Weighted Marginal Maximum Likelihood Regression Estimation

Developed by Paul Bailey and Harold Doran*

March 27, 2020

This document describes weighted Marginal Maximum Likelihood (MML) estimation for student test data in the `Dire` package. The model treats abilities as a latent variable and estimates a linear model in the latent space. In this framework, failing to account for the measurement variance will bias the regression estimates. The `Dire` package can directly estimate these models in the latent space. Aditionally, it can generate palausible values (PVs) which allow for unbiased estimation using the same formulas as multiple imputation (Mislevy et al.).

For simplicity, focusing first on a univariate model where there is only one latent ability measured, the student test data are assumed to have been generated by an Item Response Theory (IRT) model, where student i has ability θ_i. The probability of a student getting an item correct—for a dichotomous item—increases in ability but decreases in item difficulty.[1] Put together, the likelihood of student i's responss to items (R_i) is a function of the student's latent ability in the construct (θ_i), the item parameters of each item(P)

$$\mathcal{L}(\boldsymbol{R}_i|\theta_i, \boldsymbol{P}) = \prod_{h=1}^{H} \Pr(R_{ih}|\theta_i, \boldsymbol{P}_h) \tag{1}$$

where the probability function in the product is showin in the appendix and depends on the IRT model used for the item. it is possible that a student will not have been shown an item and then R_{ij} is missing. The missing response does not change the probability. This can be thought of as $\Pr(R_{ih}|\theta_i, \boldsymbol{P}_h)$ being 1 for all values of θ_i and $\boldsymbol{P}_h$.

This latent trait is then modeled as a function of a row vector of explanatory variables $\boldsymbol{X}_i$ so that.

$$\theta_i = \boldsymbol{X}_i\boldsymbol{\beta} + \epsilon_i \tag{2}$$

Where

$$\epsilon_i \sim N(0, \sigma) \tag{3}$$

This creates two sets of likelihoods for person i, one associated with the covariates ($\boldsymbol{X}_i$) and another associated with the observed item responses ($\boldsymbol{R}_i$). The likelihood for the student is the product of these two.

$$\mathcal{L}(\boldsymbol{R}_i|\boldsymbol{\beta}, \sigma, \boldsymbol{X}_i, \boldsymbol{P}) = \int f(\theta_i) \cdot \prod_{h=1}^{H} \Pr(R_{ih}|\theta_i, \boldsymbol{P}_h) d\theta_i \tag{4}$$

where $f(\theta_i)$ is the density of student is latent ability θ_i. Substituting the formula for above,

$$\mathcal{L}(\boldsymbol{R}_i|\boldsymbol{\beta}, \sigma, \boldsymbol{X}_i, \boldsymbol{P}) = \int \frac{1}{\sigma\sqrt{2\pi}} \exp(\frac{-\epsilon_i^2}{2\sigma}) \prod_{h=1}^{H} \Pr(R_{ih}|\boldsymbol{X}_i\beta + \epsilon_i, \boldsymbol{P}_h) d\epsilon_i \tag{5}$$

*This publication was prepared for NCES under Contract No. ED-IES-12-D-0002 with the American Institutes for Research. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

[1] For a polytimious item, the story is more complex, see the appendix for details of the likelihood function.

The term in the integral is the convolusion of two functions and there is no closed form solution. `Dire` follows the AM statistical software (Cohen & Jiang 1999) and uses fixed quadrature points set by the user.

In addition, Dire allows for the construct being scored to not be a simple univariate construct but to incorporate several correlated subscales or constructs. If these are labeled from one to J then each student has a vector $\boldsymbol{\theta}_i$ with components θ_{ij}. This leads to J latent regression equations and requires a J-demensional integral.

