

# Postestimation Analysis with Stata by SPost13 commands of Survey Data analyzed by MNLM

Debora Giovannelli – debora.giovannelli@gmail.com

## 1-Introduction

Data coming back from a brand survey have been analyzed by a regression model for nominal outcomes, also known as the Multinomial Logit Model. The Multinomial Logit Model (MNLM) belongs to a multivariate version of Generalized Linear Models (GLM), a class of models popularized by McCullagh and Nelder (1982) and widely used in many different fields (Social Sciences, Biomedical Sciences, Epidemiology, Public Health, Genetic, Zoology, Education, but also Marketing Researches, Survey Analysis and Product/Process/Service Quality Control). The interpretation of these regression models requires a background knowledge that is not always common, especially in business application fields. **Data must be "readable" to anyone who has the responsibility to take serious decision**, which can strongly influence not only the business of a company but also the safety and the quality of its products/processes and services. The scope of this presentation is to show and highlight the advantages of the implementation of SPost13 commands, setup by J. Scott Long and J. Freese, as very useful tools for making easier the interpretation of results coming from the implementation of this regression model for nominal response variables.

## 2-Objectives

The interpretation of regression models for categorical response variables is complex because of their nonlinearity. Models for nominal and ordinal outcomes may be interpreted using odds ratios (for logit models) and quantities based on predicted probabilities (*predictive margins*). While odds ratios do not depend on the values of the predictors (multiplicative effects), the meaning of odds ratios in terms of probabilities depends on the values of all the regressors (the magnitude of probability change depends on the values of all the explicative variables in the model). Because of nonlinearity these models require postestimation analysis and computation of predicted probabilities and related quantities as marginal effects, in order to fully describe the effects of all predictors. **Methods for the interpretation of nonlinear regression models for categorical outcomes** have been proposed by **J. Scott Long and Jeremy Freese** [7]. The statistical analyses here referred have been implemented by **Stata®15.1** and **SPost13** (Stata postestimation commands for version 13), a suite of programs for the postestimation interpretation of regression models for categorical outcomes, developed by J.S. Long and J. Freese, with the object to **give evidence on how SPost13 postestimation commands make easier the interpretation of nonlinear models as the MNLM**.

## 3-Dataset Description and Explorative Data Analysis

These statistical analyses are based on data coming from a **survey** conducted for assessing **Customer orientation** in the **professional audio market**, and previously analyzed by modelling the probability of respondents choice (favourite brand selection) by a Multinomial Logit Model (**MNLM**), where some characteristics of the respondents where included as explicative variables [6]. The response variable **Brand** is a multilevel nominal categorical variable with **5 outcomes** (5 brands coded A, B, C, D, Others), while the two categorical explanatory variables, specified in the model, are a binary variable **X1 (Age)**, with two levels (age over 50 years old, age under or equal to 50 years old), and a categorical variable **X2 (PVM, Primary Vertical Market)** with 4 levels: Ent (Entertainment), GER (Government Institution, Educational, Religious Institutions), Oth (others), R&S (Rental & Staging).

**Variables description**  
codebook Brand X1 X2

Brand	X1	X2
type: numeric (long) label: Brand	type: numeric (long) label: X1	type: numeric (long) label: X2
range: [1,5] unique values: 5	range: [1,2] unique values: 2	range: [1,4] unique values: 4
tabulation: Freq. Numeric Label 243 1 A 156 2 B 45 3 C 194 4 D 103 5 Others	tabulation: Freq. Numeric Label 163 1 Ent 249 2 GER 161 3 Oth 168 4 R&S	tabulation: Freq. Numeric Label 163 1 Ent 249 2 GER 161 3 Oth 168 4 R&S
type: numeric (long) label: X1	type: numeric (long) label: X2	
range: [1,2] unique values: 2	range: [1,4] unique values: 4	
tabulation: Freq. Numeric Label 398 1 Over 50 343 2 50	tabulation: Freq. Numeric Label 32 15 3 31 8 55 13 4 35 23 35 8 5 27 22 26 17 8 20 11	
type: numeric (long) label: X1	type: numeric (long) label: X2	
range: [1,2] unique values: 2	range: [1,4] unique values: 4	
tabulation: Freq. Numeric Label 398 1 Over 50 343 2 50	tabulation: Freq. Numeric Label 32 15 3 31 8 55 13 4 35 23 35 8 5 27 22 26 17 8 20 11	

