stat > r > library > contrast_coding.htm

R Library: Contrast Coding Systems for categorical variables

A categorical variable of K categories is usually entered in a regression analysis as a sequence of K-1 variables, e.g. as a sequence of K-1 dummy variables. Subsequently, the regression coefficients of these K -1 variables correspond to a set of linear hypotheses on the cell means. When coding categorical variables, there are a variety of coding systems we can choose for testing different set of linear hypotheses. On this page, we will cover some of the coding schemes for categorical variables. We will show how these coding schemes are constructed and interpreted.

In R there are four built-in contrasts (dummy, deviation, helmert, orthogonal polynomial) which we will demonstrate. We will also show how to create the coding schemes using a little bit of matrix manipulation. This page is done using R 2.11 and is updated in January, 2011.

Coding schemes covered

Coding Scheme	Comparisons made
Dummy Coding	Compares each level to the reference level, intercept being the cell mean of the reference group
Simple Coding	Compares each level to the reference level, intercept being the grand mean
Deviation Coding	Compares each level to the grand mean
Orthogonal Polynomial Coding	Orthogonal polynomial contrasts
Helmert Coding	Compare levels of a variable with the mean of the subsequent levels of the variable
Reverse Helmert Coding	Compares levels of a variable with the mean of the previous levels of the variable
Forward Difference Coding	Compares adjacent levels of a variable (each level minus the next level)
Backward Difference Coding	Compares adjacent levels of a variable (each level minus the prior level)
User-Defined Coding	User-defined contrast

The Example Data File

The examples in this page will use data frame called **hsb2** and we will focus on the categorical variable **race**, which has four levels (1 = Hispanic, 2 = Asian, 3 = African American and 4 = Caucasian) 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 a contrast matrix with one less column than levels of the original variable. In our example, our categorical variable has four levels so we will have contrast matrices with three columns and four rows.
First, we read in the data frame over the internet and then we create a factor variable, **race.f**, based on **race**.

```
hsb2 = read.table('http://www.ats.ucla.edu/stat/data/hsb2.csv', header=T, sep=",")

#creating the factor variable race.f
hsb2$race.f = factor(hsb2$race, labels=c("Hispanic", "Asian", "African-Am", "Caucasian"))
```

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.

```
tapply(hsb2$write, hsb2$race.f, mean)
 Hispanic Asian African-Am Caucasian
 46.45833    58       48.2  54.05517
```

1. Dummy Coding

Dummy coding is probably the most commonly used coding scheme. It compares each level of the categorical variable to a fixed reference level. For example, we can choose race = 1 as the reference group and compare the mean of variable **write** for each level of race 2, 3 and 4 to the reference level of 1. This is the default for unordered factors in R.

Dummy Coding

Level of race	race.f1 (1 vs. 2)	race.f2 (1 vs. 3)	race.f3 (1 vs. 4)
1 (Hispanic)	0	0	0
2 (Asian)	1	0	0
3 (African American)	0	1	0
4 (Caucasian)	0	0	1

```
#the contrast matrix for categorical variable with four levels
contr.treatment(4)
  2 3 4
1 0 0 0
2 1 0 0
3 0 1 0
```

```
4 0 0 1

#assigning the treatment contrasts to race.f
contrasts(hsb2$race.f) = contr.treatment(4)
#the regression
summary(lm(write ~ race.f, hsb2))

Residuals:
    Min     1Q Median 3Q  Max
 -23.06 -5.458 0.9724  7 18.8

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept) 46.4583  1.8422    25.2184  0.0000
    race.f2 11.5417  3.2861     3.5122  0.0006
    race.f3  1.7417  2.7325     0.6374  0.5246
    race.f4  7.5968  1.9889     3.8197  0.0002

Residual standard error: 9.025 on 196 degrees of freedom
Multiple R-Squared: 0.1071
F-statistic: 7.833 on 3 and 196 degrees of freedom, the p-value is 0.00005785
```

The parameter estimate for the first contrast compares the mean of the dependent variable, **write**, for levels 1 and 2 yielding 11.5417 and is statistically significant (p<.000). The t-value associated with this test is 3.5122. The results of the second contrast, comparing the mean of **write** for levels 1 and 3. The expected difference in variable **write** between group 1 and 3 is 1.7417 and is not statistically significant (t = 0.6374, p = .5246), while the third contrast is statistically significant. Notice that the intercept corresponds to the cell mean for **race** = Hispanic group.

