**Regression with SAS  
Chapter 5: Additional coding systems for categorical variables in regression analysis**

**Chapter Outline**    5.1 Simple Coding  
    5.2 Forward Difference Coding  
    5.3 Backward Difference Coding  
    5.4 Helmert Coding  
    5.5 Reverse Helmert Coding  
    5.6 Deviation Coding  
    5.7 Orthogonal Polynomial Coding  
    5.8 User-Defined Coding  
    5.9 Summary

Categorical variables require special attention in regression analysis because, unlike dichotomous or continuous variables, they cannot by entered into the regression equation just as they are.  For example, if you have a variable called **race** that is coded 1 = Hispanic, 2 = Asian 3 = Black 4 = White, then entering **race** in your regression will look at the linear effect of race, which is probably not what you intended. Instead, categorical variables like this need to be recoded into a series of variables which can then be entered into the regression model.  There are a variety of coding systems that can be used when coding categorical variables.  Ideally, you would choose a coding system that reflects the comparisons that you want to make.  In [Chapter 3](http://www.ats.ucla.edu/stat/sas/webbooks/reg/chapter3/sasreg3.htm) of the [Regression with SAS Web Book](http://www.ats.ucla.edu/stat/sas/webbooks/reg/default.htm) we covered the use of categorical variables in regression analysis focusing on the use of dummy variables, but that is not the only coding scheme that you can use.  For example, you may want to compare each level to the next higher level, in which case you would want to use "forward difference" coding, or you might want to compare each level to the mean of the subsequent levels of the variable, in which case you would want to use "Helmert" coding.  By deliberately choosing a coding system, you can obtain comparisons that are most meaningful for testing your hypotheses.  Regardless of the coding system you choose, the test of the overall effect of the categorical variable (i.e., the overall effect of **race**) will remain the same.  Below is a table listing various types of contrasts and the comparison that they make.  

|  |  |
| --- | --- |
| **Name of contrast** | **Comparison made** |
| [Simple Coding](http://www.ats.ucla.edu/stat/sas/webbooks/reg/chapter5/sasreg5.htm#SIMPLE) | Compares each level of a variable to the reference level |
| [Forward Difference Coding](http://www.ats.ucla.edu/stat/sas/webbooks/reg/chapter5/sasreg5.htm#forward) | Adjacent levels of a variable (each level minus the next level) |
| [Backward Difference Coding](http://www.ats.ucla.edu/stat/sas/webbooks/reg/chapter5/sasreg5.htm#backward) | Adjacent levels of a variable (each level minus the prior level) |
| [Helmert Coding](http://www.ats.ucla.edu/stat/sas/webbooks/reg/chapter5/sasreg5.htm#HELMERT) | Compare levels of a variable with the mean of the subsequent levels of the variable |
| [Reverse Helmert Coding](http://www.ats.ucla.edu/stat/sas/webbooks/reg/chapter5/sasreg5.htm#reverse) | Compares levels of a variable with the mean of the previous levels of the variable |
| [Deviation Coding](http://www.ats.ucla.edu/stat/sas/webbooks/reg/chapter5/sasreg5.htm#DEVIATION) | Compares deviations from the grand mean |
| [Orthogonal Polynomial Coding](http://www.ats.ucla.edu/stat/sas/webbooks/reg/chapter5/sasreg5.htm#ORTHOGONAL) | Orthogonal polynomial contrasts |
| [User-Defined Coding](http://www.ats.ucla.edu/stat/sas/webbooks/reg/chapter5/sasreg5.htm#User) | User-defined contrast |

There are a couple of notes to be made about the coding systems listed above.  The first is that they represent planned comparisons and not post hoc comparisons.  In other words, they are comparisons that you plan to do before you begin analyzing your data, not comparisons that you think of once you have seen the results of preliminary analyses.  Also, some forms of coding make more sense with ordinal categorical variables than with nominal categorical variables. Below we will show examples using **race** as a categorical variable, which is a nominal variable.  Because simple effect coding compares the mean of the dependent variable for each level of the categorical variable to the mean of the dependent variable at for the reference level, it makes sense with a nominal variable.  However, it may not make as much sense to use a coding scheme that tests the linear effect of **race**.  As we describe each type of coding system, we note those coding systems with which it does not make as much sense to use a nominal variable.  Also, you may notice that we follow several rules when creating the contrast coding schemes.  For more information about these rules, please see the section on [User-Defined Coding](http://www.ats.ucla.edu/stat/sas/webbooks/reg/chapter5/sasreg5.htm#User).

This page will illustrate two ways that you can conduct analyses using these coding schemes: 1) using **proc glm** with **estimate** statements to define "contrast" coefficients that specify levels of the categorical variable that are to be compared**,** and 2) using **proc** **reg**. When using **proc reg** to do contrasts, you first need to create k-1 new variables (where k is the number of levels of the categorical variable) and use these new variables as predictors in your regression model.  Method 1 uses a type of coding we will call "contrast coding" while method 2 uses a type of coding we will call "regression coding".

**The Example Data File**

The examples in this page will use dataset called [hsb2.sas7bdat](http://www.ats.ucla.edu/stat/sas/webbooks/reg/hsb2.sas7bdat) and we will focus on the categorical variable **race**, which has four levels (1 = Hispanic, 2 = Asian, 3 = African American and 4 = white) and we will use **write** as our dependent variable.  Although our example uses a variable with four levels, these coding systems work with variables that have more or fewer categories. No matter which coding system you select, you will always have one fewer recoded variables than levels of the original variable.  In our example, our categorical variable has four levels so we will have three new variables (a variable corresponding to the final level of the categorical variables would be redundant and therefore unnecessary).

Before considering any analyses, let's look at the mean of the dependent variable, **write**, for each level of **race**.  This will help in interpreting the output from later analyses.

**proc means data = c:\sasreg\hsb2 mean n;**

**class race;**

**var write;**

**run;**

The MEANS Procedure

Analysis Variable : write writing score

N

race Obs Mean N

------------------------------------------

1 24 46.4583333 24

2 11 58.0000000 11

3 20 48.2000000 20

4 145 54.0551724 145

------------------------------------------

**5.1 Simple Coding**

The results of simple coding are very similar to dummy coding in that each level is compared to the reference level. In the example below, level 4 is the reference level and the first comparison compares level 1 to level 4, the second comparison compares level 2 to level 4, and the third comparison compares level 3 to level 4.