**Three-way cross-tabulation table**  
table X2 Brand, by(X1) center subwidth(12)

Age and PVM	A	B	C	D	Others
Over 50	Ent: 32 GER: 55 Oth: 35 R&S: 26	15 13 8 17	3 4 5 8	31 35 27 20	8 23 22 11
50	Ent: 17 GER: 32 Oth: 24 R&S: 22	28 31 11 33	5 8 3 3	14 37 14 9	10 11 12 6

## 4-Model fitting and Selection

**Estimation using mlogit command**  
The MNLM has been fit using **mlogit** command. The dependent variable Brand has 5 nominal outcomes (A, B, C, D, Others). The model has been parameterized setting **category A** as **base outcome** (reference group). The independent variables, both categorical, have been included in the model by using **factor-variable notation**.

- Four models have been fitted:
- Full model "mfull": with two regressors with interaction terms (saturated model)
  - Main model "mmain": model with two regressors X1 (Age) and X2 (PVM) but no interaction terms
  - Restricted model "mX1": model with the regressor X1 (X2 omitted)
  - Restricted model "mX2": model with the regressor X2 (X1 omitted)

The following table summarizes the Information Criteria for all fitted models: estimates stats m\*

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
mfull	741	-1103.296	-1063.711	32	2191.421	2338.877
mmain	741	-1103.296	-1069.208	20	2178.415	2270.575
mX1	741	-1103.296	-1085.362	8	2186.724	2223.588
mX2	741	-1103.296	-1085.931	16	2203.862	2277.59

**Estimation results for the main effects model**  
mlogit Brand ib(1), X1 ib(2), X2, base(1) vsquish nolag

Multinomial logit regression

Brand	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
A (base outcome)					
B					
X1					
s50	1.095761	.2167432	5.06	0.000	.6709519 1.52057
X2					
Ent	.604048	.2848636	2.12	0.034	.0457256 1.16237
GER	-.3798146	.3271769	-1.16	0.246	-1.02107 2634403
Oth	-.7279691	.2794202	-2.61	0.009	-1.083156 1.275623
R&S	-.11286183	.2285211	-0.63	0.523	-1.734036 -.8382894
C					
s50	.6500734	.3296256	1.97	0.049	-.0404191 1.296128
X1					
s50	-.005366	.4915536	-0.01	0.983	-.7628908 1.163964
X2					
Ent	.0258795	.4892302	0.05	0.958	-.9310341 .9627931
GER	-.9483166	.4319929	-2.28	0.024	-1.34976 1.767857
Oth	-.2304349	.3590409	-0.64	0.520	-1.080056 1.600642
D					
X1					
s50	-.1081333	.1966082	-0.55	0.582	-.2772116 .4934782
X2					
Ent	.1092942	.2610228	0.42	0.676	-.4023311 .6208595
GER	-.1677803	.2586654	-0.65	0.517	-.6747552 .3391945
Oth	-.0978976	.2720755	-0.36	0.719	-.6311588 .4353607
R&S	-.2357516	.1805266	-1.31	0.192	-.5899773 -.118074
Others					
X1					
s50	-.0348311	.2428998	-0.14	0.886	-.5109081 .4412418
X2					
Ent	-.0635457	.3420584	-0.19	0.853	-.7339678 .608765
Oth	-.3861689	.2958074	-1.31	0.192	-.1396031 .9659405
R&S	-.0961928	.3472282	-0.28	0.776	-.781976 .583415
Others	-.9251729	.2253282	-4.10	0.000	-1.36722 4831262

## 5-SPost13 command fitstat

**Postestimation SPost13 command fitstat**  
The main effect model has been compared versus the full model by the SPost13 command fitstat.

