 0.

IF (religion = 4) dummy3 = 1.
IF (religion ne 4) dummy3 = 0.
```

  - When the categorical variable has more than two levels (meaning that more than 1 dummy variable is required), it is essential that all the dummy variables be entered into the regression equation.

$$ATA = b_0 + (b_1 * D_1) + (b_2 * D_2) + (b_3 * D_3)$$

  - Using this equation, we can obtain separate regression lines for each religion by substituting appropriate values for the dummy variables.

Reference Group = Protestant

Religion	D_1	D_2	D_3
Protestant	0	0	0
Catholic	1	0	0
Jewish	0	1	0
Other	0	0	1

For Protestant: $D_1 = 0; D_2 = 0; D_3 = 0$

$$\begin{aligned} ATA &= b_0 + (b_1 * 0) + (b_2 * 0) + (b_3 * 0) \\ &= b_0 \end{aligned}$$

For Catholic: $D_1 = 1; D_2 = 0; D_3 = 0$

$$\begin{aligned} ATA &= b_0 + (b_1 * 1) + (b_2 * 0) + (b_3 * 0) \\ &= b_0 + b_1 \end{aligned}$$

For Jewish: $D_1 = 0; D_2 = 1; D_3 = 0$

$$\begin{aligned} ATA &= b_0 + (b_1 * 0) + (b_2 * 1) + (b_3 * 0) \\ &= b_0 + b_2 \end{aligned}$$

For Other: $D_1 = 0; D_2 = 0; D_3 = 1$

$$\begin{aligned} ATA &= b_0 + (b_1 * 0) + (b_2 * 0) + (b_3 * 1) \\ &= b_0 + b_3 \end{aligned}$$

- Interpreting the parameters:
  - b_0 = The average ATA of Protestants (the reference group)

    The test of b_0 tells us whether the mean score on the outcome variable for the reference group differs from zero.
  - b_1 = The difference in ATA between Protestants and Catholics

    The test of b_1 tells us whether the mean score on the outcome variable differs between the reference group and the group identified by D_1.
  - b_2 = The difference in ATA between Protestants and Jews

    The test of b_2 tells us whether the mean score on the outcome variable differs between the reference group and the group identified by D_2.
  - b_3 = The difference in ATA between Protestants and Others

    The test of b_3 tells us whether the mean score on the outcome variable differs between the reference group and the group identified by D_3.
  - If we wanted a test of whether the ATA of Catholics, Jews, or others differed from zero, we could re-run the analysis with those groups as the reference group. Likewise if we wanted to test for differences in attitudes between Catholics and Jews, we could reparameterize the model with either Catholics or Jews as the reference group.

- Interpreting Pearson (zero-order) correlation coefficients:
  - The Pearson correlation between D_1 and Y, r_{D_1Y}, is the point biserial correlation between the Catholic/non-Catholic dichotomy and Y.
  - $r^2_{D_1Y}$ is the percentage of variance (of the outcome variable) that can be accounted for my the Catholic/non-Catholic dichotomy
  - The Pearson correlation between D_2 and Y, r_{D_2Y}, is the point biserial correlation between the Jewish/non-Jewish dichotomy and Y.
  - $r^2_{D_2Y}$ is the percentage of variance (of the outcome variable) that can be accounted for my the Jewish/non-Jewish dichotomy
  - The Pearson correlation between D_3 and Y, r_{D_3Y}, is the point biserial correlation between the Other/non-Other dichotomy and Y.
  - $r^2_{D_3Y}$ is the percentage of variance (of the outcome variable) that can be accounted for my the Other/non-Other dichotomy
  - R^2 is the percentage of variance in Y (ATA) in the sample that is associated with religion. $R^2_{Adjusted}$ is the percentage of Y (ATA) variance accounted for by religion in the population

- Running the analysis in SPSS

```
REGRESSION
  /STATISTICS COEFF OUTS R ANOVA ZPP
  /DEPENDENT ATA
  /METHOD=ENTER dummy1 dummy2 dummy3.
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.596[a]	.355	.294	23.41817

a. Predictors: (Constant), dummy3, dummy2, dummy1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	93.308	6.495		14.366	.000			
	dummy1	-32.641	10.155	-.514	-3.214	.003	-.442	-.494	-.456
	dummy2	10.192	11.558	.138	.882	.384	.355	.154	.125
	dummy3	-23.183	10.523	-.351	-2.203	.035	-.225	-.363	-.313

a. Dependent Variable: ata

$$ATA = 93.308 + (-32.641 * D_1) + (10.192 * D_2) + (-23.183 * D_3)$$

- The test of b_0 indicates that the mean ATA score for Protestants, $\overline{Y}_{Protestants} = 93.31$, is significantly above zero, $b = 93.31, t(32) = 14.37, p < .01$. In this case, b_0 is probably meaningless (the scale responses have to be greater than zero).