$$\mathcal{L}(\boldsymbol{R}_i|\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \boldsymbol{\Sigma}, \boldsymbol{X}_i, \boldsymbol{P}) = \int ... \int \frac{1}{(2\pi)^{\frac{J}{2}}\sqrt{\det(\boldsymbol{\Sigma})}} \exp(-\frac{1}{2}\boldsymbol{\epsilon}_i^T \boldsymbol{\Sigma} \boldsymbol{\epsilon}_i) \prod_{j=1}^{J} \prod_{h_j=1}^{H_j} \Pr(R_{ijh}|\theta_{ij}, \boldsymbol{P}_{hj}) d\theta_{i1} ... d\theta_{iJ} \tag{6}$$

where $\boldsymbol{R}_{ijh}$, $\boldsymbol{P}_{hj}$, and $\boldsymbol{\beta}_j$ all have a subscript j added because there is now a set of them for each subscale j, $\boldsymbol{\epsilon}_i$ is now a vector with elements ϵ_{ij}, and $\boldsymbol{\Sigma}$ is the covariance matrix for the $\boldsymbol{\epsilon}$ vector, which allows for covariance between constructs at the student level in that when $\Sigma_{jj'}$ is positive than a student with a higher score on construct j, conditional on $\boldsymbol{X}_i$ will also be expected to have a higher score on construct j'.

This problem suffers from the curse of dimensionality. However, as is pointed out by Cohen and Jiang (1999) this is identical to seemingly unrelated regression (SUR) and can be solved by fitting β_j and σ_j once for each subscale and then identifying each of the $\binom{J}{2}$ covariance terms in a pairwise fashion.

This comes from the fact that the multivariate normal can be decomposed into two multivariate normals, one conditional on the other. First re-writing, eq 10 with the normal rewritten as $f(\epsilon_i)$

$$\mathcal{L}(\boldsymbol{R}_i|\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \boldsymbol{\Sigma}, \boldsymbol{X}_i, \boldsymbol{P}) = \int ... \int f(\epsilon_i) \prod_{j=1}^{J} \prod_{h_j=1}^{H_j} \Pr(R_{ijh}|\theta_{ij}, \boldsymbol{P}_{hj}) d\theta_{i1} ... d\theta_{iJ} \tag{7}$$

the multivariate normal can be decomposed into the product of two distributions, an unconditional (univariate-)normal and a condional (potentially multivariate when $J > 2$) normal. Using the first integration dimension, without loss of generality, this becomes

$$\mathcal{L}(\boldsymbol{R}_i|\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \boldsymbol{\Sigma}, \boldsymbol{X}_i, \boldsymbol{P}) = \int ... \int f_1(\epsilon_{i1}) f_{-1}(\epsilon_{i,-1}|\epsilon_{i1}) \prod_{j=1}^{J} \prod_{h_j=1}^{H_j} \Pr(R_{ijh}|\theta_{ij}, \boldsymbol{P}_{hj}) d\theta_{i1} ... d\theta_{iJ} \tag{8}$$

where $f_1(\epsilon_{i1})$ is the normal density function for a single variable and $f_{-1}(\epsilon_{i,-1}|\epsilon_{i1})$ is the density function of the other variables, conditional on ϵ_{i1}. This can then be rearanged to

$$\begin{aligned}\mathcal{L}(\boldsymbol{R}_i|\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \boldsymbol{\Sigma}, \boldsymbol{X}_i, \boldsymbol{P}) = &\int f_1(\epsilon_{i1}) \left[\prod_{h_j=1}^{H_j} \Pr(R_{ijh}|\theta_{ij}, \boldsymbol{P}_{hj}) \right] d\theta_{i1} \\ &\int ... \int f_{-1}(\epsilon_{i,-1}|\epsilon_{i1}) \prod_{j=2}^{J} \prod_{h_j=1}^{H_j} \Pr(R_{ijh}|\theta_{ij}, \boldsymbol{P}_{hj}) d\theta_{i2} ... d\theta_{iJ}\end{aligned} \tag{9}$$