	Current	Saved	Difference
Log-likelihood			
Model	-1069.296	-1063.711	-5.497
Intercept-only	-1103.296	-1103.296	0.000
Chi-square			
D(df=721/709/12)	2138.415	2127.421	10.994
LR(df=16/28/-12)	68.176	79.170	-10.994
p-value	0.000	0.000	0.529
R2			
McFadden	0.031	0.036	-0.005
McFadden(adjusted)	0.013	0.007	0.006
Cox-Snell/Mc	0.088	0.101	-0.013
Cragg-Uhler/Nagelkerke	0.093	0.107	-0.014
Count	0.358	0.364	-0.007
Count(adjusted)	0.044	0.054	-0.010
AIC	2178.415	2191.421	-13.006
AIC divided by N	2.940	2.957	-0.018
BIC(df=20/32/-12)	2270.575	2338.877	-68.302

Note: Likelihood-ratio test assumes current model nested in saved model.

Log-likelihood

Model	Current	Saved	Difference
Model	-1069.296	-1063.711	-5.497
Intercept-only	-1103.296	-1103.296	0.000

Chi-square

Model	Current	Saved	Difference
D(df=721/709/12)	2138.415	2127.421	10.994
LR(df=16/28/-12)	68.176	79.170	-10.994
p-value	0.000	0.000	0.529

R2

Model	Current	Saved	Difference
McFadden	0.031	0.036	-0.005
McFadden(adjusted)	0.013	0.007	0.006
Cox-Snell/Mc	0.088	0.101	-0.013
Cragg-Uhler/Nagelkerke	0.093	0.107	-0.014
Count	0.358	0.364	-0.007
Count(adjusted)	0.044	0.054	-0.010

AIC

Model	Current	Saved	Difference
AIC	2178.415	2191.421	-13.006
AIC divided by N	2.940	2.957	-0.018
BIC(df=20/32/-12)	2270.575	2338.877	-68.302

Note: Likelihood-ratio test assumes current model nested in saved model.

**SPost13 command listcoef** provides in a single table the estimates for all the comparisons of outcome categories for each variable included in the model. By specific options the output may be suitably simplified.

## 6-SPost13 command listcoef

**Comparisons across categories by listcoef**  
listcoef, pvalue(0.05) positive

```
mlogit (N=741): Factor change in the odds of Brand (P<0.05)
```

Variable: 2.X1 (ado=0.499)	b	z	P> z	e*b	e*bstdx
B vs A	1.0958	5.056	0.000	2.991	1.728
B vs D	0.9976	4.384	0.000	2.685	1.437
B vs Others	1.1306	4.213	0.000	3.097	1.758
C vs A	0.6501	1.972	0.049	1.916	1.383

Variable: 1.X2 (ado=0.415)