2 Simple Coding

The results of simple coding are very similar to dummy coding in that each level is compared to the reference level. The difference in the regression output between dummy coding and simple coding scheme is in the intercepts. In the dummy coding scheme, the intercept corresponds to the cell mean of the reference group, while in the simple coding scheme, the intercept corresponds to the mean of cell means.

In our example below, level 1 is the reference level and **race.f1** compares level 1 to level 2, **race.f2** compares level 1 to level 3, and **race.f3** compares level 1 to level 4. For **race.f1** the coding is 3/4 for level 2, and -1/4 for all other levels. Likewise, for **race.f2** the coding is 3/4 for level 3, and -1/4 for all other levels, and for **race.f3** the coding is 3/4 for level 4, and -1/4 for all other levels. The general rule is that the reference group is never coded anything but -1/4 and for each contrast the level that is being contrasted is coded 3/4. Thus, for the first contrast it is level 2 which is coded 3/4 and all other level are -1/4. Since there are four groups and the values have to add to one there must be three levels coded as -1/4 and one level as 3/4.

SIMPLE Coding

Level of race	race.f1 (1 vs 2)	race.f2 (1 vs.3)	race.f3 (1 vs. 4)
1 (Hispanic)	-1/4	-1/4	-1/4
2 (Asian)	3/4	-1/4	-1/4
3 (African American)	-1/4	3/4	-1/4
4 (Caucasian)	-1/4	-1/4	3/4

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 the case of the variable **race.f** k = 4).

SIMPLE Coding

Level of race	level 1 vs. 2	level 1 vs. 3	level 1 vs. 4
1 (Hispanic)	-1 / k	-1 / k	-1 / k
2 (Asian)	(k-1) / k	-1 / k	-1 / k
3 (African American)	-1 / k	(k-1) / k	-1 / k
4 (Caucasian)	-1 / k	-1 / k	(k-1) / k

Let's create the contrast matrix manually using the scheme shown above. Essentially, the difference between the dummy coding scheme and the simple coding scheme is a constant matrix whose each element is 1/k if our categorical variable has k levels.

```
#creating the contrast matrix manually by modifying the dummy coding scheme
c<-contr.treatment(4)
my.coding<-matrix(rep(1/4, 12), ncol=3)
my.simple<-c-my.coding
my.simple

      2     3     4
1 -0.25 -0.25 -0.25
2  0.75 -0.25 -0.25
3 -0.25  0.75 -0.25
4 -0.25 -0.25  0.75

contrasts(hsb2$race.f)<-my.simple
summary(lm(write~race.f, hsb2))

Residuals:
    Min     1Q Median 3Q  Max
 -23.06 -5.458 0.9724  7 18.8

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept) 51.6784  0.9821    52.6191  0.0000
    race.f1 11.5417  3.2861     3.5122  0.0006
    race.f2  1.7417  2.7325     0.6374  0.5246
    race.f3  7.5968  1.9889     3.8197  0.0002

Residual standard error: 9.025 on 196 degrees of freedom
Multiple R-Squared: 0.1071
F-statistic: 7.833 on 3 and 196 degrees of freedom, the p-value is 0.00005785
```

The interpretation of this output is almost the same as for the case of dummy coding. The difference lies in the intercept.
The intercept in this case is 51.6784 = (46.45833 + 58 + 48.2 + 54.05517)/4, which is the mean of cell means, sometimes referred as grand mean.

3 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 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.
We will not create the contrast matrix manually because the **contr.sum** function creates it for us.

