StatNews #93

How are Estimated Marginal Means Calculated?

December 2017

In a linear model with categorical variables, the table of model parameter estimates can be difficult to interpret for categorical variables. 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.

Consider the simulated dataset containing information about employees of a company, with information on their salary, age, gender, and job category (see variable information in Table 1).

Employee Dataset					
Variable	**Type**	**n**	**Mean**		**St. Dev**
Salary	Continuous	474	6806.44		3148.25
Age	Continuous	474	39.15		45.71
			Values		**Proportion**
Gender	Categorical	258	0	male	54.43%
		216	1	female	45.57%
Job Category	Categorical	227	0	clerical	47.90%
		168	1	trainee	35.44%
		32	2	security	6.75%
		47	3	technical	9.92%

Table 1: Overview of the data

We investigate the question: how is Salary associated with Gender controlling for Job Category and Age? Consider in table 2 the coefficients of the linear model used to answer this question, where Salary is the dependent variable, and Age, Gender, and Job Category are independent variables. Note that in our example, we are applying dummy coding for categorical variables, and 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.

Linear Model: Coefficients			
	Estimate	**SE**	**p-value**
Intercept (β_0)	6963.7	235.9	<0.001
Gender: female (β_1)	-2456.7	240.9	<0.001
Age (β_2)	0.81	2.52	0.747
Job: trainee (β_3)	1302.5	254	<0.001
Job: security (β_4)	167.8	481.2	0.727
Job: technical (β_5)	4613.4	407.1	<0.001

Table 2: Linear model summary with salary as the response and age, gender, job as fixed effects

To calculate marginal means, the procedure uses the coefficients obtained from the linear model. For Gender, our independent variable of interest, 0 is substituted for males and 1 for females. 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 substitutes the overall mean values for calculations (unless the user specifies otherwise); in our example, 39.15 is used for 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. A balanced population, in terms of the 4-valued categorical variable Job Category, would mean that 25% of the population falls into each category. 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 1).

We see that our data is not balanced in terms of the Job Category variable; the job category percentages range from 6.75% to 47.90% 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.

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, along with the appropriate values for Gender (0 for males, 1 for females), Age (the mean value, 39.15), 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:

$$6963.7 + \mathbf{1} * (-2456.7) + \mathbf{39.15} * (0.81) + 1302.5 * \mathbf{1} + 167.8 * \mathbf{0} + (4613.4) * \mathbf{0} = 5841.34$$

Below is a table for the marginal means for each job, for each gender (Table 3).

Estimated Marginal Means by Gender and Job				
	Job Category			
	Clerical	Trainee	Security	Technical
Male	6995.51	8298.03	7163.34	11608.94
Female	4538.81	5841.34	4706.64	9152.24

Table 3: Marginal means for each job category, for each gender

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 (see Table 4).

$$\boldsymbol{Males}: \quad \frac{6995.51 + 8298.03 + 7163.34 + 11608.94}{4} = 8516.5$$

$$\boldsymbol{Females}: \quad \frac{4538.81 + 5841.34 + 4706.64 + 9152.24}{4} = 6059.8$$

Linear Model: Marginal Means from R, SAS, SPSS, JMP	
Gender	**Marginal Mean**
Male	8516.5
Female	6059.8

Table 4: Marginal means by gender in R, SAS, SPSS, JMP

Alternatively these marginal means can also be obtained directly from the coefficients of the linear equation by substituting 0.25 or $\frac{1}{4}$ for each level of the job variable.

Marginal Means for Males (Gender = 0)

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

Marginal Means for Females (Gender = 1)

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

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

Unbalanced Estimated Marginal Means

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

In our example, instead of weighing the means for each job category equally, the marginal means for each job category from Table 3 are weighed according to the proportions in Table 1. The unbalanced marginal means are calculated below:

$$\boldsymbol{Males}: (0.4790) \times 6995.51 + (0.3544) \times 8298.03 + (0.0675) \times 7163.34 + (0.0992) \times 11608.94 = 7925.9$$

$$\boldsymbol{Females}: (0.4790) \times 4538.81 + (0.3544) \times 5841.34 + (0.0675) \times 4706.64 + (0.0992) \times 9152.24 = 5469.2$$

These calculations match the marginal means output that we get from Stata (see Table 5).

Linear Model: Marginal Means from Stata	
Gender	**Marginal Mean**
Male	7925.9
Female	5469.2

Table 5: Marginal means results from Stata

The same marginal means can be obtained directly from the coefficients of the linear equation by substituting for each job category the percentage that it represents in the sample.

Marginal Means for Male

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

Marginal Means for Female

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

We see that marginal means in Stata assumes an unbalanced population using the distribution of the sample by default. However, by using Stata command *asbalanced*, Stata can replicate what the other software do, and treat the data as balanced. Table 6 summarizes these findings.

Estimated marginal means are often used to make pairwise comparisons and contrasts between groups, and so it is important to know the assumptions that each software package uses behind its predictions. By default, in R, SAS, JMP, and SPSS, unbalanced data is treated as if it is balanced, whereas in Stata, the distribution of the sample is taken into account.

Software	Treatment of Categorical Variables	LSMEANS Command
R	Balanced (default)	*emmeans (…)* *lsmeans (…)*
	Unbalanced	*emmeans (… , weights="proportional")* *lsmeans (…, weights="proportional")*
SAS	Balanced (default)	*lsmeans …*
	Unbalanced	*lsmeans … /om*
JMP	Balanced	*Analyze → Fit Model*
SPSS	Balanced	*EMMEANS*
Stata	Unbalanced (default)	*margins…*
	Balanced	*margins…, asbalanced*

Table 6: Commands to compute estimated marginal means for different software

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

If you need assistance with estimated marginal means or have any other statistical consulting questions, please feel free to contact the statistical consultants at CSCU.

Author: Michael Ko