b	z	P> z	e*b	e*bstdx	
B vs A	0.6040	2.120	0.034	1.830	1.285

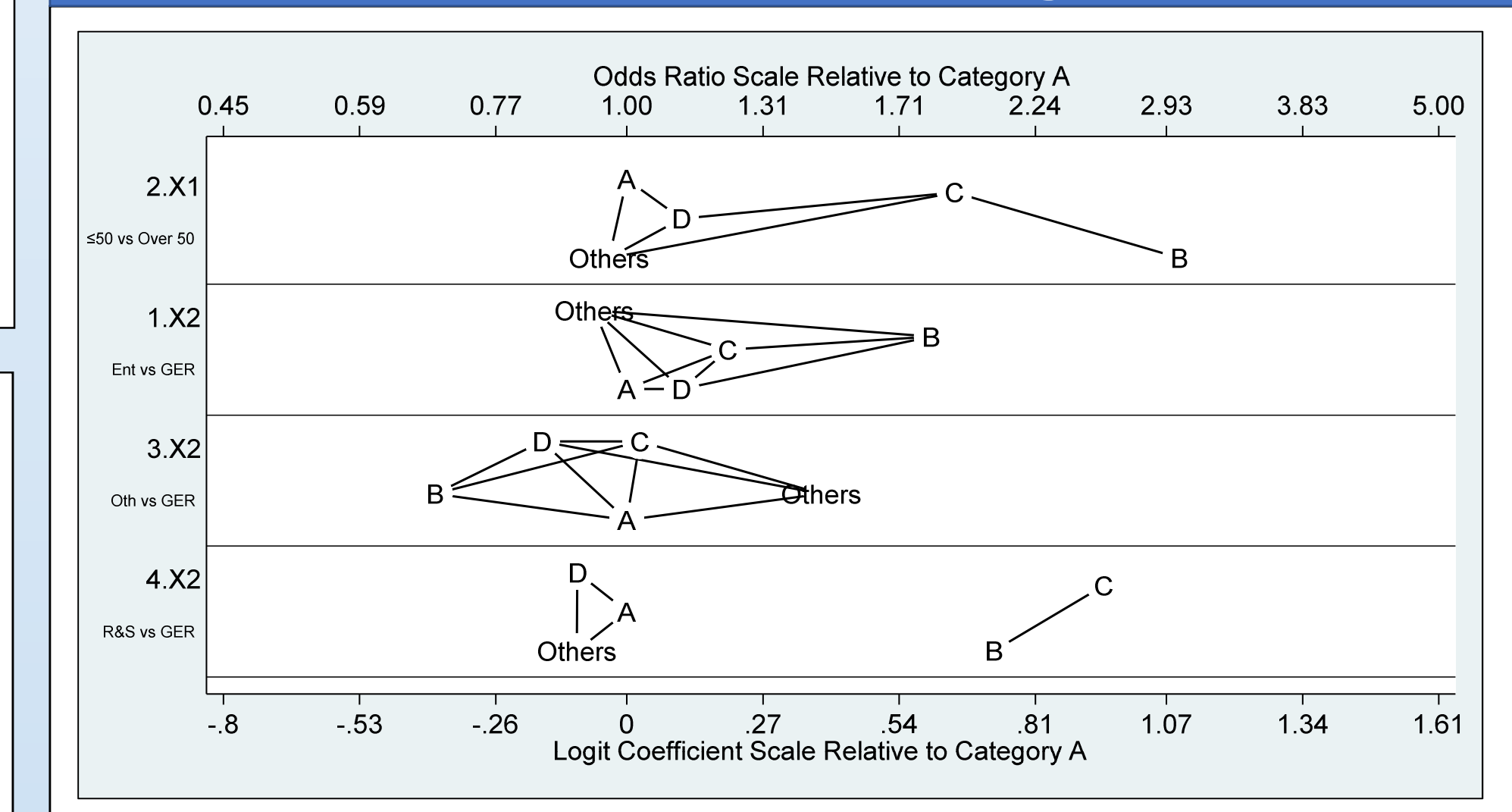
Variable: 3.X2 (ado=0.413)

b	z	P> z	e*b	e*bstdx	
Others vs B	0.7660	2.065	0.039	2.151	1.372

Variable: 4.X2 (ado=0.419)

b	z	P> z	e*b	e*bstdx	
B vs A	0.7080	2.605	0.009	2.071	1.357
B vs D	0.8259	2.802	0.005	2.284	1.413
B vs Others	0.8266	2.257	0.024	2.285	1.414
C vs A	0.9463	2.258	0.024	2.616	1.467
C vs D	1.0442	2.430	0.015	2.841	1.549
C vs Others	1.0449	2.171	0.030	2.843	1.549

## 7-SPost13 command mlogitplot



**SPost13 command mlogitplot** provides a plot that synthesizes the effects of all regressors on all contrasts in odds ratio scale and in logit scale, giving also evidence of their significance.

