HOW DO I INTERPRET THE PARAMETER ESTIMATES FOR DUMMY VARIABLES IN PROC REG OR PROC GLM? | SAS FAQ

Consider this simple data file having nine subjects (**sub**) in three groups (**iv**) with a score on the dv (**dv**).

**DATA dummy;**

**INPUT sub iv dv;**

**CARDS;**

**1 1 48**

**2 1 49**

**3 1 50**

**4 2 17**

**5 2 20**

**6 2 23**

**7 3 28**

**8 3 30**

**9 3 32**

**;**

**RUN;**

Below we do a **proc means** to find the overall mean, and another **proc means** to find the means for the three groups.

**PROC MEANS DATA=dummy;**

**VAR dv;**

**RUN;**

**PROC MEANS DATA=dummy;**

**CLASS iv;**

**VAR dv;**

**RUN;**

As we see below, the overall mean is 33, and the means for groups 1, 2 and 3 are 49, 20 and 30 respectively.

Analysis Variable : DV

N Mean Std Dev Minimum Maximum

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

9 33.0000000 12.8937970 17.0000000 50.0000000

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

Analysis Variable : DV

IV N Obs N Mean Std Dev Minimum Maximum

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

1 3 3 49.0000000 1.0000000 48.0000000 50.0000000

2 3 3 20.0000000 3.0000000 17.0000000 23.0000000

3 3 3 30.0000000 2.0000000 28.0000000 32.0000000

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

Let’s run a standard ANOVA on this data using **proc glm**.

**PROC GLM DATA=dummy;**

**CLASS iv ;**

**MODEL dv = iv ;**

**RUN;**

The results of the ANOVA are shown below.

General Linear Models Procedure

Class Level Information

Class Levels Values

IV 3 1 2 3

Number of observations in data set = 9

General Linear Models Procedure

Dependent Variable: DV

Sum of Mean

Source DF Squares Square F Value Pr > F

Model 2 1302.0000000 651.0000000 139.50 0.0001

Error 6 28.0000000 4.6666667

Corrected Total 8 1330.0000000

R-Square C.V. Root MSE DV Mean

0.978947 6.546203 2.1602469 33.000000

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

IV 2 1302.0000000 651.0000000 139.50 0.0001>

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

IV 2 1302.0000000 651.0000000 139.50 0.0001

Now, let’s take this information we have found, and relate it to the results that we get when we run a similar analysis using dummy coding. Let’s make a data file called **dummy2** that has dummy variables called **iv1** (1 if iv=1), **iv2** (1 if iv=2) and **iv3** (1 if iv=3).  Note that **iv3** is not really necessary, but it could be useful for further exploring the meaning of dummy variables. We will then use **proc reg** to predict **dv** from **iv1** and **iv2**.

**DATA dummy2;**

**SET dummy;**

**IF (iv = 1) THEN iv1 = 1; ELSE iv1 = 0;**

**IF (iv = 2) THEN iv2 = 1; ELSE iv2 = 0;**

**IF (iv = 3) THEN iv3 = 1; ELSE iv3 = 0;**

**RUN;**

**PROC REG DATA=dummy2;**

**MODEL dv = iv1 iv2 ;**

**RUN;**

The output is shown below.

Model: MODEL1

Dependent Variable: DV

Analysis of Variance

Sum of Mean

Source DF Squares Square F Value Prob>F

Model 2 1302.00000 651.00000 139.500 0.0001

Error 6 28.00000 4.66667

C Total 8 1330.00000

Root MSE 2.16025 R-square 0.9789

Dep Mean 33.00000 Adj R-sq 0.9719

C.V. 6.54620

Parameter Estimates

Parameter Standard T for H0:

Variable DF Estimate Error Parameter=0 Prob > |T|

INTERCEP 1 30.000000 1.24721913 24.054 0.0001

IV1 1 19.000000 1.76383421 10.772 0.0001

IV2 1 -10.000000 1.76383421 -5.669 0.0013

First, note that from the ANOVA using**proc glm** that the F value was 139.5 and for the regression using **proc reg** the F value (for the model) is also 139.5. This illustrates that the overall test of the model using regression is really the same as doing an ANOVA.  
After the **Analysis of Variance** section, there is a section titled **Parameter Estimates**. What is the interpretation of the values listed there, the 30, 19 and -10? Notice how we have **iv1** and **iv2** that refer to group 1 and group 2, but we did not include any dummy variable referring to group 3. Group 3 is often called the **omitted group** or **reference group**. Recall that the means of the 3 groups were 49, 20 and 30 respectively. The **intercept** term is the mean of the **omitted group**, and indeed the parameter estimate from the output is the mean of group 3, 30. The parameter estimate for **iv1** is the mean of group 1 minus the mean of group 3, 49 – 30 = 19, and indeed that is the parameter estimate for **iv1**. Likewise, the parameter estimate for **iv2** is the mean of group 2 – the mean of group 3, 20 – 30 = -10, the parameter estimate for**iv2**.  
So, in summary:

|  |  |
| --- | --- |
| **Intercept** | mean of group 3 (mean of omitted group) |
| **iv1** | mean of group 1 – group 3 (omitted group) |
| **iv2** | mean of group 2 – group 3 (omitted group) |

Try running this example, but use **iv2** and **iv3** in **proc reg** (making group 1 the omitted group) and see what happens.

Finally, consider how the parameter estimates can be used in the regression model to obtain the means for the groups (the predicted values). The regression model is:

Ypredicted = 30 + iv1\*19 + iv2\*-10

For group 1: Ypredicted = 30 + 1 \* 19 + 0 \* -10 = 49

For group 2: Ypredicted = 30 + 0 \* 19 + 1 \* -10 = 20

For group 3: Ypredicted = 30 + 0 \* 19 + 0 \* -10 = 30

As you see, the regression formula predicts that each group will have the mean value of its group.