**Method 1: PROC GLM**

The table below shows the simple coding making the comparisons described above.  The first contrast compares level 1 to level 4, and level 1 is coded as 1 and level 4 is coded as -1.  Likewise, the second contrast compares level 2 to level 4 by coding level 2 as 1 and level 4 as -1.  As you can see with contrast coding, you can discern the meaning of the comparisons simply by inspecting the contrast coefficients.  For example, looking at the contrast coefficients for c3, you can see that it compares level 3 to level 4.

SIMPLE contrast coding

|  |  |  |  |
| --- | --- | --- | --- |
| Level of race | New variable 1 (c1) | New variable 2 (c2) | New variable 3 (c3) |
| 1 (Hispanic) | 1 | 0 | 0 |
| 2 (Asian) | 0 | 1 | 0 |
| 3 (African American) | 0 | 0 | 1 |
| 4 (white) | -1 | -1 | -1 |

Below we illustrate how to form these comparisons using **proc glm**.  As you see, a separate **estimate** statement is used for each contrast.

**proc glm data = c:\sasreg\hsb2;**

**class race;**

**model write = race;**

**estimate 'level 1 versus level 4' race 1 0 0 -1;**

**estimate 'level 2 versus level 4' race 0 1 0 -1;**

**estimate 'level 3 versus level 4' race 0 0 1 -1;**

**run;**

**quit;**

The contrast estimate for the first contrast compares the mean of the dependent variable, **write**, for levels 1 and 4 yielding -7.597 and is statistically significant (p<.000). The t-value associated with this test is -3.82.  The results of the second contrast, comparing the mean of **write** for levels 2 and 4 is not statistically significant (t = 1.40, p = .1638), while the third contrast is statistically significant.  Please note that while we have included the full SAS output for this example, we will only show the relevant output in later examples to conserve space.

The GLM Procedure

Dependent Variable: write writing score

Sum of

Source DF Squares Mean Square F Value Pr > F

Model 3 1914.15805 638.05268 7.83 <.0001

Error 196 15964.71695 81.45264

Corrected Total 199 17878.87500

R-Square Coeff Var Root MSE write Mean

0.107063 17.10111 9.025111 52.77500

Source DF Type I SS Mean Square F Value Pr > F

race 3 1914.158046 638.052682 7.83 <.0001

Source DF Type III SS Mean Square F Value Pr > F

race 3 1914.158046 638.052682 7.83 <.0001

Standard

Parameter Estimate Error t Value Pr > |t|

level 1 versus level 4 -7.59683908 1.98886958 -3.82 0.0002

level 2 versus level 4 3.94482759 2.82250377 1.40 0.1638

level 3 versus level 4 -5.85517241 2.15275967 -2.72 0.0071

**Method 2: Regression**

The regression coding is a bit more complex than contrast coding.  In our example below, level 4 is the reference level and **x1** compares level 1 to level 4, **x2** compares level 2 to level 4, and **x3** compares level 3 to level 4.  For **x1** the coding is 3/4 for level 1, and -1/4 for all other levels.  Likewise, for **x2** the coding is 3/4 for level 2, and -1/4 for all other levels, and for **x3** the coding is 3/4 for level 3, and -1/4 for all other levels.  It is not intuitive that this regression coding scheme yields these comparisons; however, if you desire simple comparisons, you can follow this general rule to obtain these comparisons.

SIMPLE regression coding

|  |  |  |  |
| --- | --- | --- | --- |
| Level of race | New variable 1 (x1) | New variable 2 (x2) | New variable 3 (x3) |
| 1 (Hispanic) | 3/4 | -1/4 | -1/4 |
| 2 (Asian) | -1/4 | 3/4 | -1/4 |
| 3 (African American) | -1/4 | -1/4 | 3/4 |
| 4 (white) | -1/4 | -1/4 | -1/4 |

Below we show the more general rule for creating this kind of coding scheme using regression coding, where k is the number of levels of the categorical variable (in this instance, k = 4).

SIMPLE regression coding

|  |  |  |  |
| --- | --- | --- | --- |
| Level of race | New variable 1 (x1) | New variable 2 (x2) | New variable 3 (x3) |
| 1 (Hispanic) | (k-1) / k | -1 / k | -1 / k |
| 2 (Asian) | -1 / k | (k-1) / k | -1 / k |
| 3 (African American) | -1 / k | -1 / k | (k-1) / k |
| 4 (white) | -1 / k | -1 / k | -1 / k |

Below we illustrate how to create **x1**, **x2** and **x3** and enter these new variables into the regression model using **proc** **reg**.

**data simple;**

**set c:\sasreg\hsb2;**

**if race = 1 then x1 = 3/4; else x1 = -1/4;**

**if race = 2 then x2 = 3/4; else x2 = -1/4;**

**if race = 3 then x3 = 3/4; else x3 = -1/4;**

**run;**

**proc reg data = simple;**

**model write = x1 x2 x3;**

**run;**

**quit;**

You will notice that the regression coefficients in the table below are the same as the contrast coefficients that we saw using **proc glm**.  Both the regression coefficient for **x1** and the contrast estimate for c1 are the mean of **write** for level 1 of **race** (Hispanic) minus the mean of **write** for level 4 (white). Likewise, the regression coefficient for **x2** and the contrast estimate for c2 are the mean of **write** for level 2 (Asian) minus the mean of **write** for level 4 (white). You also can see that the t values and significance levels are also the same as those from the **proc glm** output.  Please note that while we have included the full SAS output for this example, we will only show the relevant output in later examples to conserve space.

The REG Procedure

Model: MODEL1

Dependent Variable: write writing score

Analysis of Variance

Sum of Mean

Source DF Squares Square F Value Pr > F

Model 3 1914.15805 638.05268 7.83 <.0001

Error 196 15965 81.45264

Corrected Total 199 17879

Root MSE 9.02511 R-Square 0.1071

Dependent Mean 52.77500 Adj R-Sq 0.0934

Coeff Var 17.10111

Parameter Estimates

Parameter Standard

Variable Label DF Estimate Error t Value Pr > |t|

Intercept Intercept 1 51.67838 0.98212 52.62 <.0001

x1 1 -7.59684 1.98887 -3.82 0.0002

x2 1 3.94483 2.82250 1.40 0.1638

x3 1 -5.85517 2.15276 -2.72 0.0071

**5.2 Forward Difference Coding**

In this coding system, the mean of the dependent variable for one level of the categorical variable is compared to the mean of the dependent variable for the next (adjacent) level.  In our example below, the first comparison compares the mean of **write** for level 1 with the mean of **write** for level 2 of **race** (Hispanics minus Asians).  The second comparison compares the mean of **write** for level 2 minus level 3, and the third comparison compares the mean of **write** for level 3 minus level 4.  This type of coding may be useful with either a nominal or an ordinal variable.