- The test of b_1 indicates that the mean ATA score for Catholics is significantly less than (because the sign is negative) the mean ATA score for Protestants, $b = -32.64, t(32) = 3.21, p < .01$. The mean ATA score for Catholics is $\overline{Y}_{Catholics} = b_0 + b_1 = 93.31 - 32.64 = 60.67$. The Catholic/non-Catholic distinction accounts for 19.5% of the variance in ATA scores ($r^2_{D_1Y} = .442^2 = .195$).

- The test of b_2 indicates that the mean ATA score for Jews is not significantly different from the mean ATA score for Protestants, $b = 10.19, t(32) = 0.88, p = .38$. The mean ATA score for Jews is $\overline{Y}_{Jews} = b_0 + b_2 = 93.31 + 10.19 = 103.50$. The Jew/non-Jew distinction accounts for 19.5% of the variance in ATA scores ($r^2_{D_2Y} = .355^2 = .126$).

- The test of b_3 indicates that the mean ATA score for Others is significantly lower than the mean ATA score for Protestants, $b = -23.18, t(32) = -2.20, p = .04$. The mean ATA score for Others is $\bar{Y}_{Others} = b_0 + b_3 = 93.31 - 23.18 = 70.13$. The Other/non-Other distinction accounts for 5.1% of the variance in ATA scores ($r^2_{D_3Y} = -.225^2 = .051$)

- Overall 35.5% of the variability in ATA scores in the sample is associated with religion; 29.4% of the variability in ATA scores in the population is associated with religion.

  Note that the dummy variables are not orthogonal to each other. As a result, the model R^2 does not equal (and in fact must be less than) the sum of the variance accounted for by each dummy variable.

$$R^2 < r^2_{D_1Y} + r^2_{D_2Y} + r^2_{D_3Y}$$
$$.355 < .195 + .126 + .051 = .372$$

- If we want other pairwise comparisons, we need to re-parameterize the model and run another regression. For example, to compare each group to the mean response of Jewish respondents, we need Jewish respondents to be the reference category.

Reference Group = Jewish

Religion	D_1	D_2	D_3
Protestant	1	0	0
Catholic	0	1	0
Jewish	0	0	0
Other	0	0	1

- Cautions about dummy coding

- In some dummyesque coding systems, people use 1 or 2 coding or 0 or 2 coding rather than a 0 or 1 coding. You should not do this – it changes the interpretation of the model parameters.
- We have (implicitly) assumed that the groups are mutually exclusive, but in some cases, the groups may not be mutually exclusive. For example, a bi-racial individual may indicate more than one ethnicity. This, too, affects the model parameters and extreme care must be taken to avoid erroneous conclusions.

- Comparing alternative parameterization of the model.

  - Omitting all the details, let's compare the four possible dummy code parameterizations of the model.

Reference Group		*b*	*β*	*p*	*r*	*Model* R^2
Protestant	b_0	93.31		< .001		
	Dummy 1	-32.64	-.51	.003	-.44	
	Dummy 2	10.19	.14	.384	.36	.355
	Dummy 3	-23.18	-.35	.035	-.23	
Catholic	b_0	60.67		< .001		
	Dummy 1	32.64	.57	.003	.32	
	Dummy 2	42.83	.58	.002	.36	.355
	Dummy 3	9.46	.14	.412	-.23	
Jewish	b_0	103.50		< .001		
	Dummy 1	-42.83	-.68	.002	-.44	
	Dummy 2	-10.19	-.18	.384	.32	.355
	Dummy 3	-33.38	-.51	.013	-.23	
Other	b_0	70.13		< .001		
	Dummy 1	-9.46	-.15	.412	-.44	
	Dummy 2	23.18	.41	.035	.32	.355
	Dummy 3	33.38	.45	.013	.36	

  - Note that model parameters, p-values, and correlations are all different.
  - In all cases, it is the unstandardized regression coefficients that have meaning. We should not interpret or report standardized regression coefficients for dummy code analyses (This general rule extends to all categorical variable coding systems).
  - So long as the a-1 indicator variables are entered into the regression equations, the Model R^2 is the same regardless of how the model is parameterized.

