4
Modelos de Equações Estruturais

4.1
Introdução

Este capítulo é dedicado aos fundamentos teóricos sobre os Modelos de Equações Estruturais baseados em Estruturas de Covariâncias (CSM) e em Mínimos Quadrados Parciais (PLS), e também sobre os seus respectivos métodos de estimação de escores de variáveis latentes. A ênfase será dada aos problemas estatísticos da Identificação, Estimação e Validação do modelo. No caso dos modelos CSM, existem diferentes tipos de estruturas de covariâncias que são consideradas casos especiais do Modelo Geral. Além dos Modelos de Equações Estruturais, estão incluídos nesta relação os Modelos de Testes Congenéricos, Modelos de Análise Fatorial (exploratórios e confirmatórios), Modelos de Estimação de Componentes de Variâncias e Covariâncias e Modelos de Regressão com erros de medidas (*cf*. Jöreskog, 1978).

Os Modelos de Equações Estruturais ainda são objetos de intensa pesquisa. Uma de suas características básicas é que se pode testar uma teoria de natureza causal entre um conjunto de variáveis. No caso da Satisfação do Consumidor, a teoria estabelece que o Desempenho influencia a Desconfirmação e que esta pode levar à Satisfação. Esta técnica oferece ao pesquisador a possibilidade de investigar o poder de explicação das variáveis preditoras em relação à variável dependente e ainda avaliar a importância dessas variáveis.

O capítulo está organizado da seguinte maneira. A seção 4.2 revisa os modelos baseados em estruturas de covariâncias (CSM), em particular, o Modelo Geral e os Modelos de Equações Estruturais. A seção 4.3 apresenta os métodos de estimação de escores de variáveis latentes para os modelos CSM. O capítulo termina na seção 4.4 com uma revisão dos Modelos de Equações Estruturais baseados no PLS, inclusive dos seus métodos de estimação de escores de variáveis latentes.

4.2
Análise Estrutural de Matrizes de Covariância e de Correlação

De acordo com Jöreskog (1978), a pesquisa por estruturas em variáveis psicológicas correlacionadas é um dos grandes objetivos na psicometria. Tradicionalmente esta pesquisa era realizada utilizando a Análise Exploratória de Fatores para detectar e avaliar fontes latentes de variações e covariações nas medidas observadas. No entanto, verificou-se que este tipo de análise tem maior utilidade nos primeiros estágios da experimentação ou do desenvolvimento dos testes, quando se tem pouco conhecimento acerca da natureza da medida psicológica. Freqüentemente existem estruturas nos dados que podem ser postuladas previamente e estas estruturas podem não ser consistentes com o Modelo de Análise Fatorial. Tais estruturas podem surgir, por exemplo, por causa de uma teoria baseada em hipóteses especificadas, ou por condições experimentais conhecidas, ou por resultados provenientes de estudos anteriores com inúmeros dados. Existem casos onde as variáveis observadas são ordenadas através do tempo, como nos estudos com dados longitudinais, ou definidas de acordo com um determinado esquema causal, como nos Modelos de Equações Estruturais, ou classificadas como variáveis dependentes e independentes, como nos estudos de predição. O fato é que, nestes casos, a Análise Exploratória de Fatores pode conduzir a conclusões enganosas, havendo necessidade de se aplicar outras técnicas, como no caso da estimação da Satisfação do Consumidor, onde a Satisfação foi definida como uma variável latente de caráter multidimensional.

4.2.1
O Modelo Geral

Qualquer estrutura de covariância pode ser definida especificando as variâncias e covariâncias populacionais das variáveis observadas como certas funções dos parâmetros θ_1, θ_2,..., θ_t a serem estimados a partir dos dados, isto é, $\sigma_{ij} = \sigma_{ij}(\theta)$, ou em forma matricial, $\Sigma = \Sigma(\theta)$[10]. Este modelo assume, por hipótese,

[10] Σ é a matriz de covariância e σ_{ij} é um elemento dessa matriz que corresponde à covariância entre a variável da linha i com a variável da coluna j.

que as funções $\sigma_{ij}(\boldsymbol{\theta})$ e a suas respectivas derivadas de primeira ordem são contínuas, a matriz $\boldsymbol{\Sigma}$ é positivo-definida em cada ponto $\boldsymbol{\theta}$ do espaço paramétrico admissível e a distribuição das variáveis observadas é multivariada com o seu vetor de médias $\boldsymbol{\mu}$ e a sua matriz de covariância $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ não-restringidos. Além disso, a distribuição dos dados deve ser suficientemente bem descrita pelos momentos de primeira e segunda ordem, de forma que as informações sobre $\boldsymbol{\theta}$ contidas nos momentos de ordem superior possam ser ignoradas. Na maioria dos casos, espera-se que a distribuição dos dados seja Normal Multivariada.