**Method 1: PROC GLM**

FORWARD DIFFERENCE contrast coding

|  |  |  |  |
| --- | --- | --- | --- |
| Level of race | New variable 1 (c1) | New variable 2 (c2) | New variable 3 (c3) |
|  | Level 1 v. Level 2 | Level 2 v. Level 3 | Level 3 v. Level 4 |
| 1 (Hispanic) | 1 | 0 | 0 |
| 2 (Asian) | -1 | 1 | 0 |
| 3 (African American) | 0 | -1 | 1 |
| 4 (white) | 0 | 0 | -1 |

**proc glm data = c:\sasreg\hsb2;**

**class race;**

**model write = race;**

**estimate 'level 1 versus level 2' race 1 -1 0 0;**

**estimate 'level 2 versus level 3' race 0 1 -1 0;**

**estimate 'level 3 versus level 4' race 0 0 1 -1;**

**run;**

**quit;**

Standard

Parameter Estimate Error t Value Pr > |t|

level 1 versus level 2 -11.5416667 3.28612920 -3.51 0.0006

level 2 versus level 3 9.8000000 3.38783369 2.89 0.0043

level 3 versus level 4 -5.8551724 2.15275967 -2.72 0.0071

With this coding system, adjacent levels of the categorical variable are compared.  Hence, the mean of the dependent variable at level 1 is compared to the mean of the dependent variable at level 2:  46.4583 - 58 = -11.542, which is statistically significant.  For the comparison between levels 2 and 3, the calculation of the contrast coefficient would be 58 - 48.2 = 9.8, which is also statistically significant.  Finally, comparing levels 3 and 4, 48.2 - 54.0552 = -5.855, a statistically significant difference.  One would conclude from this that each adjacent level of **race** is statistically significantly different.

**Method 2: Regression**

For the first comparison, where the first and second levels are compared, **x1** is coded 3/4 for level 1 and the other levels are coded -1/4.  For the second comparison where level 2 is compared with level 3, **x2** is coded 1/2 1/2 -1/2 -1/2, and for the third comparison wherelevel 3 is compared with level 4, **x3** is coded 1/4 1/4 1/4 -3/4.

FORWARD DIFFERENCE regression coding

|  |  |  |  |
| --- | --- | --- | --- |
| Level of race | New variable 1 (x1) | New variable 2 (x2) | New variable 3 (x3) |
|  | Level 1 v. Level 2 | Level 2 v. Level 3 | Level 3 v. Level 4 |
| 1 (Hispanic) | 3/4 | 1/2 | 1/4 |
| 2 (Asian) | -1/4 | 1/2 | 1/4 |
| 3 (African American) | -1/4 | -1/2 | 1/4 |
| 4 (white) | -1/4 | -1/2 | -3/4 |

The general rule for this regression coding scheme is shown below, where k is the number of levels of the categorical variable (in this case k = 4).

FORWARD DIFFERENCE regression coding

|  |  |  |  |
| --- | --- | --- | --- |
| Level of race | New variable 1 (x1) | New variable 2 (x2) | New variable 3 (x3) |
|  | Level 1 v. Level 2 | Level 2 v. Level 3 | Level 3 v. Level 4 |
| 1 (Hispanic) | (k-1)/k | (k-2)/k | (k-3)/k |
| 2 (Asian) | -1/k | (k-2)/k | (k-3)/k |
| 3 (African American) | -1/k | -2/k | (k-3)/k |
| 4 (white) | -1/k | -2/k | -3/k |

**data forward;**

**set c:\sasreg\hsb2;**

**if race = 1 then x1 = 3/4; else x1 = -1/4;**

**if race = 1 or race = 2 then x2 = 1/2;**

**if race = 3 or race = 4 then x2 = -1/2;**

**if race = 4 then x3 = -3/4; else x3 = 1/4;**

**run;**

**proc reg data = forward;**

**model write = x1 x2 x3;**

**run;**

**quit;**

Parameter Estimates

Parameter Standard

Variable Label DF Estimate Error t Value Pr > |t|

Intercept Intercept 1 51.67838 0.98212 52.62 <.0001

x1 1 -11.54167 3.28613 -3.51 0.0006

x2 1 9.80000 3.38783 2.89 0.0043

x3 1 -5.85517 2.15276 -2.72 0.0071

You can see the regression coefficient for **x1** is the mean of **write** for level 1 (Hispanic) minus the mean of **write** for level 2 (Asian).  Likewise, the regression coefficient for **x2** is the mean of **write** for level 2 (Asian) minus the mean of **write** for level 3 (African American), and the regression coefficient for **x3** is the mean of **write** for level 3 (African American) minus the mean of **write** for level 4 (white).

**5.3 Backward Difference Coding**

In this coding system, the mean of the dependent variable for one level of the categorical variable is compared to the mean of the dependent variable for the prior adjacent level.  In our example below, the first comparison compares the mean of **write** for level 2 with the mean of **write** for level 1 of **race** (Hispanics minus Asians).  The second comparison compares the mean of **write** for level 3 minus level 2, and the third comparison compares the mean of **write** for level 4 minus level 3.  This type of coding may be useful with either a nominal or an ordinal variable.

**Method 1: PROC GLM**

BACKWARD DIFFERENCE contrast coding

|  |  |  |  |
| --- | --- | --- | --- |
| Level of race | New variable 1 (c1) | New variable 2 (c2) | New variable 3 (c3) |
|  | Level 1 v. Level 2 | Level 2 v. Level 3 | Level 3 v. Level 4 |
| 1 (Hispanic) | -1 | 0 | 0 |
| 2 (Asian) | 1 | -1 | 0 |
| 3 (African American) | 0 | 1 | -1 |
| 4 (white) | 0 | 0 | 1 |

**proc glm data = c:\sasreg\hsb2;**

**class race;**

**model write = race;**

**estimate 'level 1 versus level 2' race -1 1 0 0;**

**estimate 'level 2 versus level 3' race 0 -1 1 0;**

**estimate 'level 3 versus level 4' race 0 0 -1 1;**

**run;**

**quit;**

Standard

Parameter Estimate Error t Value Pr > |t|

level 1 versus level 2 11.5416667 3.28612920 3.51 0.0006

level 2 versus level 3 -9.8000000 3.38783369 -2.89 0.0043

level 3 versus level 4 5.8551724 2.15275967 2.72 0.0071

With this coding system, adjacent levels of the categorical variable are compared, with each level compared to the prior level.  Hence, the mean of the dependent variable at level 2 is compared to the mean of the dependent variable at level 1:  58 - 46.4583 = 11.542, which is statistically significant.  For the comparison between levels 3 and 2, the calculation of the contrast coefficient is 48.2 - 58 = -9.8, which is also statistically significant.  Finally, comparing levels 4 and 3, 54.0552 - 48.2 = 5.855, a statistically significant difference.  One would conclude from this that each adjacent level of **race** is statistically significantly different.