## 8-Interpretation in terms of Odds Ratios

- Based on this graph, we may conclude the following:
- for individuals with age s50, compared with individuals with age over 50, holding PVM constant, the odds of selecting brand C or B relative to brand A significantly increase by a factor of 1.92 for C and 2.99 for B, while for the other contrasts (D vs A and Others vs A) the effects are statistically not significant
  - for individuals with PVM = Ent, compared with individuals with PVM = GER, holding Age constant, just one contrast (B vs A), is statistically significant, with an increase by a factor of 1.83
  - for individuals with PVM = Oth, compared with individuals with PVM = GER, holding Age constant, all the contrasts respect to A category, are statistically not significant
  - for individuals with PVM = R&S, compared with individuals with PVM = GER, holding Age constant, the odds of selecting brand C and brand B relative to brand A significantly increase by a factor of 2.58 for C and 2.07 for B, while the odds of selecting the brands D and Others relative to brand A do not significantly change
- Moreover this graph provides evidence on the effects for all the other contrasts (different base outcomes). As an example:
- for individuals with PVM = R&S, when compared with individuals with PVM = GER, holding Age constant, the odds of selecting brand C, rather than one of the other brands, is significant for the contrasts C vs A, C vs D and C vs Others, while the contrast C vs B is not statistically significant (as provided by listcoef command output).

## 9-Interpretation based on Adjusted Predictions and Marginal Effects

The **MNLM** is linear in the logit but is **nonlinear in probability**: while the factor change in the odds is constant across the levels of all variables, the effect of each predictor on the probability of an outcome of the response variable depends on the value (for continuous predictors) or level (for categorical predictors) of the specific predictor and on the level of all the other independent variables specified in the model (marginal effects depends on the values of all variables). This makes the interpretation complex so that the evaluation of the effects of all the explicative variables for all the logits, just based on the estimated coefficients, represents a limitation. Moreover, models for nominal outcomes are even more complex because they provide more parameters to interpret respect to the models for ordinal outcomes, where constraints are imposed (the effect of each regressor is constrained to be equal in all equations). The interpretation using predictions as Predictive Margins or Adjusted Predictions and summary measures based on predictions as Marginal Effects is more informative for assessing the impact of each independent variable on each outcome of the response variable. **Adjusted Predictions and Marginal Effects** computation is provided by Stata command **margins**. Margins provides three different types of Marginal Effects (three different approaches of computation), which depends on the different ways of controlling for the other variables in the model while computing Adjusted Predictions:

- Average Marginal Effects (AMEs) are computed as difference between two Average Adjusted Predictions (AAPs)
- Marginal Effects at Means (MEMs) are computed as difference between two Adjusted Predictions at Means (APMs)
- Marginal Effects at Representative values (MERs) are computed as difference between two Adjusted Predictions at specific values of the other variables (APRs)

In this specific context where all regressors specified in the model are categorical, the use of **factor-variable notation** in model specification is critical in order to **guarantee correct results by using margins** (this way Stata recognizes any interdependencies between variables). Moreover, considering the categorical nature of both regressors, two types of marginal effects, **AMEs** and **MERs**, have been **computed** as statistics to **interpret the effect of the characteristics of the respondents on the choice of the favourite brand**.

## 10-SPost13 command mtbl for tabulating Predictive Margins

**Table of Adjusted Predictions**  
SPost13 command mtbl is a wrapper of margins: it uses margins for building tables of Adjusted Predictions and tables of Marginal Effects (dydx). If the outcome has multiple categories, mtbl automatically submits multiple margins commands for all outcomes and combines the results in the table. Results from multiple calls of mtbl may be combined into a single compact table:

```
. quietly mtbl, at (X2 = 1 X1 = 2) rown(PVM Ent Age s50) dec(4) below  
. quietly mtbl, at (X2 = 2 X1 = 2) rown(PVM Ger Age s50) dec(4) below  
. quietly mtbl, at (X2 = 3 X1 = 2) rown(PVM Oth Age s50) dec(4) below  
. quietly mtbl, at (X2 = 4 X1 = 2) rown(PVM R&S Age s50) dec(4) below  
. quietly mtbl, at (X2 = 1 X1 = 1) rown(PVM Ent Age Over 50) dec(4) below  
. quietly mtbl, at (X2 = 2 X1 = 1) rown(PVM Ger Age Over 50) dec(4) below  
. quietly mtbl, at (X2 = 3 X1 = 1) rown(PVM Oth Age Over 50) dec(4) below  
. quietly mtbl, at (X2 = 4 X1 = 1) rown(PVM R&S Age Over 50) dec(4) below
```

