

Grove, W.M. (2004). The MAXSLOPE taxometric procedure: Mathematical derivation, parameter estimation, consistency tests. *Psychological Reports*, 95, 517–550.

ADDENDUM

Explanation of How Bias in Estimators of Latent Group Means is Removed

William M. Grove
University of Minnesota, Twin Cities
July 15, 2005

In Grove (2004, section entitled “Use of MAXSLOPE to Estimate Parameters of the Taxometric Model”), a multi-step procedure is given for estimating parameters of the common-slope two-component mixture regression model. The sixth step estimates μ_{cX} (i.e., the complement class mean on X) from the average of those X values whose Y values are in a region of Y where the regression of X on Y is relatively flat (once b_{YX} , the slope of the within-population linear regression of X on Y , has been estimated and the influence of X on Y within latent populations approximately removed). This addendum explains how bias in estimating latent population means (μ_{cX} , μ_{cY} , μ_{tX} , and μ_{tY}) is removed by the MAXSLOPE program code.

By definition, the bias in estimating μ_{cX} and μ_{cY} , using the procedure of Step 6, equals

$$\text{bias}(\tilde{\mu}_{cX}) = \tilde{\mu}_{cX} - \mu_{cX}$$

$$\text{bias}(\tilde{\mu}_{cY}) = \tilde{\mu}_{cY} - \mu_{cY}$$

where $\tilde{\mu}_{cX}$ is the biased estimate obtained via Step 6 of the algorithm. Briefly, this estimator is given by the local mean of X in a region of Y near the sample minimum of Y , where essentially all observations are expected to belong to the complement class, as long as P (the taxon base rate) does not greatly exceed $1/2$. In the MAXSLOPE code, by default X values corresponding to the lowest 5% of the sample values of Y are used; but this parameter is adjustable by the user. This value was chosen after examining results of small experiments with various simulated mixture regression datasets: Gaussian, gamma, and beta mixtures base rates having $.1 \leq P \leq .5$, separation between taxon and complement class means $1/2 \leq \delta \leq 3$ (measured in within-population pooled standard

deviations), and within-population standardized regression coefficients $0 \leq b_{YX} \leq 1/2$ and $0 \leq b_{XY} \leq 1/2$.

The Step 6 latent mean estimator is biased because it removes the regression of X on Y before using the local behavior of Y to estimate μ_{cX} , but Y itself has a regression on X that has not been removed. This uncompensated effect of X on Y still influences $\tilde{\mu}_{cY}$, because we rely on X values (contaminated by Y , as it were) to pick out the observations used for estimating μ_{cY} .

From now on, we will deal with estimating and correcting the bias in $\tilde{\mu}_{cX}$ alone, which will be intertwined with removing bias in $\tilde{\mu}_{cY}$. Dealing with $\tilde{\mu}_{cX}$ and $\tilde{\mu}_{cY}$ is precisely analogous.

Starting with the within-complement class regression of Y on X , the definition of a biased estimate, and solving for μ_{cX} , we obtain

$$\begin{aligned}\mu_{cX} &= \tilde{\mu}_{cX} - b_{YX} (\tilde{\mu}_{cY} - \text{bias}(\tilde{\mu}_{cY})) \\ &= \tilde{\mu}_{cX} - b_{YX} \tilde{\mu}_{cY} + b_{YX} \text{bias}(\tilde{\mu}_{cY}).\end{aligned}$$

We can solve the latter equation for an estimator of μ_{cX} less biased than $\tilde{\mu}_{cX}$ itself. This revised estimator is less biased because it removes part of the confounding effect of X (with which Y is correlated) from Y , before using these Y -values to choose relevant X -values for estimating the complement class mean on Y . This less biased μ_{cX} -estimator is given by plugging in an estimator of b_{YX} into the last equation above, temporarily neglecting the term in $\text{bias}(\tilde{\mu}_{cY})$, and solving for a second-iteration approximation of μ_{cX} we will denote $\tilde{\mu}_{cX}^{i=2}$:

$$\begin{aligned}\tilde{\mu}_{cX}^{i=2} &= \tilde{\mu}_{cX} - b_{YX} \tilde{\mu}_{cY} + \text{neglected} \\ &\approx \tilde{\mu}_{cX} - \tilde{b}_{YX} \tilde{\mu}_{cY}\end{aligned}$$

Where $\tilde{\mu}_{cX}$ can itself be regarded as a first-step iterative estimator of μ_{cX} . $\tilde{\mu}_{cX}^{i=2}$ still has some remaining bias due to the neglected term in $\text{bias}(\tilde{\mu}_{cY})$; this bias can be appreciable.