**Method 2: Regression**

For the first comparison, where the first and second levels are compared, **x1** is coded 3/4 for level 1 while the other levels are coded -1/4.  For the second comparison where level 2 is compared with level 3, **x2** is coded 1/2 1/2 -1/2 -1/2, and for the third comparison wherelevel 3 is compared with level 4, **x3** is coded 1/4 1/4 1/4 -3/4.

BACKWARD DIFFERENCE regression coding

|  |  |  |  |
| --- | --- | --- | --- |
| Level of race | New variable 1 (x1) | New variable 2 (x2) | New variable 3 (x3) |
|  | Level 2 v. Level 1 | Level 3 v. Level 2 | Level 4 v. Level 3 |
| 1 (Hispanic) | - 3/4 | -1/2 | -1/4 |
| 2 (Asian) | 1/4 | -1/2 | -1/4 |
| 3 (African American) | 1/4 | 1/2 | -1/4 |
| 4 (white) | 1/4 | 1/2 | 3/4 |

The general rule for this regression coding scheme is shown below, where k is the number of levels of the categorical variable (in this case, k = 4).

BACKWARD DIFFERENCE regression coding

|  |  |  |  |
| --- | --- | --- | --- |
| Level of race | New variable 1 (x1) | New variable 2 (x2) | New variable 3 (x3) |
|  | Level 1 v. Level 2 | Level 2 v. Level 3 | Level 3 v. Level 4 |
| 1 (Hispanic) | -(k-1)/k | -(k-2)/k | -(k-3)/k |
| 2 (Asian) | 1/k | -(k-2)/k | -(k-3)/k |
| 3 (African American) | 1/k | 2/k | -(k-3)/k |
| 4 (white) | 1/k | 2/k | 3/k |

**data backward;**

**set c:\sasreg\hsb2;**

**if race = 1 then x1 = -3/4; else x1 = 1/4;**

**if race = 1 or race = 2 then x2 = -1/2;**

**if race = 3 or race = 4 then x2 = 1/2;**

**if race = 4 then x3 = 3/4; else x3 = -1/4;**

**run;**

**proc reg data = backward;**

**model write = x1 x2 x3;**

**run;**

**quit;**

Parameter Estimates

Parameter Standard

Variable Label DF Estimate Error t Value Pr > |t|

Intercept Intercept 1 51.67838 0.98212 52.62 <.0001

x1 1 11.54167 3.28613 3.51 0.0006

x2 1 -9.80000 3.38783 -2.89 0.0043

x3 1 5.85517 2.15276 2.72 0.0071

In the above example, the regression coefficient for **x1** is the mean of **write** for level 2 minus the mean of **write** for level 1 (58- 46.4583 = 11.542).  Likewise, the regression coefficient for **x2** is the mean of **write** for level 3 minus the mean of **write** for level 2, and the regression coefficient for **x3** is the mean of **write** for level 4 minus the mean of **write** for level 3.

**5.4 Helmert Coding**

Helmert coding compares each level of a categorical variable to the mean of the subsequent levels.  Hence, the first contrast compares the mean of the dependent variable for level 1 of **race** with the mean of all of the subsequent levels of **race** (levels 2, 3, and 4), the second contrast compares the mean of the dependent variable for level 2 of **race** with the mean of all of the subsequent levels of **race** (levels 3 and 4), and the third contrast compares the mean of the dependent variable for level 3 of **race** with the mean of all of the subsequent levels of **race** (level 4). While this type of coding system does not make much sense with a nominal variable like **race**, it is useful in situations where the levels of the categorical variable are ordered say, from lowest to highest, or smallest to largest, etc.

For Helmert coding, we see that the first comparison comparing level 1 with levels 2, 3 and 4 is coded 1, -1/3, -1/3 and -1/3, reflecting the comparison of level 1 with all other levels.  The second comparison is coded 0, 1, -1/2 and -1/2, reflecting that it compares level 2 with levels 3 and 4.  The third comparison is coded 0, 0, 1 and -1, reflecting that level 3 is compared to level 4.

**Method 1: PROC GLM**

HELMERT contrast coding

|  |  |  |  |
| --- | --- | --- | --- |
| Level of race | New variable 1 (c1) | New variable 2 (c2) | New variable 3 (c3) |
|  | Level 1 v. Later | Level 2 v. Later | Level 3 v. Later |
| 1 (Hispanic) | 1 | 0 | 0 |
| 2 (Asian) | -1/3 | 1 | 0 |
| 3 (African American) | -1/3 | -1/2 | 1 |
| 4 (white) | -1/3 | -1/2 | -1 |

Below we illustrate how to form these comparisons using **proc glm** with **estimate** statements.  Note that on the first estimate statement we indicate -.33333 and not just -.33.  We need to use this many decimals so the sum of all of the contrast coefficients (i.e., 1 + -.333333 + -.333333 + -.333333) is sufficiently close to zero, otherwise SAS will say that the term cannot be estimated.

**proc glm data = c:\sasreg\hsb2;**

**class race;**

**model write = race;**

**estimate 'level 1 versus levels 2, 3 & 4' race 1 -.33333 -.33333 -.33333;**

**estimate 'level 2 versus levels 3 & 4' race 0 1 -.5 -.5;**

**estimate 'level 3 versus level 4' race 0 0 1 -1;**

**run;**

**quit;**

Standard

Parameter Estimate Error t Value Pr > |t|

level 1 versus levels 2, 3 & 4 -6.96006384 2.17520603 -3.20 0.0016

level 2 versus levels 3 & 4 6.87241379 2.92632513 2.35 0.0198

level 3 versus level 4 -5.85517241 2.15275967 -2.72 0.0071

The contrast estimate for the comparison between level 1 and the remaining levels is calculated by taking the mean of the dependent variable for level 1 and subtracting the mean of the dependent variable for levels 2, 3 and 4: 46.4583 - [(58 + 48.2 + 54.0552) / 3] = -6.960, which is statistically significant.  This means that the mean of **write** for level 1 of **race** is statistically significantly different from the mean of **write** for levels 2 through 4.  As noted above, this comparison probably is not meaningful because the variable **race** is nominal.  This type of comparison would be more meaningful if the categorical variable was ordinal.