Changing to the numerical quadrature representation

$$\begin{aligned}\mathcal{L}(\boldsymbol{R}_i|\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \boldsymbol{\Sigma}, \boldsymbol{X}_i, \boldsymbol{P}) = &\sum_{q_1=1}^{Q} \delta f_1(q_1 - \boldsymbol{X}_i\beta_1) \left[\prod_{h_j=1}^{H_j} \Pr(R_{ijh}|q_1, \boldsymbol{P}_{hj}) \right] \\ &\sum_{q_2=1}^{Q} ... \sum_{q_J=1}^{Q} \delta^{J-1} f_{-1}(\epsilon_{i,-1}(q_{-1})|q_{i1}) \prod_{j=2}^{J} \prod_{h_j=1}^{H_j} \Pr(R_{ijh}|q_{ij}, \boldsymbol{P}_{hj})\end{aligned} \tag{10}$$

this is additively seperable—the max with respect to $\boldsymbol{\beta_1}$ can be found only with data for the first construct. The other constructs can be relabeled and each can be maximized in this way. Similarly, the variances can be independently found in this way.

The `Dire` package helps analyists estimate coefficients in two potentially useful ways. First the parameters in the regression can be directly estimated, where the β values are used from the above model fit. Second, `Dire` can generate plausible values from the fitted likelihood surface.

The direct estimation method has the advantage of using the latent space to identify coefficients. However, it can be time consuming to fit a latent parameter model.

The plausible value approach has two advantages. First, it can be used to fit a saturated model (with all possible coefficients of interest) and then other models that use subsets or linear combinations of those coefficients can be fit to the plausible values. Second, doing this saves time relative to fitting a direct estimation model per regression specification.

This document first describes how parameters are estimated in the latest space. The third section describes the variance estimator. The subsequent section describes estimation of degrees of freedom. Finally, it describes plausible value generation. Each section starts by describing the univariate case and then describes the case with J potentially correlated constructs or subscales. Additionally, while the introduction has not mentioned weights, they are incorporated in the subsequent sections.

Parameter Estimation

Starting with the single construct MML model for test data for n individuals, conditional on a set of parameters for a set of H test items, the likelihood of a regression equation is

$$\mathcal{L}(\boldsymbol{\beta}, \sigma | \boldsymbol{w}, \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{P}) = \prod_{i=1}^{N} \left[\int \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(\theta_i - \boldsymbol{X}_i\beta)^2}{2\sigma^2} \prod_{h=1}^{H} \Pr(\boldsymbol{R}_{ih} | \theta_i, \boldsymbol{P}_h) d\theta_i \right]^{\boldsymbol{w}_i} \tag{11}$$

where $\mathcal{L}$ is the likelihood[2] of the regression parameters $\boldsymbol{\beta}$ with full sample weights $\boldsymbol{w}_i$ conditional on item score matrix $\boldsymbol{R}$, student covariate matrix $\boldsymbol{X}$, and item parameter data $\boldsymbol{P}$; σ^2 is the variance of the regression residual; θ_i is the ith student's latent ability measure that is being integrated out; $\Pr(\boldsymbol{R}_{ih} | \theta_i, \boldsymbol{P}_h)$ is the probability of individual i's score on test item h, conditional on the student's ability and item parameters $\boldsymbol{P}_h$—see the appendix for example forms of $\Pr(\boldsymbol{R}_{ih} | \theta_i, \boldsymbol{P}_h)$.

The integral is evaluated using the trapezoid rule[3] at quadrature points t_q and quadrature weights δ so that

$$\mathcal{L}(\boldsymbol{\beta}, \sigma | \boldsymbol{w}, \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{P}) = \prod_{i=1}^{N} \left[\sum_{q=1}^{Q} \delta \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(t_q - \boldsymbol{X}_i\boldsymbol{\beta})^2}{2\sigma^2} \prod_{h=1}^{H} \Pr(\boldsymbol{R}_{ih} | t_q, \boldsymbol{P}_h) \right]^{\boldsymbol{w}_i} \tag{12}$$

where δ is the distance between any two uniformly spaced quadrature points so that $\delta = t_{q+1} - t_q$ for any q that is at least one and less than Q. The range and value of Q parameterize the quadrature, and its accuracy and should be varied to ensure convergence. The advantage of the trapezoidal rule is that the fixed quadrature points allow the values of the probability (the portion inside the product) to be calculated once per student.