DEVIATION Coding

Level of race	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 (Caucasian)	-1	-1	-1

```
#the contrast matrix for categorical variable with four levels
contr.sum(4)
  [,1] [,2] [,3]
1    1    0    0
2    0    1    0
3    0    0    1
4   -1   -1   -1

#assigning the deviation contrasts to race.f
contrasts(hsb2$race.f) = contr.sum(4)
#the regression
summary(lm(write ~ race.f, hsb2))

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  51.6784   0.9821    52.6191   0.0000
    race.f1  -5.2200   1.6314    -3.1997   0.0016
    race.f2   6.3216   2.1603     2.9263   0.0038
    race.f3  -3.4784   1.7323    -2.0079   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.

4 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. In R it is not necessary to compute these values since this contrast can be obtained for any categorical variable by using the **contr.poly** function. This is also the default contrast used for ordered factor variables.

For the purpose of illustration, let's create an ordered categorical variable based on the variable **read**. Notice that this is the only example on this page that does not use **race** as the categorical variable, since it is not ordered.

```
hsb2$readcat<-cut(hsb2$read, 4, ordered = TRUE)
table(readcat)
readcat
(28,40] (40,52] (52,64] (64,76]
     22      93      55      30

tapply(hsb2$write, hsb2$readcat, mean)
 (28,40]  (40,52]  (52,64]  (64,76]
42.77273 49.97849 56.56364 61.83333
```

POLYNOMIAL coding

Level of readcat	Linear (readcat.L)	Quadratic (readcat.Q)	Cubic (readcat.C)
1 (28,40]	-.671	.5	-.224
2 (40,52]	-.224	-.5	.671
3 (52,64]	.224	-.5	-.671
4 (64,76]	.671	.5	.224

```
#the contrast matrix for categorical variable with four levels
```

```
contr.poly(4)
             .L   .Q          .C
1 -0.6708204  0.5 -0.2236068
2 -0.2236068 -0.5  0.6708204
3  0.2236068 -0.5 -0.6708204
4  0.6708204  0.5  0.2236068

#assigning the orthogonal polynomial contrasts to readcat
contrasts(hsb2$readcat) = contr.poly(4)
summary(lm(write ~ readcat, hsb2))

Call:
lm(formula = write ~ readcat, data = hsb2)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.978  -5.824   1.227   5.436  17.022 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  52.7870     0.6339  83.268   <2e-16 ***
readcat.L    14.2587     1.4841   9.607   <2e-16 ***
readcat.Q    -0.9680     1.2679  -0.764    0.446    
readcat.C    -0.1554     1.0062  -0.154    0.877    
---
Signif. codes:  0 �***� 0.001 �**� 0.01 �*� 0.05 �.� 0.1 � � 1

Residual standard error: 7.726 on 196 degrees of freedom
Multiple R-squared: 0.3456,     Adjusted R-squared: 0.3356 
F-statistic: 34.51 on 3 and 196 DF,  p-value: < 2.2e-16
```

The regression results indicate a strong linear effect of **readcat** on the outcome variable **write**. There is not a significant quadratic effect nor a cubic effect of **readcat** on the outcome variable **write**.

5 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).

In R there is a built-in function, **contr.helmert**, which, up to a scale, corresponds to the reverse Helmert coding discussed in next section. We will create the contrast matrix for Helmert coding manually below.

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 Coding

	race.f1	race.f2	race.f3
Level of Race	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 (Caucasian)	-1/4	-1/3	-1/2

```
#helmert for factor variable with 4 levels
my.helmert = matrix(c(3/4, -1/4, -1/4, -1/4, 0, 2/3, -1/3, -1/3, 0, 0, 1/
        2, -1/2), ncol = 3)
my.helmert
      [,1]       [,2] [,3]
[1,]  0.75  0.0000000  0.0
[2,] -0.25  0.6666667  0.0
[3,] -0.25 -0.3333333  0.5
[4,] -0.25 -0.3333333 -0.5

#assigning the new helmert coding to race.f
contrasts(hsb2$race.f) = my.helmert
#the regression
summary(lm(write ~ race.f, hsb2))

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  51.6784   0.9821    52.6191   0.0000
    race.f1  -6.9601   2.1752    -3.1997   0.0016
    race.f2   6.8724   2.9263     2.3485   0.0198
    race.f3  -5.8552   2.1528    -2.7198   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. 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.