To calculate the contrast coefficient for the comparison between level 2 and the later levels, you subtract the mean of the dependent variable for levels 3 and 4 from the mean of the dependent variable for level 2:  58 - [(48.2 + 54.0552) / 2] = 6.872, which is statistically significant.  The contrast estimate for the comparison between level 3 and level 4 is the difference between the mean of the dependent variable for the two levels:  48.2 - 54.0552 = -5.855, which is also statistically significant.

**Method 2: Regression**

Below we see an example of Helmert regression coding.  For the first comparison (comparing level 1 with levels 2, 3 and 4) the codes are 3/4 and -1/4 -1/4 -1/4.  The second comparison compares level 2 with levels 3 and 4 and is coded 0 2/3 -1/3 -1/3.  The third comparison compares level 3 to level 4 and is coded 0 0 1/2 -1/2.

HELMERT regression coding

|  |  |  |  |
| --- | --- | --- | --- |
| Level of race | New variable 1 (x1) | New variable 2 (x2) | New variable 3 (x3) |
|  | Level 1 v. Later | Level 2 v. Later | Level 3 v. Later |
| 1 (Hispanic) | 3/4 | 0 | 0 |
| 2 (Asian) | -1/4 | 2/3 | 0 |
| 3 (African American) | -1/4 | -1/3 | 1/2 |
| 4 (white) | -1/4 | -1/3 | -1/2 |

Below we illustrate how to create **x1**, **x2** and **x3** and enter these new variables into the regression model using **porc reg**.

**data helmert;**

**set c:\sasreg\hsb2;**

**if race = 1 then x1 = .75; else x1 = -.25;**

**if race = 1 then x2 = 0;**

**if race = 2 then x2 = 2/3;**

**if race = 3 or race = 4 then x2 = -1/3;**

**if race = 1 or race = 2 then x3 = 0;**

**if race = 3 then x3 = 1/2;**

**if race = 4 then x3 = -1/2;**

**run;**

**proc reg data = helmert;**

**model write = x1 x2 x3;**

**run;**

**quit;**

As you see below, the regression coefficient for **x1** is the mean of **write** for level 1 (Hispanic) versus all subsequent levels (levels 2, 3 and 4).  Likewise, the regression coefficient for **x2** is the mean of **write** for level 2 minus the mean of **write** for levels 3 and 4.  Finally, the regression coefficient for **x3** is the mean of **write** for level 3 minus the mean of **write** for level 4.

Parameter Estimates

Parameter Standard

Variable Label DF Estimate Error t Value Pr > |t|

Intercept Intercept 1 51.67836 0.98212 52.62 <.0001

x1 1 -6.96003 2.17521 -3.20 0.0016

x2 1 6.87241 2.92633 2.35 0.0198

x3 1 -5.85517 2.15276 -2.72 0.0071

**5.5 Reverse Helmert Coding**

Reverse Helmert coding (also know as difference coding) is just the opposite of Helmert coding: instead of comparing each level of categorical variable to the mean of the subsequent level(s), each is compared to the mean of the previous level(s).  In our example, the first contrast codes the comparison of the mean of the dependent variable for level 2 of **race** to the mean of the dependent variable for level 1 of **race**.  The second comparison compares the mean of the dependent variable level 3 of **race** with both levels 1 and  2 of **race**, and the third comparison compares the mean of the dependent variable for level 4 of **race** with levels 1, 2 and 3. Clearly, this coding system does not make much sense with our example of **race** because it is a nominal variable.  However, this system is useful when the levels of the categorical variable are ordered in a meaningful way.  For example, if we had a categorical variable in which work-related stress was coded as low, medium or high, then comparing the means of the previous levels of the variable would make more sense.

For reverse Helmert coding, we see that the first comparison comparing levels 1 and 2 are coded -1 and 1 to compare these levels, and 0 otherwise.  The second comparison comparing levels 1, 2 with level 3 are coded -1/2, -1/2,  1 and 0, and the last comparison comparing levels 1, 2 and 3 with level 4 are coded -1/3, -1/3, -1/3 and 1.

**Method 1: PROC GLM**

REVERSE HELMERT contrast coding

|  |  |  |  |
| --- | --- | --- | --- |
|  | New variable 1 (c1) | New variable 2 (c2) | New variable 3 (c3) |
|  | Level 2 v. Level 1 | Level 3 v. Previous | Level 4 v. Previous |
| 1 (Hispanic) | -1 | -1/2 | -1/3 |
| 2 (Asian) | 1 | -1/2 | -1/3 |
| 3 (African American) | 0 | 1 | -1/3 |
| 4 (white) | 0 | 0 | 1 |

Below we illustrate how to form these comparisons using **proc** **glm** with **estimate** statements.  Note that on the third estimate statement we indicate -.33333 and not just -.33.  We need to use this many decimals so the sum of all of the contrast coefficients (i.e., -.333333 + - .333333 + - .333333 + 1) is sufficiently close to zero, otherwise SAS will say that the term cannot be estimated.

**proc glm data = c:\sasreg\hsb2;**

**class race;**

**model write = race;**

**estimate 'level 2 versus level1' race -1 1 0 0;**

**estimate 'level 3 versus levels 1 & 2' race -.5 -.5 1 0;**

**estimate 'level 4 versus levels 1, 2 & 4' race -.33333 -.33333 -.33333 1;**

**run;**

**quit;**

An alternate way, which solves the problem of the repeating decimals, is shown below.  Only one output is shown because the two outputs are identical.

**proc glm data = c:\sasreg\hsb2;**

**class race;**

**model write = race;**

**estimate 'level 2 versus level 1' race -1 1 0 0;**

**estimate 'level 3 versus levels 1 & 2' race -.5 -.5 1 0;**

**estimate 'level 4 versus levels 1, 2 & 4' race -1 -1 -1 3 / divisor=3;**

**run;**

**quit;**

Standard

Parameter Estimate Error t Value Pr > |t|

level 2 versus level1 11.5416667 3.28612920 3.51 0.0006

level 3 versus levels 1 & 2 -4.0291667 2.60236299 -1.55 0.1232

level 4 versus levels 1, 2 & 4 3.1690296 1.48797250 2.13 0.0344

The contrast estimate for the first comparison shown in this output was calculated by subtracting the mean of the dependent variable for level 2 of the categorical variable from the mean of the dependent variable for level 1:  58 - 46.4583 = 11.542.  This result is statistically significant.  The contrast estimate for the second comparison (between level 3 and the previous levels) was calculated by subtracting the mean of the dependent variable for levels 1 and 2 from that of level 3:  48.2 - [(46.4583 + 58) / 2] = -4.029.  This result is not statistically significant, meaning that there is not a reliable difference between the mean of **write** for level 3 of **race** compared to the mean of **write** for levels 1 and 2 (Hispanics and Asians).  As noted above, this type of coding system does not make much sense for a nominal variable such as **race**.  For the comparison of level 4 and the previous levels, you take the mean of the dependent variable for the those levels and subtract it from the mean of the dependent variable for level 4:  54.0552 - [(46.4583 + 58 + 48.2) / 3] = 3.169.  This result is statistically significant.