The log-likelihood is given by

$$\ell(\boldsymbol{\beta}, \sigma | \boldsymbol{w}, \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{P}) = \sum_{i=1}^{N} \boldsymbol{w}_i \log \left[\delta \sum_{q=1}^{Q} \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(t_q - \boldsymbol{X}_i\boldsymbol{\beta})^2}{2\sigma^2} \prod_{j=1}^{K} \Pr(\boldsymbol{R}_{ij} | t_q, \boldsymbol{P}_j) \right] \tag{13}$$

Note that δ can be removed for optimization, and its presence adds $\log(\delta) \sum \boldsymbol{w}_i$ to the log-likelihood.

[2] When survey weights are applied, the likelihoods in this document are all pseudo-likelihoods.

[3] Using Big-O notation (Black, 2019), the trapezoid rule's convergence is in $O(\delta^2)$, meaning that the convergence is proportional to δ^2. If the bounds are set wide enough such that every student's likelihood is essentially zero at the edges, the convergence rate is faster than polynomial because the function is periodic and analytic (Johnson, 2010).

Composite Score Estimation

When the outcome of interest is composite scores, the parameters are estimated by separately estimating the coefficients for each subscale ($\boldsymbol{\beta}_j$ for subscale j) and then calculating the composite scores ($\boldsymbol{\beta}_c$) using subscale weights (ω_j).[4]

$$\boldsymbol{\beta}_c = \sum_{j=1}^{J} \omega_j \boldsymbol{\beta}_j \tag{14}$$

The full covariance matrix for the residuals ($\boldsymbol{\epsilon}$ vector) is then

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1J} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1J} & \sigma_{2J} & \cdots & \sigma_J^2 \end{bmatrix} \tag{15}$$

The likelihood is then

$$\ell\left(\sigma_{jj'} \,|\boldsymbol{\beta}_j, \boldsymbol{\beta}_{j'}, \sigma_j, \sigma_{j'}; \boldsymbol{w}, \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{P}\right) = \sum_{n=1}^{N} \boldsymbol{w}_n \log \left\{ \int\int \frac{1}{2\pi\sqrt{\left|\boldsymbol{\Sigma}_{(jj')}\right|}} \exp\left(\hat{\boldsymbol{\epsilon}}_{jj'}^T \boldsymbol{\Sigma}_{(jj')}^{-1} \hat{\boldsymbol{\epsilon}}_{jj'}\right) \right. \tag{16}$$

$$\left. \times \left[\prod_{h=1}^{H_j} \Pr(\boldsymbol{R}_{njh}|\theta_j, \boldsymbol{P}_{h'})\right] \left[\prod_{h'=1}^{H_{j'}} \Pr(\boldsymbol{R}_{nj'h'}|\theta_{j'}, \boldsymbol{P}_{h'})\right] \right\} d\theta_j d\theta_{j'} \tag{17}$$

where $|\boldsymbol{\Sigma}_{(jj')}|$ is the determinant of $\boldsymbol{\Sigma}_{(jj')}$,

$$\tilde{\boldsymbol{\Sigma}}_{(jj')} \equiv \begin{bmatrix} \sigma_j^2 & \sigma_{jj'} \\ \sigma_{jj'} & \sigma_{j'}^2 \end{bmatrix} \tag{18}$$

and the residual term is defined as

$$\hat{\boldsymbol{\epsilon}}_{jj'} \equiv \begin{pmatrix} \theta_j - \boldsymbol{X}_i\boldsymbol{\beta}_j \\ \theta_{j'} - \boldsymbol{X}_i\boldsymbol{\beta}_{j'} \end{pmatrix} \tag{19}$$

Notice that the parameters $\boldsymbol{\beta}_j$, $\boldsymbol{\beta}_{j'}$, σ_j^2, and $\sigma_{j'}^2$ are taken from the subscale estimation, so the only parameter not fixed is the covariance term σ_{ij}.

Variance Estimation

Starting with the univariate case, estimating variance of the parameters $\boldsymbol{\beta}$ can be done in one of several ways.