3. (Unweighted) Effects coding

- Dummy coding allows us to test for differences between levels of a (categorical) predictor variable. In some cases, the main question of interest is whether or not the mean of a specific group differs from the overall sample mean. Effects coding allow us to test these types of hypotheses.
- These indicator variables are called "effects codes" because they reflect the treatment effect (think α terms in ANOVA).
- For effects coded indicator variables, one group is specified to be the base group and is given a value of -1 for each of the (a-1) indicator variables.

Effects Coding of Gender (a = 2)

Gender	E_1
Male	1
Female	-1

Effects Coding of Treatment Groups (a = 3)

Group	E_1	E_2
Treatment 1	0	1
Treatment 2	1	0
Control	-1	-1

Effects Coding of Religion (a = 4)

Religion	E_1	E_2	E_3
Protestant	-1	-1	-1
Catholic	1	0	0
Jewish	0	1	0
Other	0	0	1

- Once again, the choice of the base group is statistically arbitrary, but it affects how you interpret the resulting regression parameters. In contrast to dummy coding, the base group is often the group of *least* interest because the regression analysis does not directly inform us about the base group.
  - For each of the other groups, the effects coded parameters inform us about the difference between the mean of each group and the grand mean.

- Effects coding for a dichotomous variable
  - Again, let's use women as the reference group:

```
IF (gender = 2) effect1 = -1.
IF (gender = 1) effect1 = 1.
```

Effects Coding of Gender (a = 2)

Gender	E_1
Male	1
Female	-1

  - Now, we can predict implicit self esteem from the effects coded gender variable in an OLS regression

$$Implicit\ \ Self - Esteem = b_0 + b_1 * Effect1$$

  - Using this equation, we can get separate regression lines for women and men by substituting appropriate values for the effects coded variable.

For women: *Effect1* = -1

$$Implicit\ \ Self - Esteem = b_0 - b_1$$

For men: *Effect1* = 1

$$Implicit\ \ Self - Esteem = b_0 + b_1$$

  - Interpreting the parameters:
    - b_0 = The average self-esteem of all the group means

      The test of b_0 tells us whether the grand mean (calculated as the average of all the group means) on the outcome variable of the reference group differs from zero.

      If the sample sizes in each group are equivalent, then b_0 is the grand mean..
    - b_1 = The difference between men's average self-esteem and the mean level of self-esteem.

      The test of b_1 tells us whether the mean score for the group coded 1 differs from the grand mean (the calculated as the average of all the group means).

      In an ANOVA framework, we would call the group effect for men, α_{Men}.

- Interpreting other regression output:
  - When $a = 2$, the Pearson correlation between E_1 and Y, r_{E_1Y}, is the point biserial correlation between gender (male vs. female) and Y. When $a > 2$, the interpretation of r_{E_1Y} is ambiguous.
  - When $a = 2$, $r^2_{E_1Y}$ is the percentage of variance (of the outcome variable) that can be accounted for by the female/male dichotomy, When $a > 2$, the interpretation of $r^2_{E_1Y}$ is ambiguous.