Expression: Pr(Brand), predict(outcome(i))

	A	B	C	D	Others
PVM Ent Age s50	0.2447	0.3700	0.0572	0.2402	0.0879
PVM Ger Age s50	0.3048	0.2519	0.0583	0.2683	0.1167
PVM Oth Age s50	0.3258	0.1842	0.0639	0.2425	0.1836
PVM R&S Age s50	0.2299	0.3936	0.1133	0.1835	0.0798
PVM Ent Age Over 50	0.3471	0.1755	0.0423	0.3059	0.1292
PVM Ger Age Over 50	0.3902	0.1078	0.0390	0.3083	0.1547
PVM Oth Age Over 50	0.3933	0.0743	0.0403	0.2627	0.2294
PVM R&S Age Over 50	0.3442	0.1970	0.0885	0.2466	0.1237

**SPost13 command mtbl** allows tabulating Adjusted Predictions for multiple outcomes in a compact table. By multiple calls of mtbl and suitable labelling options, a synthetic and informative table may be provided.

## 11-SPost13 command mchange for AMEs computation

**AMEs** are marginal effects computed as difference between two Average Adjusted Predictions (AAPs).

```
. mchange  
mlogit: Changes in Pr(y) | Number of obs = 741  
Expression: Pr(Brand), predict(outcome(i))
```