The inverse Hessian matrix is a consistent estimator when the estimator of $\boldsymbol{\beta}$ is consistent[5] (Green, 2003, p. 520):

$$\text{Var}(\boldsymbol{\beta}) = -\boldsymbol{H}(\boldsymbol{\beta})^{-1} = -\left[\frac{\partial^2 \ell(\boldsymbol{\beta}, \sigma|\boldsymbol{w}, \boldsymbol{R}, \boldsymbol{X})}{\partial \boldsymbol{\beta}^2}\right]^{-1} \tag{20}$$

This variance is returned when the variance method is set to `consistent` or left as the default.

[4] We use the term *composite score* to mean those scores that are weighted sums of subscale scores, as in Eq. 14. Overall scores that use a unidimensional model are calculated according to the methods already described by simply pooling items into a single construct.

[5] Strictly speaking, σ^2 also is a parameter, but we are rarely interested in the variance of the variance. Nevertheless, the package generates an estimate of σ^2 along with the coefficients themselves. For notational simplicity, all formulas ignore this.

A class of variance estimators typically called "sandwich" or "robust" variance estimators allow for variation in the residual and are of the form

$$\text{Var}(\boldsymbol{\beta}) = H(\boldsymbol{\beta})^{-1}\boldsymbol{V}H(\boldsymbol{\beta})^{-1} \tag{21}$$

where V is an estimate of the variance of the summed score function (Binder, 1983).

For a convenience sample, we provide two robust estimators. First, the so-called `robust` (Huber or Huber-White) variance estimator uses

$$\boldsymbol{V} = \sum_{i=1}^{N}\left[\frac{\partial\ell(\beta,\sigma|\boldsymbol{w}_i,\boldsymbol{R}_i,\boldsymbol{X}_i)}{\partial\beta}\right]\left[\frac{\partial\ell(\beta,\sigma|\boldsymbol{w}_i,\boldsymbol{R}_i,\boldsymbol{X}_i)}{\partial\beta}\right]' \tag{22}$$

Second, for the `cluster robust` case, the partial derivatives are summed within the cluster so that

$$\boldsymbol{V} = \sum_{c=1}^{n'}\left[\frac{\partial\ell(\beta,\sigma|\boldsymbol{w}_c,\boldsymbol{R}_c,\boldsymbol{X}_c)}{\partial\beta}\right]\left[\frac{\partial\ell(\beta,\sigma|\boldsymbol{w}_c,\boldsymbol{R}_c,\boldsymbol{X}_c)}{\partial\beta}\right]' \tag{23}$$

where there are n' clusters, indexed by c, and the partial derivatives are summed within the group of which there are n_c members:

$$\frac{\partial\ell(\beta,\sigma|\boldsymbol{w}_c,\boldsymbol{R}_c,\boldsymbol{X}_c)}{\partial\beta} = \sum_{i=1}^{n_c}\frac{\partial\ell(\beta,\sigma|\boldsymbol{w}_i,\boldsymbol{R}_i,\boldsymbol{X}_i)}{\partial\beta} \tag{24}$$

Finally, `Dire` impelments the survey sampling method called the `Taylor series` method and uses the same formula as Eq. 21, but $\boldsymbol{V}$ is the estimate of the variance of the score vector (Binder, 1983). Our implementation assumes a two-stage design with n_a primary sampling units (PSUs) in stratum a and summed across the A strata according to

$$\boldsymbol{V} = \sum_{a=1}^{A}\boldsymbol{V}_a \tag{25}$$

where $\boldsymbol{V}_a$ is a variance estimate for stratum a and is defined by

$$\boldsymbol{V}_a = \frac{n_a}{n_a-1}\sum_{p=1}^{n_a}\left(\boldsymbol{s}_p-\bar{\boldsymbol{s}}_a\right)\left(\boldsymbol{s}_p-\bar{\boldsymbol{s}}_a\right)' \tag{26}$$

where s_p is the sum of the weighted (or pseudo-) score vector that includes all units in PSU p in stratum a and $\bar{\boldsymbol{s}}_a$ is the (unweighted) mean of the $\boldsymbol{s}_p$ terms in stratum a so that

$$s_p = \sum_{i\in\text{PSU }p}\frac{\partial\ell(\beta,\sigma|\boldsymbol{w}_i,\boldsymbol{R}_i,\boldsymbol{X}_i)}{\partial\beta} \qquad \bar{\boldsymbol{s}}_a = \frac{1}{n_a}\sum_{p\in\text{stratum }a}s_p \tag{27}$$