- Running the analysis in SPSS

```
REGRESSION
  /STATISTICS COEFF OUTS R ANOVA ZPP
  /DEPENDENT implicit
  /METHOD=ENTER effect1.
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.404[a]	.163	.151	.28264

a. Predictors: (Constant), effect1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations		
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part
1	(Constant)	.348	.039		8.887	.000			
	effect1	-.145	.039	-.404	-3.716	.000	-.404	-.404	-.404

a. Dependent Variable: implicit

$$Implicit\ \ Self - Esteem = .348 + .145 * Effect1$$

- The test of b_0 indicates that the average self-esteem score (the average of men's self-esteem and of women's self-esteem) is greater than zero, $b = .35, t(71) = 8.89, p < .01$.
- The test of b_1 indicates that men's self-esteem differs from the average self-esteem score by -0.145 units (that is, they have lower self-esteem), $b = -.15, t(71) = 3.72, p < .01$.
- Thus, we know that $\overline{Y}_{Men} = .348 - .145 = .203$ and that $\overline{Y}_{Women} = .348 + .145 = .493$.

- The female/male distinction accounts for 16% of the variance in implicit self-esteem scores, $r^2_{D_1Y} = .16$.

- Note that with unequal sample sizes, the "grand mean" is the average of the group means $\bar{Y}_{Unweighted} = .2024 + .4932 = .348 = b_0$ and is not the grand mean of all N observations $\bar{Y} = .426$.

Descriptives

implicit

	N	Mean	Std. Deviation
Male	17	.2024	.31041
Female	56	.4932	.27403
Total	73	.4255	.30676

- Previously, we called this approach the *unique* Sums of Squares (or *Type III* SS approach to unbalanced data). Sometimes this approach is also called the *regression* approach to unbalanced data. This is the default/favored approach to analyzing unbalanced data (see pp. 9-6 to 9-13).
- When you have unbalanced designs, be careful about interpreting effects coded variables!

- Effect coding a categorical variable with more than 2 levels
  - Let's (again) return to our example of the relationship between religion and attitudes toward abortion.

  - Because religion has four levels, we need to create 4 effects coded variables. We have a choice of four possible base levels:

Reference Group = Protestant

Religion	E_1	E_2	E_3
Protestant	-1	-1	-1
Catholic	1	0	0
Jewish	0	1	0
Other	0	0	1

Reference Group = Jewish

Religion	E_1	E_2	E_3
Protestant	1	0	0
Catholic	0	1	0
Jewish	-1	-1	-1
Other	0	0	1

Reference Group = Catholic

Religion	E_1	E_2	E_3
Protestant	1	0	0
Catholic	-1	-1	-1
Jewish	0	1	0
Other	0	0	1

Reference Group = Other

Religion	E_1	E_2	E_3
Protestant	1	0	0
Catholic	0	1	0
Jewish	0	0	1
Other	-1	-1	-1

- For this example, we will use Protestant as the base group. In practice, it would probably be better to use Other as the base group, but for the purposes of comparing effects coding output to dummy coding output, we will stick with Protestant as the base group.

```
IF (religion = 1) effect1 = -1 .
IF (religion = 2) effect1 = 1 .
IF (religion = 3) effect1 = 0 .
IF (religion = 4) effect1 = 0 .
```

```
IF (religion = 1) effect2 = -1 .
IF (religion = 2) effect2 = 0 .
IF (religion = 3) effect2 = 1 .
IF (religion = 4) effect2 = 0 .
```

```
IF (religion = 1) effect3 = -1 .
IF (religion = 2) effect3 = 0 .
IF (religion = 3) effect3 = 0 .
IF (religion = 4) effect3 = 1 .
```

  - As with dummy variables, it is essential that all the effect coded variables be entered into the regression equation.

$$ATA = b_0 + (b_1 * E_1) + (b_2 * E_2) + (b_3 * E_3)$$

  - Using this equation, we can get separate regression lines for each religion by substituting appropriate values for the effect coded variables.

Reference Group = Protestant

Religion	E_1	E_2	E_3
Protestant	-1	-1	-1
Catholic	1	0	0
Jewish	0	1	0
Other	0	0	1

For Protestant: $E_1 = -1; E_2 = -1; E_3 = -1$

$$\begin{aligned} ATA &= b_0 + (b_1 * -1) + (b_2 * -1) + (b_3 * -1) \\ &= b_0 - (b_1 + b_2 + b_3) \end{aligned}$$

For Jewish: $E_1 = 0; E_2 = 1; E_3 = 0$

$$\begin{aligned} ATA &= b_0 + (b_1 * 0) + (b_2 * 1) + (b_3 * 0) \\ &= b_0 + b_2 \end{aligned}$$

For Catholic: $E_1 = 1; E_2 = 0; E_3 = 0$

$$\begin{aligned} ATA &= b_0 + (b_1 * 1) + (b_2 * 0) + (b_3 * 0) \\ &= b_0 + b_1 \end{aligned}$$

For Other: $E_1 = 0; E_2 = 0; E_3 = 1$

$$\begin{aligned} ATA &= b_0 + (b_1 * 0) + (b_2 * 0) + (b_3 * 1) \\ &= b_0 + b_3 \end{aligned}$$

- Interpreting the parameters:
  - b_0 = The average ATA (averaging the mean of the four groups)

    The test of b_0 tells us whether the average score of the outcome variable differs from zero.
  - b_1 = The difference in ATA between Catholics and the mean

    The test of b_1 tells us whether the mean score on the outcome variable for the group identified by E_1.differs from the grand mean.
  - b_2 = The difference in ATA between Protestants and the men

    The test of b_2 tells us whether the mean score on the outcome variable for the group identified by E_2.differs from the grand mean.
  - b_3 = The difference in ATA between Protestants and Others

    The test of b_3 tells us whether the mean score on the outcome variable for the group identified by E_3.differs from the grand mean.
  - If we wanted a test of whether the ATA of Protestants differed from the mean, we could re-run the analysis with a different group as the base group.
  - Again, be careful about interpreting "average" when the cell sizes are unequal; average refers to the average of the group means, not the average of the N observations.

- Interpreting correlation coefficients:
  - With more than two groups for an effects coded predictor variable, we should refrain from interpreting r_{E_1Y}, r_{E_2Y}, or r_{E_3Y}.
  - So long as all effects coded indicators are entered into the same regression equation, R^2 is still interpretable as the percentage of variance in Y (ATA) in the sample that is associated with religion. $R^2_{Adjusted}$ is the percentage of Y (ATA) variance accounted for by religion in the population.

- Running the analysis in SPSS