6 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.

The regression coding for reverse Helmert coding is shown below. For the first comparison, where the first and second level are compared, **race.f1** is coded -1/2 and 1/2 and 0 otherwise. For the second comparison, the values of **race.f2** are coded -1/3 -1/3 2/3 and 0. Finally, for the third comparison, the values of **race.f3** are coded -1/4 -1/4 -/14 and 3/4. The built-in Helmert coding in R is equivalent to this coding scheme up to a constant in each column. So both give the same significance test. We will show both ways below

Reverse HELMERT Coding

	race.f1	race.f2	race.f3
Level of Race	Level 2 v. 1	Level 3 v. 1 and 2	Level 4 v. 1, 2 and 3
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 (Caucasian)	0	0	3/4

```
# manually reverse helmert for factor variable with 4 leves
my.rev.helmert = matrix(c(-1/2, 1/2, 0, 0, -1/3, -1/3, 2/3, 0, -1/4, -1/4,
        -1/4, 3/4), ncol = 3)
my.rev.helmert
     [,1]       [,2]  [,3]
[1,] -0.5 -0.3333333 -0.25
[2,]  0.5 -0.3333333 -0.25
[3,]  0.0  0.6666667 -0.25
[4,]  0.0  0.0000000  0.75

#assigning the reverse helmert coding to race.f
contrasts(hsb2$race.f) = my.rev.helmert
#the regression
summary(lm(write ~ race.f, hsb2))

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  51.6784   0.9821    52.6191   0.0000
    race.f1  11.5417   3.2861     3.5122   0.0006
    race.f2  -4.0292   2.6024    -1.5483   0.1232
    race.f3   3.1691   1.4880     2.1298   0.0344

# using built-in reverse Helmert coding
contrasts(hsb2$race.f) = contr.helmert(4)
summary(lm(write ~ race.f, hsb2))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  51.6784     0.9821  52.619  < 2e-16 ***
race.f1       5.7708     1.6431   3.512 0.000552 ***
race.f2      -1.3431     0.8675  -1.548 0.123170
race.f3       0.7923     0.3720   2.130 0.034439 *
```

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.

7 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.

For the first comparison, where the first and second levels are compared, **race.f1** 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, **race.f2** is coded 1/2 1/2 -1/2 -1/2, and for the third comparison where level 3 is compared with level 4, **race.f3** is coded 1/4 1/4 1/4 -3/4.

FORWARD DIFFERENCE coding

	race.f1	race.f2	race.f3
Level of race	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 (Caucasian)	-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

	contrast 1	contrast 2	contrast 3
Level of race	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 (Caucasian)	-1/k	-2/k	-3/k

```
#forward difference for factor variable with 4 leves
my.forward.diff = matrix(c(3/4, -1/4, -1/4, -1/4, 1/2, 1/2, -1/2, -1/2, 1/
        4, 1/4, 1/4, -3/4), ncol = 3)
my.forward.diff
      [,1] [,2]  [,3]
[1,]  0.75  0.5  0.25
[2,] -0.25  0.5  0.25
[3,] -0.25 -0.5  0.25
[4,] -0.25 -0.5 -0.75

#assigning the forward difference coding to race.f
contrasts(hsb2$race.f) = my.forward.diff
#the regression
summary(lm(write ~ race.f, hsb2))

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  51.6784   0.9821    52.6191   0.0000
    race.f1 -11.5417   3.2861    -3.5122   0.0006
    race.f2   9.8000   3.3878     2.8927   0.0043
    race.f3  -5.8552   2.1528    -2.7198   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.

8 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.

For the first comparison, where the first and second levels are compared, **race.f1** 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, **race.f2** is coded -1/2 -1/2 1/2 1/2, and for the third comparison where level 3 is compared with level 4, **race.f3** is coded -1/4 -1/4 -1/4 3/4.

BACKWARD DIFFERENCE Coding

	race.f1	race.f2	race.f3
Level of race	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 (Caucasian)	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 Coding

	contrast 1	contrast 2	contrast 3
Level of race	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 (Caucasian)	1/k	2/k	3/k