The neglected term obviously has expectation

$$E[\text{bias}(\tilde{\mu}_{cX}^{i=2})] = -b_{YX} \text{bias}(\tilde{\mu}_{cY})$$

if the error in estimating b_{YX} is presumed to nil, for the moment.

The immediately preceding equation suggests that we could again estimate and then remove a bias term, viz. $b_{YX} \text{bias}(\tilde{\mu}_{cY})$, to obtain a still less biased estimator. By definition,

$$\mu_{cY} = E[Y_c] \approx E[Y = b_0 + b_{XY}X + \varepsilon | X < X_t].$$

From the foregoing, it is clear that we can obtain an initial, albeit biased, estimate of μ_{cY} , the Y -mean for compliment class members, by examining the mean of Y -values for observations having especially low values of X , just as we looked for extremely low

values of Y to get an initial estimate of μ_{cX} .

With $\tilde{\mu}_{cY}$ in hand, we can then develop an equation precisely analogous to the one above for $\text{bias}(\tilde{\mu}_{cX})$, this estimating the bias in $\tilde{\mu}_{cY}$. This estimated bias then permits a less biased estimator $\tilde{\mu}_{cY}^{i=2}$. The remaining bias of $\tilde{\mu}_{cY}^{i=2}$ in turn has the same form as that for $\tilde{\mu}_{cX}^{i=2}$ given above. That is, from we obtain

$$\begin{aligned}\mu_{cY} &= \tilde{\mu}_{cY} - b_{XY} (\tilde{\mu}_{cX} - \text{bias}(\tilde{\mu}_{cX})) \\ &= \tilde{\mu}_{cY} - b_{XY} \tilde{\mu}_{cX} + b_{XY} \text{bias}(\tilde{\mu}_{cX}).\end{aligned}$$

Hence

$$\begin{aligned}\tilde{\mu}_{cY}^{i=2} &= \tilde{\mu}_{cY} - \tilde{b}_{XY} \tilde{\mu}_{cX} + \text{neglected}, \\ E[\text{bias}(\tilde{\mu}_{cY}^{i=2})] &\approx b_{XY} (E[\tilde{\mu}_{cY}^{i=2}] - E[\tilde{\mu}_{cY}]), \\ \widetilde{\text{bias}}(\tilde{\mu}_{cY}^{i=2}) &\approx b_{XY} (\tilde{\mu}_{cY}^{i=2} - \tilde{\mu}_{cY}).\end{aligned}$$

Thus an estimator of μ_{cX} that is less biased (in expectation) than either $\tilde{\mu}_{cX}$ or $\tilde{\mu}_{cX}^{i=2}$ is given by

$$\begin{aligned}\tilde{\mu}_{cX}^{i=3} &\approx \tilde{\mu}_{cX}^{i=2} - \tilde{b}_{YX} \widetilde{\text{bias}}(\tilde{\mu}_{cY}) \\ &\approx \tilde{\mu}_{cX}^{i=2} - \tilde{b}_{YX} \text{bias}(\tilde{\mu}_{cY}) \\ &\approx \tilde{\mu}_{cX}^{i=2} - \tilde{b}_{YX} \tilde{\mu}_{cY}^{i=2} + \tilde{b}_{YX} (\tilde{b}_{XY} (\tilde{\mu}_{cY} - \tilde{\mu}_{cY}^{i=2})) \\ &\approx \tilde{\mu}_{cX}^{i=2} - \tilde{b}_{YX} \tilde{\mu}_{cX} + \tilde{b}_{YX} \tilde{b}_{XY} \tilde{\mu}_{cY} - \tilde{b}_{YX} \tilde{b}_{XY} \tilde{\mu}_{cY}^{i=2}\end{aligned}$$

Iterating such bias corrections indefinitely, one reaches the asymptotically unbiased estimator

$$\hat{\mu}_{cX} = (r^0 + r^2 + r^4 + \dots) \tilde{\mu}_{cX} + (r^1 + r^3 + r^5 + \dots) \tilde{\mu}_{cY},$$

where we let $r = b_{YX}$ for brevity's sake.

Now, the sum of the geometric series

$$r^0 + r^2 + r^4 + \dots + r^1 + r^3 + r^5 + \dots = 1/(1-r)$$

whenever $-1 < r < 1$, and in that case the sum of the truncated series is given by $r + r^2 + \dots = 1/(1-r) - 1$.