```
REGRESSION
  /STATISTICS COEFF OUTS R ANOVA ZPP
  /DEPENDENT ATA
  /METHOD=ENTER effect1 effect2 effect3.
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.596[a]	.355	.294	23.41817

a. Predictors: (Constant), effect3, effect1, effect2

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	81.900	4.055		20.198	.000			
	effect1	-21.233	6.849	-.598	-3.100	.004	-.444	-.481	-.440
	effect2	21.600	7.883	.550	2.740	.010	-.029	.436	.389
	effect3	-11.775	7.122	-.322	-1.653	.108	-.328	-.281	-.235

a. Dependent Variable: ata

$$ATA = 81.90 + (-21.233 * E_1) + (21.60 * E_2) + (-11.76 * E_3)$$

- The test of b_0 indicates that the mean ATA score, $\overline{Y}_{Group\ Means} = 81.90$ (calculated as the average of the four group means), is significantly above zero, $b = 81.90, t(32) = 20.20, p < .01$.

- The test of b_1 indicates that the mean ATA score for Catholics is significantly less than (because the sign is negative) the mean ATA score, $b = -21.23, t(32) = 3.10, p < .01$. The mean ATA score for Catholics is $\overline{Y}_{Catholics} = b_0 + b_1 = 81.900 - 21.233 = 60.67$.

- The test of b_2 indicates that the mean ATA score for Jews is significantly greater than (because the sign is positive) the mean ATA score, $b = 21.60, t(32) = 2.74, p = .01$. The mean ATA score for Jews is $\overline{Y}_{Jews} = b_0 + b_2 = 81.900 + 21.600 = 103.50$.

- The test of b_3 indicates that the mean ATA score for Others is not significantly different than the mean ATA score, $b = -11.78, t(32) = -1.65, p = .11$. The mean ATA score for Others is $\overline{Y}_{Others} = b_0 + b_3 = 81.900 - 11.775 = 70.13$.

- Overall 35.5% of the variability in ATA scores in the sample is associated with religion; 29.4% of the variability in ATA scores in the population is associated with religion.

- Unweighted vs. Weighted effects codes
  - We have considered unweighted effects coding. That is, each group mean is unweighted (or treated equally) regardless of the number of observations contributing to the group mean.
  - It is also possible to consider weighted effects coding in which each group mean is weighted by the number of observations contributing to the group mean.
    - The construction of the indictor variables takes into account the various group sizes.
    - Weighted effects codes correspond with Type I Sums of Squares in ANOVA
    - In general, you would only want to consider weighted effects codes if you have a representative sample.

4. Contrast coding
   - Contrast coding allows us to test specific, focused hypotheses regarding the levels of the (categorical) predictor variable and the outcome variable.
   - Contrast coding in regression is equivalent to conducting contrasts in an ANOVA framework.

   - Let's suppose a researcher wanted to compare the attitudes toward abortion in the following ways:
     - Judeo-Christian religions vs. others
     - Christian vs. Jewish
     - Catholic vs. Protestant

   - We need to convert each of these hypotheses to a set of contrast coefficients

Religion	C_1	C_2	C_3
Catholic	1	1	1
Protestant	1	1	-1
Jewish	1	-2	0
Other	-3	0	0

     - For each contrast, the sum of the contrast coefficients should equal zero
     - The contrasts should be orthogonal (assuming equal *n*)
       - If the contrast codes are not orthogonal, then you need to be very careful about interpreting the regression coefficients.

```
IF (religion = 1) cont1 = 1.
IF (religion = 2) cont1 = 1 .
IF (religion = 3) cont1 = 1 .
IF (religion = 4) cont1 = -3.

IF (religion = 1) cont2 = 1 .
IF (religion = 2) cont2 = 1.
IF (religion = 3) cont2 = -2 .
IF (religion = 4) cont2 = 0.

IF (religion = 1) cont3 = -1 .
IF (religion = 2) cont3 =  1 .
IF (religion = 3) cont3 =  0 .
IF (religion = 4) cont3 =  0.
```

