How are estimated marginal means calculated?

Michael Ko

1 Introduction

In a linear model with categorical variables, the table of model parameter estimates can be difficult to interpret. One approach to understand these estimates is to calculate the estimated marginal means (sometimes referred to as least square means, predicted means, or expected means). Most statistical software packages offer procedures to obtain predictions of the response variable for the different levels of categorical variables after fitting linear models. However, these procedures should be used carefully as the results obtained can be very different depending on the statistical software package used.

2 Example

Consider a simulated dataset containing information about employees of a company, with information on their salary, age, gender, and job category. The continuous variables are summarized in Table 2.1; the categorical variables are summarized in Tables 2.2 and 2.3.

Table 2.1: Mean and standard deviation of the continuous variables in the employee dataset

	Mean	SD
Salary	6806.43	3148.26
Age	39.16	45.71

Table 2.2: Summary of the job category variable

Values	Count	Proportion
0 (clerical)	227	0.479
1 (trainee)	168	0.354
2 (security)	32	0.068
3 (technical)	47	0.099

Table 2.3: Summary of the gender variable

Values	Count	Proportion
0 (male)	258	0.544
1 (female)	216	0.456

In this newsletter, we will investigate the relationship between salary and gender controlling for job category and age. Table 2.4 contains the results of a linear model with salary as the

dependent variable with gender, job category, and age as predictor variables. Note that in our example, we are applying dummy coding for categorical variables; we are considering the reference level to be the lowest level of these categorical variable (i.e. male (0) for gender and clerical (0) for job category). For more information about dummy coding, please refer to our *Dummy and Effect Coding Newsletter*.

Table 2.4: Linear model summary with salary as the response and age, gender, job as predictors.

Coefficient	Estimate	SE	p-value
Intercept (β_0)	6963.73	235.94	<0.001
Gender: female (β_1)	-2456.7	240.92	<0.001
Age (β_2)	0.81	2.52	0.747
Job: trainee (β_3)	1302.53	254.02	<0.001
Job: security (β_4)	167.83	481.22	0.727
Job: technical (β_5)	4613.43	407.11	<0.001

Coefficients obtained from the linear model are used to estimated marginal means. For *gender*, our independent variable of interest, 0 represents a male subject while as 1 represents female subject. But what values are used for the other variables in the model: *age* and *job category*? For continuous variables like *age*, marginal means procedures typically substitute the overall mean values for calculations (unless the user specifies otherwise); in our example, 39.16 is the average age.

For categorical variables, some software packages calculate marginal means as if the data is from a balanced population, while others assume an unbalanced population. The term "balanced population" means that the sample is uniformly split across the different bins of the categorical variable; in terms of the 4-valued categorical variable *job category*, that would mean that 25 percent of the population falls into each bin. Thus, the predicted salary values obtained for each job category would be weighted equally when calculating the marginal mean for each gender. For an unbalanced population, the predicted salaries would be weighted according to the distribution of jobs in the data (see proportions in Table 2.2).

We see that our data is not balanced in terms of the *job category* variable. The job category percentages range from 6.75 to 47.89 percent in the sample. Below we show how different software packages treat this categorical variable when calculating marginal means—specifically, whether they assume a balanced or unbalanced population.

2.1 Balanced Estimated Marginal Means

In R, SAS, SPSS, and JMP, the marginal means procedure by default assumes a balanced population.

To see this, we first calculate marginal means for each job category, for both male and female employees. We take the linear model equation and use the coefficients from Table 2.4, along with the appropriate values for gender (0 for males, 1 for females), age (the mean value, 39.16, and job category (1 for the indicated job, 0 for the others).

For example, a female trainee's predicted salary would be calculated as follows:

$$6963.73 + 1 \times (-2456.7) + 39.15 \times (0.81) + 1302.53 \times 1 + 167.83 \times 0 + (4613.43) \times 0 = 5841.3$$

Table 2.5 displays the marginal means for each job and gender combination.

Table 2.5: Marginal means for each job category, for each gender.