When a stratum has only one PSU, $\boldsymbol{V}_a$ is undefined. The best approach is for the analyst to adjust the strata and PSU identifiers, in a manner consistent with the sampling approach, to avoid singleton strata. Two simpler, automated, but less defensible options are available in `Dire`. First, the strata with single PSUs can be dropped from the variance estimation, yielding an underestimate of the variance.

The second option is for the singleton stratum to use the overall mean of s_p in place of $\bar{s}_a$. So,

$$\bar{\boldsymbol{s}} = \frac{1}{n'}\sum s_p \tag{28}$$

where the sum is across all PSUs, and n' is the number of PSUs across all strata. Then, for each singleton stratum, Eq. 26 becomes

$$\boldsymbol{V}_a = 2\left(\boldsymbol{s}_p-\bar{\boldsymbol{s}}\right)\left(\boldsymbol{s}_p-\bar{\boldsymbol{s}}\right)' \tag{29}$$

where the value 2 is used in place of $\frac{n_a}{n_a-1}$, which is undefined when $n_a = 1$. This option can underestimate the variance but is thought to more likely overestimate it.

While not advisable, it is possible to assume information equality where the hessian (eq. 20) and the score vector based covariance (eq. 22) are equal. When a user sets `gradientHessian=TRUE` the `Dire` package uses this short hand in calculating any of the above results. Using this option is necessary to get agreement with the AM software on variance terms.

Univariate degrees of freedom

For the Taylor series estimator, the degrees of freedom estimator also uses the Welch-Satterthwaite (WS) degrees of freedom estimate (dof_{WS}). The WS weights require an estimate of the degrees of variance per independent group. For a clustered sample, that is available at the stratum.

Following Binder (1983) and Cohen (2002), the contribution c_s to the degrees of freedom from stratum s is defined as $\boldsymbol{z}_{uj}$ from the section, "Estimation of Standard Errors of Weighted Means When Plausible Values Are Not Present, Using the Taylor Series Method,"

$$c_s = w_s \frac{n_s}{n_s - 1} \sum_{u=1}^{n_s} \boldsymbol{z}_{uj} \boldsymbol{z}_{uj}^T \tag{30}$$

where u indexes the PSUs in the stratum, of which there are n_s, and w_s is the stratum weight, or the sum, in that stratum, of all the unit's full sample weights. Using the c_s values, the degrees of freedom is

$$\mathrm{dof}_{\mathrm{WS}} = \frac{\left(\sum c_s\right)^2}{\sum c_s^2} \tag{31}$$

which uses the formula of Satterthwaite (1946), assuming one degree of freedom per stratum.

Composite Score Variances

In the case of composite scores the variance becomes more complex and only one method is supported, Taylor series.

The log-likelihood of composite scores is additively separable, the covariances (including the variances) can be calculated in two steps using Eq. **??**. First, the covariance matrix of $\boldsymbol{\xi}$ is formed, and then the composite covariance terms are estimated as the variance of a linear combination of the elements of $\boldsymbol{\xi}$.

In the first step, any of the methods in the section "Variance Estimation" are applied to Eq. **??**, treating $\boldsymbol{\xi}$ in the same fashion Eq. **??** treats $\boldsymbol{\beta}$. This step results in a block diagonal inverse Hessian matrix, with a block for each subscale, and a potentially dense matrix for $\boldsymbol{V}$. Each matrix is square and has $S \cdot (\zeta + 1)$ rows and columns, where ζ is the number of elements in the regression formula (each subscale), to which one is added for the σ terms.

