Stratified Models versus Interaction Terms: When are the results equivalent?

Suzanne Dufault

May 25, 2019

Abstract

Stratified models will return results that are derivable from a model with an interaction term only when the interaction of the stratifying variable and each of the main effects covariates is accounted for in the interaction model.

1 Data Generation

Consider the scenario where you have a count outcome Y (e.g. TDLU count) and are interested in its association with a binary variable A (e.g. HDP status). You are further interested in whether a 3-level variable G (e.g. genotype of GG, GT, or TT) modifies the association. Further, you adjust for some continuous confounder W (e.g. centered age).

As a structural equation model, we could write:

$$W \sim f_W(U_W) \tag{1}$$
$$G \sim f_G(U_G) \tag{2}$$
$$A \sim f_A(G, W, U_A) \tag{3}$$
$$Y \sim f_Y(G, W, A, U_Y) \tag{4}$$
$$\tag{5}$$

For the specific data generating distributions, see the code in the raw .Rnw file. For a sample of size 10, you would then have a dataset of the following form:

id	Y	A	G	W
id1	5	0	1	0.87
id2	25	1	0	-1.32
id3	20	1	0	-1.75
id4	3	0	2	0.53
id5	1	1	2	-2.22
id6	18	1	1	-3.22
id7	12	1	1	-2.34
id8	10	1	1	-3.26
id9	38	0	0	-4.11
id10	7	1	2	-3.21

Table 1: Example dataset.

2 Simple Example with Observed Descrepancy

To look for multiplicative interaction, you run the following models:

$$\log \mathrm{E}[Y|A,W,G] = \beta_0 + \beta_1 A + \beta_2 G + \beta_3 W + \beta_4 A \times G \qquad \text{Full Model} \tag{6}$$

(7)

$$\log \mathrm{E}[Y|W,G] = \gamma_0 + \gamma_1 G + \gamma_2 W \qquad \text{Stratified on } A = a \tag{8}$$

(9)

term	estimate	std.error	statistic	p.value
(Intercept)	10.18	0.06	39.64	0.00
A	1.25	0.07	3.03	0.00
G	0.79	0.05	-4.86	0.00
W	0.76	0.01	-21.06	0.00
A:G	0.54	0.08	-8.01	0.00

Table 2: Full model with interaction term.

term	estimate	std.error	statistic	p.value
(Intercept)	12.88	0.05	49.97	0.00
G	0.42	0.06	-14.50	0.00
W	0.77	0.02	-13.03	0.00

Table 3: Model with only A = 1.

term	estimate	std.error	statistic	p.value
(Intercept)	10.07	0.06	38.28	0.00
G	0.79	0.05	-4.79	0.00
W	0.75	0.02	-16.55	0.00

Table 4: Model with only A = 0.

For particular covariate combinations, we can compare whether the models produce the same results. For the following examples, we are estimating the expected log count of Y when $A = 1$, $G = 1$:

$$2.32 + 0.226 + -0.241 + -0.619 = 1.6864287 \qquad \text{Interaction Model} \tag{10}$$

$$2.556 + -0.86 = 1.6958754 \qquad \text{Stratified Model} \tag{11}$$

2.1 Changing Sample Size

Estimating the expected log count of Y when $A = 1$, $G = 1$, we see that increasing the sample size does not resolve the observed issue.

$$\log \mathrm{E}[Y|A,W,G] = \beta_0 + \beta_1 A + \beta_2 G + \beta_3 W + \beta_4 A \times G \qquad \text{Full Model} \tag{12}$$

$$\log \mathrm{E}[Y|W,G] = \gamma_0 + \gamma_1 G + \gamma_2 W \qquad \text{Stratified on } A = a \tag{13}$$

N	Full	Stratified
100	1.62004	1.61755
1000	1.65339	1.64617
10000	1.64269	1.64280

Table 5: Results from the full model with interaction term compared to the results from the stratified model in the relevant strata. Several sample sizes are shown.

3 Resolving the Discrepancy

Estimating the expected log count of Y when $A = 1$, $G = 1$, we can see that adding an interaction term for the other variable in the model (W) resolves the discrepancy. Therefore, the stratified models implicitly are estimating coefficients with respect to all covariates' interactions with the stratifying variable.

$$\log \mathrm{E}[Y|A, W, G] = \beta_0 + \beta_1 A + \beta_2 G + \beta_3 W + \beta_4 A \times G + \beta_5 A \times W \qquad \text{Full Model} \qquad (14)$$

$$\log \mathrm{E}[Y|W, G] = \gamma_0 + \gamma_1 G + \gamma_2 W \qquad \text{Stratified on } A = a \qquad (15)$$

N	Full	Stratified
100	1.61755	1.61755
1000	1.64617	1.64617
10000	1.64280	1.64280

Table 6: Results from the full model with interaction term compared to the results from the stratified model in the relevant strata. Several sample sizes are shown.