- Now, we can enter all a-1 contrast codes into a regression equation

```
REGRESSION
  /STATISTICS COEFF OUTS R ANOVA ZPP
  /DEPENDENT ATA
  /METHOD=ENTER cont1 cont2 cont3.
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.596[a]	.355	.294	23.41817

a. Predictors: (Constant), cont3, cont1, cont2

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	81.900	4.055		20.198	.000			
	cont1	3.925	2.374	.237	1.653	.108	.225	.281	.235
	cont2	-8.838	3.608	-.352	-2.449	.020	-.277	-.397	-.348
	cont3	-16.321	5.077	-.459	-3.214	.003	-.444	-.494	-.456

a. Dependent Variable: ata

$$ATA = 81.90 + (3.925 * C_1) + (-8.838 * C_2) + (-16.321 * C_3)$$

For Catholic: $C_1 = 1; C_2 = 1; C_3 = 1$

$$\begin{aligned} ATA &= b_0 + (b_1 * 1) + (b_2 * 1) + (b_3 * 1) \\ &= b_0 + b_1 + b_2 + b_3 \\ &= 60.67 \end{aligned}$$

For Jewish: $C_1 = 1; C_2 = -2; C_3 = 0$

$$\begin{aligned} ATA &= b_0 + (b_1 * 1) + (b_2 * -2) + (b_3 * 0) \\ &= b_0 + b_1 - 2 * b_2 \\ &= 103.50 \end{aligned}$$

For Protestant: $C_1 = 1; C_2 = 1; C_3 = -1$

$$\begin{aligned} ATA &= b_0 + (b_1 * 1) + (b_2 * 1) + (b_3 * -1) \\ &= b_0 + b_1 + b_2 - b_3 \\ &= 93.31 \end{aligned}$$

For Other: $C_1 = -3; C_2 = 0; C_3 = 0$

$$\begin{aligned} ATA &= b_0 + (b_1 * -3) + (b_2 * 0) + (b_3 * 0) \\ &= b_0 - 3 * b_1 \\ &= 70.13 \end{aligned}$$

- The variance accounted for by religion is 35.5%, the same as we found in other parameterizations of the model.
- The test of b_0 indicates that the mean ATA score (calculated as the average of the four group means), $\overline{Y}_{Group\ Means} = 81.90$, is significantly above zero, $b = 81.90, t(32) = 20.20, p < .01$. The intercept may be interpreted as the mean because the set of contrast coding coefficients is orthogonal. The regression coefficients are not affected by the unequal *n* because we are taking an unweighted approach to unbalanced designs.

- In general, the other regression slope parameters are not directly interpretable, but the significance test associated with each parameter tells us about the contrast of interest.
  - Judeo-Christian religions and others do not differ in their attitudes toward abortion, $b = 3.93, t(32) = 1.65, p = .11$.
  - Individuals of a Christian faith have less favorable attitudes toward abortion than Jewish individuals, $b = -8.84, t(32) = -2.45, p = .02$
  - Catholic individuals less favorable attitudes toward abortion than Protestant individuals, $b = 16.32, t(32) = -3.31, p < .01$

- We would have obtained the same results had we conducted these contrasts in a oneway ANOVA framework.