This step results in the following matrix:

$$\mathrm{Var}(\boldsymbol{\xi}) = H(\boldsymbol{\xi})^{-1} \boldsymbol{V} H(\boldsymbol{\xi})^{-1} \tag{32}$$

For the second step, the composite coefficient then has an ith variance term of

$$\mathrm{Var}(\boldsymbol{\xi}_{ci}) = \boldsymbol{e}_i H(\boldsymbol{\xi})^{-1} \boldsymbol{V} H(\boldsymbol{\xi})^{-1} \boldsymbol{e}_i \tag{33}$$

where $\boldsymbol{\xi}_{ci}$ is the composite coefficient for the ith coefficient, and $\boldsymbol{e}_i$ is the vector of weights arranged such that

$$\boldsymbol{\xi}_{ci} = \boldsymbol{e}_i^T \boldsymbol{\xi} \tag{34}$$

The covariance between two terms, i and j, is a simple extension

$$\mathrm{Cov}(\boldsymbol{\beta}_{ci}, \boldsymbol{\beta}_{cj}) = \boldsymbol{e}_i H(\boldsymbol{\beta})^{-1} \boldsymbol{V} H(\boldsymbol{\beta})^{-1} \boldsymbol{e}_j \tag{35}$$

which uses the definition,

$$\boldsymbol{\xi}_{cj} = \boldsymbol{e}_j^T \boldsymbol{\xi} \tag{36}$$

A simple example may help clarify. Imagine a composite score composed of two subscales, 1 and 2, with weights $\omega_1 = 0.4$ and $\omega_2 = 0.6$. Supposed a user is interested in a regression of the form

$$\theta = a + x_1 \cdot b + \epsilon \tag{37}$$

$$\epsilon \sim N(0, \sigma) \tag{38}$$

Then the regression in Eq. 37 would be fit once for subscale 1 and once for subscale 2; the first fit would yield estimated values $\{a_1, b_1, \sigma_1\}$, and the second fit would yield $\{a_2, b_2, \sigma_2\}$. The estimated value, for example, a_c, would be $a_c = 0.4 \cdot \alpha_1 + 0.6 \cdot \alpha_2$. By stacking the estimates together,

$$\boldsymbol{\theta} = \begin{bmatrix} a_1 \\ b_1 \\ \sigma_1 \\ a_2 \\ b_2 \\ \sigma_2 \end{bmatrix} \tag{39}$$

the covariance matrix can then be estimated and will result in a matrix $\boldsymbol{\Omega} \equiv \text{Var}(\boldsymbol{\beta})$ from Eq. 20 that has six rows and six columns. Using the vector

$$\boldsymbol{e}_1 = \begin{bmatrix} 0.4 \\ 0 \\ 0 \\ 0.6 \\ 0 \\ 0 \end{bmatrix} \tag{40}$$

it can easily be confirmed that $a_c = \boldsymbol{e}_1^T \boldsymbol{\xi}$, so $\text{Var}(a_c) = \boldsymbol{e}_1^T \boldsymbol{\Omega} \boldsymbol{e}_1$.

Composite degrees of freedom

References

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, *51*(3), 279–292.

Black, P. E. (2019). Big-O notation. In P. E. Black (Ed.), *Dictionary of algorithms and data structures*. Washington, DC: National Institute of Standards and Technology. Retrieved from https://www.nist.gov/dads/HTML/bigOnotation.html

Cohen, J. D., & Jiang, T. (1999). Comparison of partially measured latent traits across nominal subgroups. *Journal of the American Statistical Association*, 94(*448*), 1035–1044.

Green, W. H. (2003). *Econometric analysis* Upper Saddle River, NJ: Prentice Hall.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium of Mathematical Statistics and Probability*, Vol. I: *Statistics* (pp. 221–233). Berkeley, CA: University of California Press.

Johnson, S. G. (2010). *Notes on the convergence of trapezoidal-rule quadrature*. Retrieved from https://math.mit.edu/~stevenj/trapezoidal.pdf

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models*. (2nd ed.). London, UK: Chapman & Hall/CRC.

NAEP. (2008). The generalized partial credit model [NAEP Technical Documentation Website]. Retrieved from https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_models_gen.aspx.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(*4*), 817–838.