**Method 2: Regression**

The regression coding for reverse Helmert coding is shown below.  For the first comparison, where the first and second level are compared, **x1** is coded -1/2 and 1/2 and 0 otherwise.  For the second comparison, the values of **x2** are coded -1/3 -1/3  2/3 and 0.  Finally, for the third comparison, the values of **x3** are coded -1/4 -1/4 -/14 and 3/4.

REVERSE HELMERT regression coding

|  |  |  |  |
| --- | --- | --- | --- |
| Level of race | New variable 1 (x1) | New variable 2 (x2) | New variable 3 (x3) |
| 1 (Hispanic) | -1/2 | -1/3 | -1/4 |
| 2 (Asian) | 1/2 | -1/3 | -1/4 |
| 3 (African American) | 0 | 2/3 | -1/4 |
| 4 (white) | 0 | 0 | 3/4 |

Below we illustrate how to create **x1**, **x2** and **x3** and enter these new variables into the regression model using **proc** **reg**.

**data diff;**

**set c:\sasreg\hsb2;**

**if race = 1 then x1 = -1/2;**

**if race = 2 then x1 = 1/2;**

**if race = 3 or race = 4 then x1 = 0;**

**if race = 1 or race = 2 then x2 = -1/3;**

**if race = 3 then x2 = 2/3;**

**if race = 4 then x2 = 0;**

**if race = 4 then x3 = 3/4; else x3 = -1/4;**

**run;**

**proc reg data = diff;**

**model write = x1 x2 x3;**

**run;**

**quit;**

Parameter Estimates

Parameter Standard

Variable Label DF Estimate Error t Value Pr > |t|

Intercept Intercept 1 51.67839 0.98212 52.62 <.0001

x1 1 11.54167 3.28613 3.51 0.0006

x2 1 -4.02917 2.60236 -1.55 0.1232

x3 1 3.16905 1.48799 2.13 0.0344

In the above examples, both the regression coefficient for **x1** and the contrast estimate for c1 would be the mean of **write** for level 1 (Hispanic) minus the mean of **write** for level 2 (Asian).  Likewise, the regression coefficient for **x2** and the contrast estimate for c2 would be the mean of **write** for levels 1 and 2 combined minus the mean of **write** for level 3.  Finally, the regression coefficient for **x3** and the contrast estimate for c3 would be the mean of **write** for levels 1, 2 and 3 combined minus the mean of **write** for level 4.

**5.6 Deviation Coding**

This coding system compares the mean of the dependent variable for a given level to the overall mean of the dependent variable.  In our example below, the first comparison compares level 1 (Hispanics) to all levels of **race**, the second comparison compares level 2 (Asians) to all levels of **race**, and the third comparison compares level 3 (African Americans) to all levels of **race**.

As you can see, the logic of the contrast coding is fairly straightforward.  The first comparison compares level 1 to levels 2, 3 and 4.  A value of 3/4 is assigned to level 1 and a value of -1/4 is assigned to levels 2, 3 and 4.  Likewise, the second comparison compares level 2 to levels 1, 3 and 4. A value of 3/4 is assigned to level 2 and a value of -1/4 is assigned to levels 1, 3 and 4. A similar pattern is followed for assigning values for the third comparison.  Note that you could substitute 3 for 3/4 and 1 for 1/4 and you would get the same test of significance, but the contrast coefficient would be different.

**Method 1: PROC GLM**

DEVIATION contrast coding

|  |  |  |  |
| --- | --- | --- | --- |
| Level of race | New variable 1 (c1) | New variable 2 (c2) | New variable 3 (c3) |
|  | Level 1 v. Mean | Level 2 v. Mean | Level 3 v. Mean |
| 1 (Hispanic) | 3/4 | -1/4 | -1/4 |
| 2 (Asian) | -1/4 | 3/4 | -1/4 |
| 3 (African American) | -1/4 | -1/4 | 3/4 |
| 4 (white) | -1/4 | -1/4 | -1/4 |

Below we illustrate how to form these comparisons using **proc glm**.

**proc glm data = c:\sasreg\hsb2;**

**class race;**

**model write = race;**

**estimate 'level 1 versus levels 2, 3 & 4' race .75 -.25 -.25 -.25;**

**estimate 'level 2 versus levels 1, 3 & 4' race -.25 .75 -.25 -.25;**

**estimate 'level 3 versus levels 1, 2 & 4' race -.25 -.25 .75 -.25;**

**run;**

**quit;**

Standard

Parameter Estimate Error t Value Pr > |t|

level 1 versus levels 2, 3 & 4 -5.22004310 1.63140849 -3.20 0.0016

level 2 versus levels 1, 3 & 4 6.32162356 2.16031394 2.93 0.0038

level 3 versus levels 1, 2 & 4 -3.47837644 1.73230472 -2.01 0.0460

The contrast estimate is the mean for level 1 minus the grand mean.  However, this grand mean is not the mean of the dependent variable that is listed in the output of the **means** command above.  Rather it is the mean of means of the dependent variable at each level of the categorical variable:  (46.4583 + 58 + 48.2 + 54.0552) / 4 = 51.678375.  This contrast estimate is then 46.4583 - 51.678375 = -5.220.  The difference between this value and zero (the null hypothesis that the contrast coefficient is zero) is statistically significant (p = .0016), and the t-value for this test of -3.20.  The results for the next two contrasts were computed in a similar manner.

**Method 2: Regression**

As you see in the example below, the regression coding is accomplished by assigning 1 to level 1 for the first comparison (because level 1 is the level to be compared to all others), a 1 to level 2 for the second comparison (because level 2 is to be compared to all others), and 1 to level 3 for the third comparison (because level 3 is to be compared to all others).  Note that a  -1 is assigned to level 4 for all three comparisons (because it is the level that is never compared to the other levels) and all other values are assigned a 0.  This regression coding scheme yields the comparisons described above.