```
ONEWAY ata BY religion
  /CONTRAST= 1 1 1 -3
  /CONTRAST= 1 1 -2 0
  /CONTRAST= -1 1 0 0
```

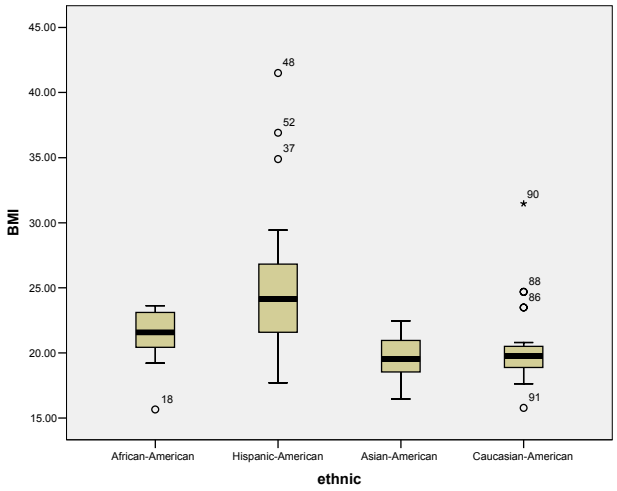
Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
ata	Assume equal variances	1	47.0994	28.48656	1.653	32	.108
		2	-53.0256	21.65011	-2.449	32	.020
		3	-32.6410	10.15480	-3.214	32	.003

- Note that the t-values and the p-values are identical to what we obtained from the regression analysis.

5. A Comparison between Regression and ANOVA
   - When the predictor variable is categorical and the outcome variable is continuous, we could run the analysis as a one-way AVOVA or as a regression. Let's compare these two approaches.
   - For this example, we'll examine ethnic differences in body mass index (BMI). First, we obtain a (stratified) random sample of Temple students, with equal numbers of participants in each of the four ethnic groups we are considering (n = 27). For each participant, we assess his or her BMI. Higher numbers indicate greater obesity.

African-American		Hispanic-American		Asian-American		Caucasian-American	
20.98	20.63	20.52	21.13	20.94	20.52	18.01	18.65
22.46	21.58	22.04	22.31	18.36	21.03	19.79	19.13
23.05	22.59	17.71	22.52	19.00	22.46	20.72	19.20
19.65	15.66	26.36	26.60	18.30	19.58	20.80	19.37
23.17	22.05	27.97	27.05	20.17	20.52	23.49	19.57
23.18	19.22	19.08	24.03	18.02	19.53	23.49	19.65
19.97	23.40	23.01	25.82	18.18	16.47	24.69	19.74
21.13	23.29	23.43	41.50	18.83	22.46	24.69	19.76
23.57	21.70	18.30	24.13	21.80	16.47	31.47	19.79
21.38	20.80	34.89	18.01	18.71	22.31	15.78	19.80
19.79	23.62	29.18	24.22	19.46	20.98	17.63	20.12
20.98	20.25	29.44	36.91	19.76	22.46	17.71	20.25
19.48	23.63	19.75	25.80	19.13	17.75	17.93	20.30
23.01		25.10		18.89		18.01	



   - We have some outliers and unequal variances, but let's ignore the assumptions for the moment and compare the ANOVA and regression outputs.

- In a regression framework, we can use effects coding to parameterize the model. I'll pick Caucasian-Americans to be the reference group.
- Effects code parameters are interpreted as deviations from the grand mean. Thus, the regression coefficients that come out of the model should match the $\hat{\alpha}_j$ terms we calculate in the ANOVA framework. Let's compute the model parameters in both models:

- ANOVA approach:

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

Descriptives

BMI

	N	Mean
African-American	27	21.4896
Hispanic-American	27	25.0670
Asian-American	27	19.7070
Caucasian-American	27	20.3533
Total	108	21.6543

$$\hat{\mu} = 21.65$$

$$\hat{\alpha}_j = \bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot}$$

$$\hat{\alpha}_1 = 21.4869 - 21.6543 = -0.165$$
$$\hat{\alpha}_2 = 25.0670 - 21.6543 = 3.413$$
$$\hat{\alpha}_3 = 19.7070 - 21.6543 = -1.947$$
$$\hat{\alpha}_4 = 20.3533 - 21.6543 = -1.301$$

- Regression Approach

Effects Coding

```
IF (ethnic = 1) effect1 = 1 .
IF (ethnic = 2) effect1 = 0 .
IF (ethnic = 3) effect1 = 0 .
IF (ethnic = 4) effect1 = -1 .

IF (ethnic = 1) effect2 = 0 .
IF (ethnic = 2) effect2 = 1 .
IF (ethnic = 3) effect2 = 0 .
IF (ethnic = 4) effect2 = -1 .

IF (ethnic = 1) effect3 = 0 .
IF (ethnic = 2) effect3 = 0 .
IF (ethnic = 3) effect3 = 1 .
IF (ethnic = 4) effect3 = -1 .
```

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error
1	(Constant)	21.654	.334
	effect1	-.165	.578
	effect2	3.413	.578
	effect3	-1.947	.578

a. Dependent Variable: BMI

  - As we expected, the coefficients match exactly!

$$\hat{\mu} = \hat{b}_0 \qquad \hat{\alpha}_2 = \hat{b}_2$$
$$\hat{\alpha}_1 = \hat{b}_1 \qquad \hat{\alpha}_3 = \hat{b}_3$$

  - These matching parameters indicate that is it possible for an ANOVA model and a regression model to be identically parameterized.

- The ANOVA tables outputted from ANOVA and regression also test equivalent hypotheses:
  - *ANOVA*: $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$ (There are no group effects)
  - *Regression*: $H_0 : b_1 = b_2 = b_3 = 0$ (The predictor variable accounts for no variability in the outcome variable).

  - Let's compare the ANOVA tables from the two analyses