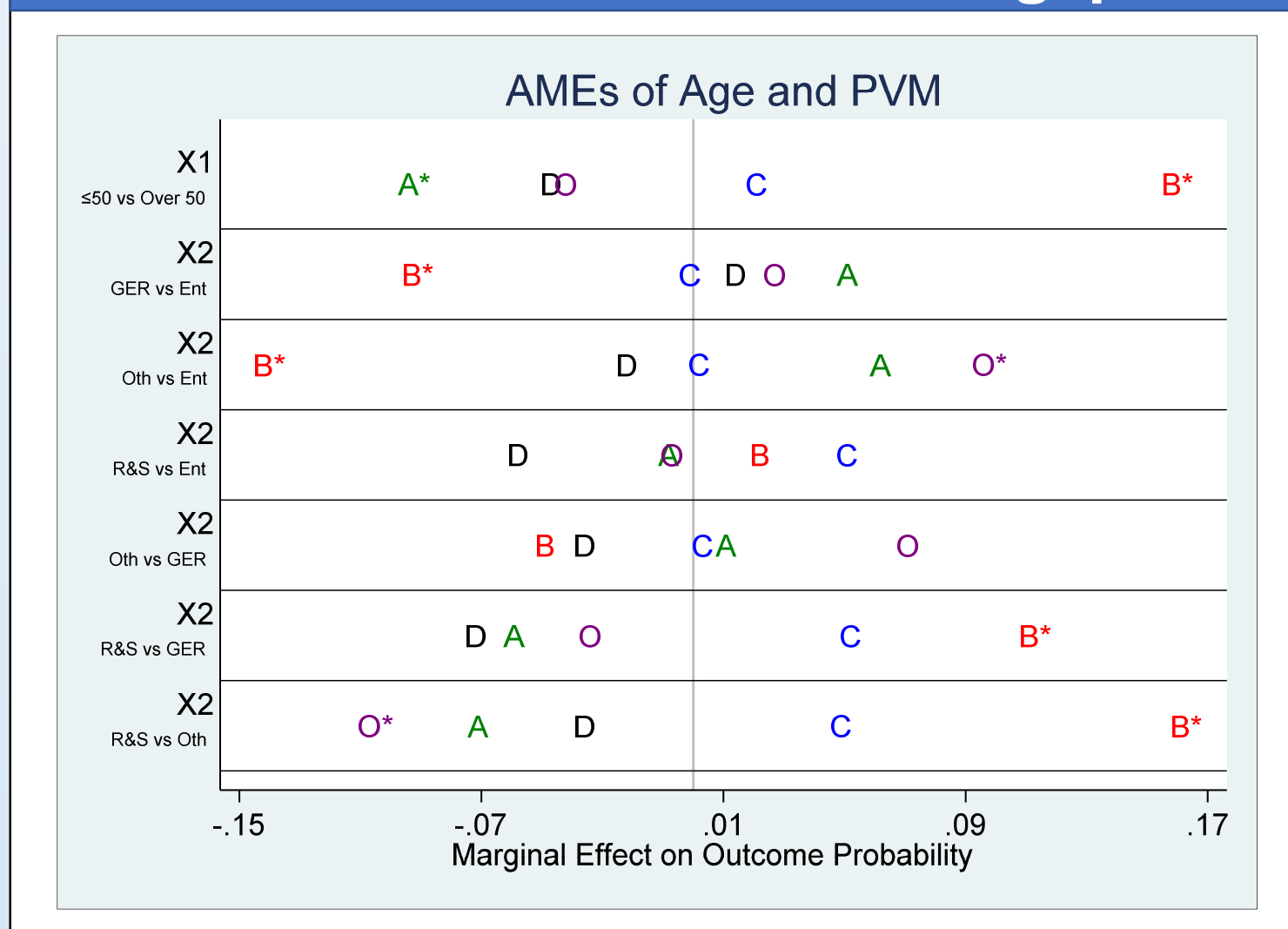
	A	B	C	D	Others
X1					
s50 vs Over 50	-0.092	0.160	0.021	-0.047	-0.042
p-value	0.007	0.000	0.248	0.149	0.097
X2					
GER vs Ent	0.051	-0.091	-0.001	0.014	0.027
p-value	0.275	0.027	0.952	0.752	0.409
Oth vs Ent	0.062	-0.140	0.002	-0.022	0.098
p-value	0.230	0.001	0.934	0.650	0.014
R&S vs Ent	-0.008	0.022	0.051	-0.058	-0.007
p-value	0.967	0.639	0.076	0.219	0.843
Oth vs GER	0.011	-0.049	0.003	-0.036	0.071
p-value	0.814	0.163	0.881	0.416	0.066
R&S vs GER	-0.059	0.113	0.052	-0.072	-0.034
p-value	0.199	0.006	0.051	0.092	0.294
R&S vs Oth	-0.071	0.163	0.049	-0.036	-0.105
p-value	0.170	0.000	0.093	0.443	0.008

Average predictions

	A	B	C	D	Others
Pr(y base)	0.328	0.211	0.061	0.262	0.139

**SPost13 command mchangeplot** provides a plot that synthesizes in terms of AMEs the effects of all regressors on all contrasts, giving also evidence of their significance.

## 12-SPost13 command mchangeplot



## 13-SPost13 command mtbl for MERs computation

**MERs** are marginal effects computed as difference between two Adjusted Predictions for a variable, conditioning at specific values of the other variables (APRs). These conditional MEs may be computed by mchange or mtbl SPost13 commands. When computing MERs of Age by mchange, conditioning on PVM, multiple calls of mchange are required, because just one value of PVM can be specified in at() option. To synthesize all the MERs in one single table **multiple calls of mtbl** have been submitted. When computing MERs for PVM by mtbl, conditioning on Age, two tables have been provided in order not to loose the labels of table's rows (one for the estimates and one for the p-values).

**MERs of Age**  
MEs of Age at specific levels of PVM

Expression: Marginal effect of Pr(Brand), predict(outcome(i))

	A	B	C	D	Others
Ent	-0.0431	0.0677	0.0034	-0.0024	-0.0256
Oth	0.0				