DEVIATION regression coding

|  |  |  |  |
| --- | --- | --- | --- |
| Level of race | New variable 1 (x1) | New variable 2 (x2) | New variable 3 (x3) |
|  | Level 1 v. Mean | Level 2 v. Mean | Level 3 v. Mean |
| 1 (Hispanic) | 1 | 0 | 0 |
| 2 (Asian) | 0 | 1 | 0 |
| 3 (African American) | 0 | 0 | 1 |
| 4 (white) | -1 | -1 | -1 |

Below we illustrate how to create **x1**, **x2** and **x3** and enter these new variables into the regression model using **proc** **reg**.

**data deviation;**

**set c:\sasreg\hsb2;**

**if race = 1 then x1 = 1;**

**if race = 2 or race = 3 then x1 = 0;**

**if race = 4 then x1 = -1;**

**if race = 2 then x2 = 1;**

**if race = 1 or race = 3 then x2 = 0;**

**if race = 4 then x2 = -1;**

**if race = 3 then x3 = 1;**

**if race = 1 or race = 2 then x3 = 0;**

**if race = 4 then x3 = -1;**

**run;**

**proc reg data = deviation;**

**model write = x1 x2 x3;**

**run;**

**quit;**

In this example, both the regression coefficient for **x1** is the mean of **write** for level 1 (Hispanic) minus the grand mean of **write.** Likewise, the regression coefficient for **x2** is the mean **write** for level 2 (Asian) minus the grand mean of **write**, and so on. As we saw in the previous analyses, all three contrasts are statistically significant.

Parameter Estimates

Parameter Standard

Variable Label DF Estimate Error t Value Pr > |t|

Intercept Intercept 1 51.67838 0.98212 52.62 <.0001

x1 1 -5.22004 1.63141 -3.20 0.0016

x2 1 6.32162 2.16031 2.93 0.0038

x3 1 -3.47838 1.73230 -2.01 0.0460

**5.7 Orthogonal Polynomial Coding**

Orthogonal polynomial coding is a form of trend analysis in that it is looking for the linear, quadratic and cubic trends in the categorical variable.  This type of coding system should be used only with an ordinal variable in which the levels are equally spaced.  Examples of such a variable might be income or education.  The table below shows the contrast coefficients for the linear, quadratic and cubic trends for the four levels.  These could be obtained from most statistics books on linear models.

POLYNOMIAL

|  |  |  |  |
| --- | --- | --- | --- |
| Level of race | Linear (x1) | Quadratic (x2) | Cubic (x3) |
| 1 (Hispanic) | -.671 | .5 | -.224 |
| 2 (Asian) | -.224 | -.5 | .671 |
| 3 (African American) | .224 | -.5 | -.671 |
| 4 (white) | .671 | .5 | .224 |

**Method 1: PROC GLM**

**proc glm data = c:\sasreg\hsb2;**

**class race;**

**model write = race;**

**estimate 'linear' race -.671 -.224 .224 .671;**

**estimate 'quadratic' race .5 -.5 -.5 .5;**

**estimate 'cubic' race -.224 .671 -.671 .224;**

**run;**

**quit;**

Standard

Parameter Estimate Error t Value Pr > |t|

linear 2.90227902 1.53520851 1.89 0.0602

quadratic -2.84324713 1.96424409 -1.45 0.1494

cubic 8.27749195 2.31648010 3.57 0.0004

To calculate the contrast estimates for these comparisons, you need to multiply the code used in the new variable by the mean for the dependent variable for each level of the categorical variable, and then sum the values.  For example, the code used in **x1** for level 1 of **race** is -.671 and the mean of **write** for level 1 is 46.4583.  Hence, you would multiply -.671 and 46.4583 and add that to the product of the code for level 2 of **x1** and its mean, and so on.  To obtain the contrast estimate for the linear contrast, you would do the following:  -.671\*46.4583 + -.224\*58 + .224\*48.2 + .671\*54.0552 = 2.905 (with rounding error).  This result is not statistically significant at the .05 alpha level, but it is close.  The quadratic component is also not statistically significant, but the cubic one is.  This suggests that, if the mean of the dependent variable was plotted against **race**, the line would tend to have two bends.  As noted earlier, this type of coding system does not make much sense with a nominal variable such as **race**.

**Method 2: Regression**

The regression coding for orthogonal polynomial coding is the same as the contrast coding.  Below you can see the SAS code for creating **x1**, **x2** and **x3** that correspond to the linear, quadratic and cubic trends for **race**.

**data poly;**

**set c:\sasreg\hsb2;**

**if race = 1 then x1 = -.671;**

**if race = 2 then x1 = -.224;**

**if race = 3 then x1 = .224;**

**if race = 4 then x1 = .671;**

**if race = 1 then x2 = .5;**

**if race = 2 then x2 = -.5;**

**if race = 3 then x2 = -.5;**

**if race = 4 then x2 = .5;**

**if race = 1 then x3 = -.224;**

**if race = 2 then x3 = .671;**

**if race = 3 then x3 = -.671;**

**if race = 4 then x3 = .224;**

**run;**

**proc reg data = poly;**

**model write = x1 x2 x3;**

**run;**

**quit;**

Parameter Estimates

Parameter Standard

Variable Label DF Estimate Error t Value Pr > |t|

Intercept Intercept 1 51.67838 0.98212 52.62 <.0001

x1 1 2.89986 1.53393 1.89 0.0602

x2 1 -2.84325 1.96424 -1.45 0.1494

x3 1 8.27059 2.31455 3.57 0.0004

The regression coefficients obtained from this analysis are the same as the contrast coefficients obtained using **proc glm**.

**5.8 User Defined Coding**

You can use SAS for any general kind of coding scheme.  For our example, we would like to make the following three comparisons:

1) level 1 to level 3    
2) level 2 to levels 1 and 4   
3) levels 1 and 2 to levels 3 and 4.

In order to compare level 1 to level 3, we use the contrast coefficients 1 0 -1 0. To compare level 2 to levels 1 and 4 we use the contrast coefficients -1/2 1 0 -1/2 .  Finally, to compare levels 1 and 2 with levels 3 and 4 we use the coefficients 1/2 1/2 -1/2 -1/2.  Before proceeding to the SAS code necessary to conduct these analyses, let's take a moment to more fully explain the logic behind the selection of these contrast coefficients.