This truncated series can be broken down into disjoint sums of terms involving odd vs. even powers of r (starting with r^1) to obtain the two series $r_{\text{odd}} = r^1 + r^3 + \dots$ and $r_{\text{even}} = r^2 + r^4 + \dots$. Let the sum of the series involving odd powers be $c_{\text{odd}} (1/(1-r_{\text{odd}}) - 1)$, and the sum of the other series $c_{\text{even}} (1/(1-r_{\text{even}}) - 1)$. We want to find c_{odd} and c_{even} , such that

$$c_{\text{odd}} (1/(1-r_{\text{odd}}) - 1) + c_{\text{even}} (1/(1-r_{\text{even}}) - 1) = (1/(1-r) - 1).$$

The c -coefficients will be required for the bias adjustment of $\tilde{\mu}_{cX}$ and kindred estimators.

It is easy to see that $c_{\text{even}}/c_{\text{odd}} = 1/r$. To find the individual coefficients, note that when $r = 1/2$, $c_{\text{odd}} = 2/3$ and $c_{\text{even}} = 1/3$. Also, when $r = 1/4$, $c_{\text{odd}} = 4/15$ and $c_{\text{even}} = 1/15$. Again, when $r = 1/8$, $c_{\text{odd}} = 8/63$ and $c_{\text{even}} = 1/63$. Each of the coefficients for even terms satisfies the general formula $c_{\text{even}} = 1/(1/r^2 - 1)$ and so by induction we conclude that this is the general formula; it then follows that the other general formula is $c_{\text{odd}} = r/(1/r^2 - 1)$.

Hence, asymptotically complete bias corrections for the complement class mean estimators are obtained as

$$\hat{\mu}_{cY} = \tilde{\mu}_{cY} + (1/(1-r)-1)(r/(1/r^2-1)\tilde{\mu}_{cX} + 1/(1/r^2-1)\mu_{cY}), \text{ and}$$

$$\hat{\mu}_{cX} = \tilde{\mu}_{cX} + (1/(1-r)-1)(r/(1/r^2-1)\tilde{\mu}_{cY} + 1/(1/r^2-1)\mu_{cX})$$

with the r s defined for the first equation as $r = b_{YX}$ and for the second as $r = b_{XY}$.

The MAXSLOPE code checks whether the geometric series will converge by determining whether $-1 < r = b_{YX} < 1$ and $-1 < r = b_{XY} < 1$, as these are necessary and sufficient conditions for convergence. The data submitted to MAXSLOPE analysis may or may not have X and Y standardized (i.e., overall means of zero, overall variances of unity). If they are not, then b_{YX} or b_{XY} or both may equal or exceed one in absolute value, so that the series does not converge. Also, even when X and Y are standardized, the b_{YX} and b_{XY} estimates are made from local, not overall, slope information. Hence, even in this case, \tilde{b}_{YX} or \tilde{b}_{XY} may exceed one in absolute value, even though the overall slope cannot (because of standardization). In case an estimated b lies outside $(-1, 1)$, the MAXSLOPE code uses the less-biased (but not asymptotically unbiased) one-step estimators $\tilde{\mu}_{cX}^{i=2}$ and $\tilde{\mu}_{cY}^{i=2}$, which are well-defined even when $b \notin (-1, 1)$.

The asymptotically unbiased estimators for taxon population means $\hat{\mu}_{tX}$ and $\hat{\mu}_{tY}$ are obtained by simple algebraic manipulation of $\hat{\mu}_{cX}$ and $\hat{\mu}_{cY}$. Consider that the overall unweighted sample averages $\hat{\mu}_X$ and $\hat{\mu}_Y$ are both asymptotically and small-sample unbiased for their respective population parameters. We have the trivial mixture identities

$$\hat{\mu}_X = \tilde{Q}\hat{\mu}_{cX} + \tilde{P}\tilde{\mu}_{tX} \text{ and}$$

$$\hat{\mu}_Y = \tilde{Q}\hat{\mu}_{cY} + \tilde{P}\tilde{\mu}_{tY},$$

where P is the taxon base rate (one of the parameters estimated by MAXSLOPE), and $\tilde{Q} = 1 - \tilde{P}$. Plugging in our already obtained MAXSLOPE estimates $\hat{\mu}_X$, $\hat{\mu}_Y$, $\hat{\mu}_{cX}$, $\hat{\mu}_{cY}$, and \tilde{P} , we simply solve for $\tilde{\mu}_{tX}$ and $\tilde{\mu}_{tY}$. Because at present \tilde{P} is only proven to be a consistent estimator of P , I do not have a proof that $\tilde{\mu}_{tX}$ and $\tilde{\mu}_{tY}$ are unbiased for μ_{tX} and μ_{tY} . However, I conjecture that absent sizeable aberration in the P -estimate, the $\tilde{\mu}_{tX}$ and $\tilde{\mu}_{tY}$ should be approximately asymptotically unbiased for their corresponding parameters, as $\hat{\mu}_X$ and $\hat{\mu}_Y$ are asymptotically unbiased for theirs.