A estrutura de correlação é definida do mesmo modo: as correlações populacionais das variáveis observadas são funções $\rho_{ij} = \rho_{ij}(\boldsymbol{\theta})$ de $\boldsymbol{\theta}$. A estrutura de correlação é definida conforme a equação 4.1.

$$\Sigma = D_{\sigma}P(\theta)D_{\sigma} \tag{4.1}$$

Onde: $\mathbf{D_{\sigma}}$ é a matriz diagonal dos desvios padrões populacionais $\sigma_1, \sigma_2, \ldots, \sigma_p$ das variáveis observadas, considerados como parâmetros livres, e $\mathbf{P(\theta)}$ é a matriz de correlação. A estrutura de covariância 4.1 tem parâmetros $\sigma_1, \sigma_2, \ldots, \sigma_p, \theta_1, \theta_2, \ldots, \theta_t$, que serão estimados a partir dos dados, cabendo observar que a estimativa de $\sigma_i \ \forall i = 1,\ldots,p$, não será necessariamente igual ao desvio padrão correspondente na amostra.

4.2.1.1
Identificação

O problema da identificação consiste essencialmente em verificar se o vetor de parâmetros $\boldsymbol{\theta}$ será exclusivamente determinado por $\boldsymbol{\Sigma}$. Cada $\boldsymbol{\theta}$ no espaço paramétrico admissível gera uma matriz $\boldsymbol{\Sigma}$, porém dois ou mais $\boldsymbol{\theta}$´s podem gerar a mesma matriz $\boldsymbol{\Sigma}$. O modelo é dito identificado se para quaisquer dois vetores $\boldsymbol{\theta_1}$ **e** $\boldsymbol{\theta_2}$ de uma região do espaço paramétrico (localmente ou globalmente), $\boldsymbol{\theta_1} \neq \boldsymbol{\theta_2}$ implica que $\boldsymbol{\Sigma(\theta_1)} \neq \boldsymbol{\Sigma(\theta_2)}$, ou seja, a matriz $\boldsymbol{\Sigma}$ é gerada por um e somente um $\boldsymbol{\theta}$. Isto implica que todos os parâmetros são identificados. Entretanto, mesmo se o modelo como um todo não for identificado, ainda assim alguns parâmetros poderão ser. Se o modelo não for completamente identificado, restrições

apropriadas deverão ser impostas a $\boldsymbol{\theta}$ para garantir a identificação, porém a escolha dessas restrições poderá afetar a interpretação dos resultados de um modelo estimado.

Para examinar o problema da identificação de um modelo, considere a equação 4.2.

$$\sigma_{ij} = \sigma_{ij}(\theta),\ i \leq j \tag{4.2}$$

Nesta equação existem ½p(p+1) equações com *t* incógnitas. Logo, uma condição necessária para a identificação de todos os parâmetros será:

$$t \leq \tfrac{1}{2}p(p+1) \tag{4.3}$$

Se um parâmetro θ_i pode ser determinado de $\boldsymbol{\Sigma}$ através da resolução da equação 4.2 ou de um subconjunto dela, então este parâmetro é identificado, do contrário, o parâmetro é não-identificado. Freqüentemente alguns parâmetros podem ser determinados de $\boldsymbol{\Sigma}$ de várias formas, isto é, utilizando diferentes conjuntos de equações. Isto dá origem às condições de sobre-identificação em $\boldsymbol{\Sigma}$, que pode acontecer se o modelo for verdadeiro.

As equações em 4.2 freqüentemente são não-lineares e, na maioria das vezes, as soluções são difíceis e demoradas. Neste caso, soluções explícitas (analíticas) raramente existem. No entanto, existem métodos empíricos baseados na Matriz de Informação de Fisher (*cf.* Bollen, 1989) desenvolvidos para testar a identificação do modelo. Se a matriz de informação é positivo-definida então é quase certo que o modelo é identificado, mas se essa matriz é singular, então o modelo será não-identificado e o rank dessa matriz mostrará os parâmetros não-identificados.

4.2.1.2
Estimação

A população é caracterizada pelo vetor de médias $\boldsymbol{\mu}$ e pela matriz de covariância $\boldsymbol{\Sigma}$, que é função de $\boldsymbol{\theta}$, ambos considerados não-restringidos. Na prática

$\boldsymbol{\theta}$ é desconhecido devendo ser estimado a partir de uma amostra contendo N observações independentes de um vetor aleatório $\boldsymbol{x}$ de ordem p (variáveis observadas). Seja $\mathbf{S} = (s_{ij})$ a matriz de covariância amostral de ordem (p×p), baseada em n = N-1 graus de liberdade. A informação proveniente de **S** também pode ser representada pela matriz de correlação $\mathbf{R} = (r_{ij})$ e um conjunto de desvios padrões $s_1, ..., s_p$, onde $s_i = (s_{ii})^{½}$ e $r_{ij} = s_{ij} / s_i s_j$. A matriz de correlação é aplicada, normalmente, em problemas onde as origens ou as unidades das escalas de medidas das variáveis observadas são arbitrárias ou irrelevantes.