Appendix. Test Probability Density Functions

For all cases scored as either correct or incorrect, we use the *three parameter logit* (3PL) model:

$$\Pr(\boldsymbol{R}_{ij}|\theta_i, \boldsymbol{P}_j) = g_j + \frac{1 - g_j}{1 + \exp\left[-D\, a_j\, (\theta_i - d_j)\right]} \tag{41}$$

where g_j is the guessing parameter, a_j is the discrimination factor, d_j is the item difficulty, and D is a constant, usually set to 1.7, to map the θ_i and d_j terms to a probit-like space; this term is applied by tradition.

When a *two parameter logit* (2PL) is used, Eq. 41 is modified to omit g_j (effectively setting it to zero):

$$\Pr(\boldsymbol{R}_{ij}|\theta_i, \boldsymbol{P}_j) = \frac{1}{1 + \exp\left[-D\, a_j\, (\theta_i - d_j)\right]} \tag{42}$$

When a *Rasch model* is used, Eq. 42 is further modified to set all a_j to a single a, and D is set to one.

$$\Pr(\boldsymbol{R}_{ij}|\theta_i, \boldsymbol{P}_j) = \frac{1}{1 + \exp\left[-a\, (\theta_i - d_j)\right]} \tag{43}$$

The *Graded Response Model* (GRM) has a probability density that generalizes an ordered logit (McCullagh & Nelder, 1989):

$$\Pr(\boldsymbol{R}_{ij}|\theta_i, \boldsymbol{P}_j) = \frac{1}{1 + \exp\left[-D\, a_j\, (\theta_i - d_{R_{ij},j})\right]} - \frac{1}{1 + \exp\left[-D\, a_j\, (\theta_i - d_{1+R_{ij},j})\right]} \tag{44}$$

Here the parameters $\boldsymbol{P}_j$ are the cut points d_{cj}, where $d_{0j} = -\infty$ and $d_{C+1,j} = \infty$. In the first term on the right side of Eq. 44, the subscript R_{ij} on $d_{R_{ij},j}$ indicates it is the cut point associated with the response level to item j for person i, whereas the last subscript (j) indicates that it is the d term for item j. In the second term, the cut point above that cut point is used.

The *Generalized Partial Credit Model* (GPCM) has a probability density that generalizes a multinomial logit (McCullagh & Nelder, 1989)

$$\Pr(\boldsymbol{R}_{ij}|\theta_i, \boldsymbol{P}_j) = \frac{\exp\left[\sum_{c=0}^{R_{ij}} Da_j(\theta_i - d_{cj})\right]}{\sum_{r=0}^{C} \exp\left[\sum_{c=0}^{r} Da_j(\theta_i - d_{cj})\right]} \tag{45}$$

where c indexes cut points, of which there are C, and j indexes the item.

The GPCM equation has an indeterminancy because all d_j terms could increase and make the values of the probability the same. We can solve the indeterminacy in several ways.

NAEP (2008) uses a mean difficulty (b_j), and the d_j values are then given by

$$d_{0j} = 0 \qquad\qquad d_{cj} = b_j - \delta_{jc}\,;\, 1 \le c \le C \tag{46}$$

where the δ_{jc} values are estimated so that $0 = \sum_{c=1}^{C} \delta_{jc}$. In this package, when the `polyParamTab` has an `itemLocation`, it serves as `b`. When there is no `itemLocation`, the package uses the δ values directly

$$d_{0j} = 0 \qquad\qquad d_{cj} = \delta_{jc}\,;\, 1 \le c \le C \tag{47}$$

When a *Partial Credit Model* (PCM) is used, and the value of D is set to one, whereas a_j is again shared across all items. So

$$\Pr(\boldsymbol{R}_{ij}|\theta_i, \boldsymbol{P}_j) = \frac{\exp\left[\sum_{c=0}^{R_{ij}} a(\theta_i - d_{cj})\right]}{\sum_{r=0}^{C} \exp\left[\sum_{c=0}^{r} a(\theta_i - d_{cj})\right]} \tag{48}$$