```
#backward difference for factor variable with 4 leves
my.backward.diff = matrix(c(-3/4, 1/4, 1/4, 1/4, -1/2, -1/2, 1/2, 1/2,
 -1/4, -1/4, -1/4, 3/4), ncol = 3)
my.backward.diff
      [,1] [,2]  [,3]
[1,] -0.75 -0.5 -0.25
[2,]  0.25 -0.5 -0.25
[3,]  0.25  0.5 -0.25
[4,]  0.25  0.5  0.75

#assigning the backward difference coding to race.f
contrasts(hsb2$race.f) = my.backward.diff
#the regression
summary(lm(write ~ race.f, hsb2))
```

```
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  51.6784   0.9821    52.6191   0.0000
    race.f1  11.5417   3.2861    -3.5122   0.0006
    race.f2  -9.8000   3.3878    -2.8927   0.0043
    race.f3   5.8552   2.1528     2.7198   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.

9 User Defined Coding

In R it is possible to use 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. All these have been expressed in terms of cell means. For convenience of the calculation, we will assume that the intercept corresponds to the mean of cell means. So our initial contrast matrix looks like the following, defined as **mat**. The final contrast matrix (or coding scheme) turns out to be the inverse of **mat** transposed.

```
#initial contrast matrix including the constant term
mat = matrix(c(1/4, 1/4, 1/4, 1/4, 1, 0, -1, 0, -1/2, 1, 0, -1/2, -1/2, -1/2, 1/2, 1/2), ncol
mat
     [,1] [,2] [,3] [,4]
[1,] 0.25    1 -0.5 -0.5
[2,] 0.25    0  1.0 -0.5
[3,] 0.25   -1  0.0  0.5
[4,] 0.25    0 -0.5  0.5

mymat = solve(t(mat))
mymat
     [,1] [,2] [,3] [,4]
[1,]    1 -0.5   -1 -1.5
[2,]    1  0.5    1  0.5
[3,]    1 -1.5   -1 -1.5
[4,]    1  1.5    1  2.5

#remove the intercept (constant) term
my.contrasts<-mymat[,2:4]
contrasts(hsb2$race.f) = my.contrasts
summary(lm(write ~ race.f, hsb2))

Call:
lm(formula = write ~ race.f, data = hsb2)

Residuals:
     Min       1Q   Median       3Q      Max
-23.0552  -5.4583   0.9724   7.0000  18.8000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  51.6784     0.9821  52.619   <2e-16 ***
race.f1       1.7417     2.7325   0.637   0.5246
race.f2       6.8724     2.9263   2.348   0.0198 *
race.f3      -2.8432     1.9642  -1.448   0.1494
---
Signif. codes:  0 �***� 0.001 �**� 0.01 �*� 0.05 �.� 0.1 � � 1

Residual standard error: 9.025 on 196 degrees of freedom
Multiple R-squared: 0.1071,     Adjusted R-squared: 0.0934
F-statistic: 7.833 on 3 and 196 DF,  p-value: 5.785e-05
```

The intercept corresponds to the mean of the cell means as shown earlier. The coefficient for **race.f1** corresponds to the difference in the mean of the variable **write** between group 3 and group 1. Similarly, the coefficient for **race.f2** corresponds to the difference between group 2 and the mean of group 1 and 4, and so on.

How to cite this page   Report an error on this page or leave a comment

The content of this web site should not be construed as an endorsement of any particular web site, book, or software product by the University of California.

IDRE RESEARCH TECHNOLOGY GROUP

High Performance Computing

Statistical Computing

GIS and Visualization

High Performance Computing
Hoffman2 Cluster
Hoffman2 Account Application
Hoffman2 Usage Statistics
UC Grid Portal
UCLA Grid Portal
Shared Cluster & Storage

GIS
Mapshare
Visualization
3D Modeling
Technology Sandbox
Tech Sandbox Access
Data Centers

Statistical Computing
Classes
Conferences
Reading Materials
IDRE Listserv
IDRE Resources
Social Sciences Data Archive

ABOUT CONTACT NEWS EVENTS OUR EXPERTS