For the first contrast, we are comparing level 1 to level 3, and the contrast coefficients are 1 0 -1 0.  This means that the levels associated with the contrast coefficients with opposite signs are being compared.  In fact, the mean of the dependent variable is multiplied by the contrast coefficient.  Hence, levels 2 and 4 are not involved in the comparison:  they are multiplied by zero and "dropped out."  You will also notice that the contrast coefficients sum to zero.  This is necessary.  If the contrast coefficients do not sum to zero, the contrast is not estimable and SAS will issue an error message. Which level of the categorical variable is assigned a positive or negative value is not terribly important:  1 0 -1 0 is the same as -1 0 1 0 in that both of these codings compare the first and the third levels of the variable.  However, the sign of the regression coefficient would change.

Now let's look at the contrast coefficients for the second and third comparisons.  You will notice that in both cases we use fractions that sum to one (or minus one).  They do not have to sum to one (or minus one).  You may wonder why we would use fractions like -1/2 1 0 -1/2 instead of whole numbers such as -1 2 0 -1.  While -1/2 1 0 -1/2 and -1 2 0 -1 both compare level 2 with levels 1 and 4 and both will give you the same t-value and p-value for the regression coefficient, the contrast estimates/regression coefficients themselves would be different, as would their interpretation.  The coefficient for the -1/2 1 0 -1/2 contrast is the mean of level 2 minus the mean of the means for levels 1 and 4:  58 - (46.4583 + 54.0552)/2 = 7.74325.  (Alternatively, you can multiply the contrasts by the mean of the dependent variable for each level of the categorical variable: -1/2\*46.4583 + 1\*58.00 + 0\*48.20 + -1/2\*54.0552 = 7.74325.  Clearly these are equivalent ways of thinking about how the contrast coefficient is calculated.)  By comparison, the coefficient for the -1 2 0 -1 contrast is two times the mean for level 2 minus the means of the dependent variable for levels 1 and 4:  2\*58 - (46.4583 + 54.0552) = 15.4865, which is the same as -1\*46.4583 + 2\*58 + 0\*48.20 - 1\*54.0552 = 15.4865. Note that the regression coefficient using the contrast coefficients -1 2 0 -1 is twice the regression coefficient obtained when -1/2 1 0 -1/2 is used.

**Method 1: PROC GLM**

In order to compare level 1 to level 3, we use the contrast coefficients 1 0 -1 0. To compare level 2 to levels 1 and 4 we use the contrast coefficients -1/2 1 0 -1/2 .  Finally, to compare levels 1 and 2 with levels 3 and 4, we use the coefficients 1/2 1/2 -1/2 -1/2.  These coefficients are used in the **estimate** statements below.

**proc glm data = c:\sasreg\hsb2;**

**class race;**

**model write = race;**

**estimate 'level 1 versus level 3' race 1 0 -1 0;**

**estimate 'level 2 versus levels 1 & 4' race -.5 1 0 -.5;**

**estimate 'levels 1 & 2 versus levels 3 & 4' race .5 .5 -.5 -.5;**

**run;**

**quit;**

Standard

Parameter Estimate Error t Value Pr > |t|

level 1 versus level 3 -1.74166667 2.73248820 -0.64 0.5246

level 2 versus levels 1 & 4 7.74324713 2.89718584 2.67 0.0082

levels 1 & 2 versus levels 3 & 4 1.10158046 1.96424409 0.56 0.5756

The contrast estimate for the first comparison is the mean of level 1 minus the mean for level 3, and the significance of this is .525, i.e., not significant.  The second contrast estimate is 7.743, which is the mean of level 2 minus the mean of level 1 and level 4, and this difference is significant, p = 0.008.  The final contrast estimate is 1.1 which is the mean of levels 1 and 2 minus the mean of levels 3 and 4, and this contrast is not statistically significant, p = .576.

**Method 2: Regression**

As in the prior example, we will make the following three comparisons:

1) level 1 to level 3,    
2) level 2 to levels 1 and 4 and     
3) levels 1 and 2 to levels 3 and 4.

For methods 1 and 2 it was quite easy to translate the comparisons we wanted to make into contrast codings, but it is not as easy to translate the comparisons we want into a regression coding scheme.  If we know the contrast coding system, then we can convert that into a regression coding system using the SAS program shown below. As you can see, we place the three contrast codings we want into the matrix **c** and then perform a set of matrix operations on **c,** yielding the matrix **x**. We then display **x** using the **print** command.

**proc iml;**

**c = { 1 -.5 .5,**

**0 1 .5,**

**-1 0 -.5,**

**0 -.5 -.5 };**

**x = c\*inv( c`\*c );**

**print x;**

**run;**

**quit;**

 Below we see the output from this program showing the regression coding scheme we would use.

X

-0.5 -1 1.5

0.5 1 -0.5

-1.5 -1 1.5

1.5 1 -2.5

This converted the contrast coding into the regression coding that we need for running this analysis with **proc reg**.  Below, we use **if-then** statements to create **x1**, **x2** and **x3** according to the coding shown above and then enter them into the regression analysis.

**data special;**

**set c:\sasreg\hsb2;**

**if race = 1 then x1 = -0.5;**

**if race = 2 then x1 = .5;**

**if race = 3 then x1 = -1.5;**

**if race = 4 then x1 = 1.5;**

**if race = 1 or race = 3 then x2 = -1;**

**if race = 2 or race = 4 then x2 = 1;**

**if race = 1 or race = 3 then x3 = 1.5;**

**if race = 2 then x3 = -.5;**

**if race = 4 then x3 =-2.5;**

**run;**

**proc reg data = special;**

**model write = x1 x2 x3;**

**run;**

**quit;**

The first comparison of the mean of the dependent variable for level 1 to level 3 of the categorical variable was not statistically significant, while the comparison of the mean of the dependent variable for level 2 to that of levels 1 and 4 was.  The comparison of the mean of the dependent variable for levels 1 and 2 to that of levels 3 and 4 also was not statistically significant.

Parameter Estimates

Parameter Standard

Variable Label DF Estimate Error t Value Pr > |t|

Intercept Intercept 1 51.67838 0.98212 52.62 <.0001

x1 1 -1.74167 2.73249 -0.64 0.5246

x2 1 7.74325 2.89719 2.67 0.0082

x3 1 1.10158 1.96424 0.56 0.5756

**5.9 Summary**

This page has described a number of different coding systems that you could use for categorical data, and two different strategies you could use for performing the analyses.  You can choose a coding system that yields comparisons that make the most sense for testing your hypotheses.  In general we would recommend using the easiest method that accomplishes your goals.