	Clerical	Trainee	Security	Technical
Male	6995.51	8298.03	7163.34	11608.93
Female	4538.81	5841.34	4706.64	9152.24

We can then obtain the marginal mean for each gender by averaging the marginal means across job categories. Taking an unweighted average of the marginal means for each job category, thus assuming a balanced population, yields the actual marginal means reported by R, SAS, SPSS, and JMP. For males, the marginal mean is

$$\frac{1}{4} \times (6995.51 + 8298.03 + 7163.34 + 11608.94) = 8516.5,$$

while for females the marginal mean is

$$\frac{1}{4} \times (6995.51 + 8298.03 + 7163.34 + 11608.94) = 6059.8$$

Alternatively, these marginal means can also be obtained directly from the coefficients of the linear equation by substituting $\frac{1}{4}$ for each level of the job variable. For males, we have

$$6963.7 + 0 \times (-2456.7) + 39.15 \times (0.81) + 1302.5 \times \frac{1}{4} + 167.8 \times \frac{1}{4} + (4613.4) \times \frac{1}{4} = 8516.5,$$

while for females this calculation is

$$6963.7 + 1 \times (-2456.7) + 39.15 \times (0.81) + 1302.5 \times \frac{1}{4} + 167.8 \times \frac{1}{4} + (4613.4) \times \frac{1}{4} = 6059.8.$$

Although R, SAS, JMP, and SPSS treat categorical variables as balanced, R and SAS have options to treat them as unbalanced (see Table 6).

2.2 Unbalanced Estimated Marginal Means

In Stata, the marginal means procedure assumes an unbalanced population by default.

In our example, instead of weighing the means for each job category equally, the marginal means from Table 2.5 are weighted according to the proportion of our sample in each job category (given in Table 2.2). The unbalanced marginal mean for males is

$$(0.4790) \times 6995.51 + (0.3544) \times 8298.03 + (0.0675) \times 7163.34 + (0.0992) \times 11608.94 = 7925.9,$$

and the unbalanced marginal mean for females is

$$(0.4790) \times 4538.81 + (0.3544) \times 5841.34 + (0.0675) \times 4706.64 + (0.0992) \times 9152.24 = 5469.2$$

These are the marginal means computed by Stata.

The same marginal means can be obtained directly from the coefficients of the linear equation by replacing each job category dummy variable with its corresponding proportion in our sample. For males, we have

$$6963.7 + 0 \times (-2456.7) + 39.15 \times (0.81) + 1302.5 \times 0.354 + 167.8 \times 0.068 + (4613.4) \times 0.099 = 7925.9,$$

and for females, we have

$$6963.7 + 1 \times (-2456.7) + 39.15 \times (0.81) + 1302.5 \times 0.354 + 167.8 \times 0.068 + (4613.4) \times 0.099 = 5469.2.$$

We see that marginal means in Stata assumes an unbalanced population using the distribution of the sample by default. However, by using the option *asbalanced*, Stata's margins command can replicate the behavior of other software packages and compute balanced marginal means. Table 6 summarizes these findings.

Table 2.6: Commands to compute estimated marginal means in each software package.

Software	Treatment of Categorical Variables	Command
R	Balanced (default)	`emmeans()`
R	Unbalanced	`emmeans (... , weights="proportional")`
SAS	Balanced (default)	`lsmeans`
SAS	Unbalanced	`lsmeans ... /om`
JMP	Balanced	Analyze, Fit Model, Effect Details
SPSS	Balanced	EMMEANS
Stata	Unbalanced (default)	`margins`
Stata	Balanced	`margins..., asbalanced`

For more information on how to use these methods, see also our handout on *Post-hoc Analyses*.

3 References

- Margins Manual from Stata: https://www.stata.com/manuals13/rmargins.pdf
- LSMeans from SAS: https://documentation.sas.com/?docsetId=statug&docsetTarget=statug_glm_syntax10.htm&docsetVersion=15.1&locale=en
- Emmeans package for R: https://cran.r-project.org/web/packages/emmeans/index.html
- Emmeans in SPSS: https://www.ibm.com/support/knowledgecenter/SSLVMB_24.0.0/spss/advanced/syn_mixed_emmeans.html
- LSMeans in JMP: https://www.jmp.com/support/help/en/15.0/index.shtml#page/jmp/effect-details.shtml

---