```
ONEWAY BMI BY ethnic
  /STATISTICS DESCRIPTIVES.
```

ANOVA

BMI

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	463.272	3	154.424	12.828	.000
Within Groups	1251.986	104	12.038		
Total	1715.258	107			

```
REGRESSION
  /DEPENDENT BMI
  /METHOD=ENTER effect1 effect2 effect3.
```

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	463.272	3	154.424	12.828	.000[a]
	Residual	1251.986	104	12.038		
	Total	1715.258	107			

a. Predictors: (Constant), effect3, effect2, effect1

b. Dependent Variable: BMI

  - The results are identical, $F(3,104) = 12.83, p < .01$

  - If the results are identical, the effect size measures should also match, and they do!

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.520[a]	.270	.249	3.46963

a. Predictors: (Constant), effect3, effect2, effect1

Tests of Between-Subjects Effects

Dependent Variable: BMI

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	463.272[a]	3	154.424	12.828	.000	.270
Intercept	50641.950	1	50641.950	4206.727	.000	.976
ethnic	463.272	3	154.424	12.828	.000	.270
Error	1251.986	104	12.038			
Total	52357.208	108				
Corrected Total	1715.258	107				

a. R Squared = .270 (Adjusted R Squared = .249)

$$R^2 = \eta^2 = .27$$

- We have not yet examined the individual parameter estimates from the regression output. All the regression output we have examined thus far is independent of how the regression model is parameterized. Thus, any parameterization of the regression model should produce identical output to the ANOVA results.

- In fact, when you use the UNIANOVA command in SPSS, SPSS constructs dummy variables, runs a regression, and converts the output to an ANOVA format. We can see this by asking for *parameter estimates* in the output.

```
UNIANOVA BMI  BY ethnic
  /PRINT = PARAMETER.
```

Parameter Estimates

Dependent Variable: BMI

Parameter	B	Std. Error	t	Sig.
Intercept	20.353	.668	30.481	.000
[ethnic=1.00]	1.136	.944	1.203	.232
[ethnic=2.00]	4.714	.944	4.992	.000
[ethnic=3.00]	-.646	.944	-.684	.495
[ethnic=4.00]	0[a]	.	.	.

a. This parameter is set to zero because it is redundant.

- These are dummy variable indicators with group 4 as the reference group

```
IF (ethnic = 1) dummy1 = 1 .
IF (ethnic ne 1) dummy1 = 0 .
IF (ethnic = 2) dummy2 = 1 .
IF (ethnic ne 2) dummy2 = 0 .
IF (ethnic = 3) dummy3 = 1 .
IF (ethnic ne 3) dummy3 = 0 .

REGRESSION
  /STATISTICS COEFF OUTS R ANOVA
  /DEPENDENT BMI
  /METHOD=ENTER dummy1 dummy2 dummy3.
```

Coefficients[a]

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	(Constant)	20.353	.668	30.481	.000
	dummy1	1.136	.944	1.203	.232
	dummy2	4.714	.944	4.992	.000
	dummy3	-.646	.944	-.684	.495

a. Dependent Variable: BMI

- The tests of these the regression parameters will be equivalent to various contrasts in ANOVA.

ANOVA	Regression
Deviation contrasts	Effects coded parameters
Simple contrasts	Dummy coded parameters
Complex contrasts	Contrast coded parameters

- We have shown that ANOVA and regression are equivalent analyses. The common framework that unites the two is called the *general linear model*. Specifically, ANOVA is a special case of regression analysis.

- Some concepts and output are easier to understand or interpret from a regression framework.
  - Oftentimes, the regression approach is conceptually easier to understand than the ANOVA approach.
  - Unequal *n* designs are more easily understood from within a regression framework than an ANOVA framework.
  - In complicated designs (with many factors and covariates), it is easier to maintain control over the analysis in a regression framework.

- At this point, you might be asking yourself the converse question – why bother with ANOVA at all?
  - With simple designs, ANOVA is easier to understand and interpret.
  - Testing assumptions is a bit easier within an ANOVA framework than in a regression framework.
  - The procedures for controlling the Type I error (especially post-hoc tests) are easier to implement in an ANOVA framework.
  - Some tests that have been developed for assumption violations (Welch's *t*-test; Brown-Forsythe *F** test; some non-parametric tests) are easier to understand from an ANOVA approach.