Já que não existem restrições para o vetor de médias e os momentos de ordem superior podem ser ignorados, o problema da estimação se resume ajustar a matriz $\boldsymbol{\Sigma}$ da forma $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ para a matriz de covariância observada (amostral) **S**. Os métodos clássicos normalmente empregados para se fazer este ajuste são: Mínimos Quadrados Ordinários, Mínimos Quadrados Generalizados e Máxima Verossimilhança[11], conforme descritos nas equações 4.4, 4.5 e 4.6 respectivamente.

$$F_{ULS} = ½ \text{ tr } (\mathbf{S} - \boldsymbol{\Sigma})^2 \tag{4.4}$$

$$F_{GLS} = ½ \text{ tr } (\mathbf{I} - \mathbf{S}^{-1}\boldsymbol{\Sigma})^2 \tag{4.5}$$

$$F_{ML} = \text{tr } (\mathbf{S}^{-1}\boldsymbol{\Sigma}) - \log|\, \boldsymbol{\Sigma}^{-1}\mathbf{S}| - p \tag{4.6}$$

Onde:

$\boldsymbol{\Sigma}$ é a matriz de covariância do modelo e **S** é a matriz de covariância amostral.

tr(A) é o traço da matriz A;

|A| representa o determinante da matriz A.

O ajuste perfeito se dá quando o valor de F = 0, ou seja, quando $\boldsymbol{\Sigma} = \mathbf{S}$.

As funções F_{ULS}, F_{GLS} e F_{ML} são minimizadas com respeito a $\boldsymbol{\theta}$ e, na ausência de uma solução analítica, este processo pode ser realizado

[11]Referenciados na literatura estrangeira como ULS (*unweighted least square*), GLS (*generalized least square*) e ML (*maximum likelihood*) respectivamente.

numericamente através de métodos computacionais, tal como o Método de *Scoring de Fisher* (*cf.* Andrade *et al*, 2000). Em geral, o processo de minimização é iniciado atribuindo-se valores arbitrários para os parâmetros $\boldsymbol{\theta}^{(1)}$. Em seguida são gerados sucessivos valores $\boldsymbol{\theta}^{(2)}$, $\boldsymbol{\theta}^{(3)}$, ... tal que $F(\boldsymbol{\theta}^{(k+1)}) < F(\boldsymbol{\theta}^{(k)})$. A convergência é obtida quando a diferença $|F(\boldsymbol{\theta}^{(k+1)}) - F(\boldsymbol{\theta}^{(k)})| < \varepsilon$, onde ε é um valor dado, normalmente menor do que zero.

Os métodos GLS e ML não dependem das escalas das variáveis observadas, ou seja, $F(\mathbf{S}, \boldsymbol{\Sigma}) = F(\mathbf{DSD}, \mathbf{D}\boldsymbol{\Sigma}\mathbf{D})$ para qualquer matriz diagonal **D** de fatores de escala positivos. Já o método ULS não goza desta propriedade.

Sob a hipótese do vetor de observações ***x*** apresentar distribuição normal multivariada e o tamanho da amostra ser "suficientemente grande", os métodos GLS e ML produzirão estimativas não-viesadas e consistentes para os parâmetros. No entanto, ambos os métodos requerem que a matriz de covariância **S** seja positivo-definida.

4.2.1.3
Validação

A validade do modelo pode ser testada através do teste da razão de verossimilhança. O logaritmo desta razão é simplesmente ($N/2$) vezes o valor mínimo das funções F_{ULS} ou F_{GLS} ou F_{ML}. Esta razão segue uma distribuição χ^2 com d = ½ p(p+1)-1 graus de liberdade, sujeita às condições do modelo e ao tamanho da amostra.

Após a validação do modelo, diversas hipóteses estruturais poderão ser testadas, por exemplo: (1) certos θ´s podem ser definidos para serem iguais a determinados valores; e (2) certos θ´s podem ser definidos para serem iguais em determinados grupos de modelos. Essas hipóteses conduzem para uma estrutura de covariâncias $\boldsymbol{\Sigma}(\boldsymbol{\nu})$, onde $\boldsymbol{\nu}$ é um subconjunto de $u < t$ elementos de $\boldsymbol{\theta}$. Seja F_ν o mínimo de F sob a hipótese estrutural e seja F_θ o mínimo de F sob o modelo geral. Então ($N/2$) (F_ν - F_θ) segue aproximadamente uma distribuição χ^2 com ($t - u$) graus de liberdade.

4.2.2
Modelos de Equações Estruturais

Os Modelos de Equações Estruturais se constituem numa vasta classe de modelos que incluem variáveis latentes, erros de medidas nas variáveis dependentes e independentes, múltiplos indicadores, causas recíprocas, simultaneidade e interdependência. Os métodos incluem como casos especiais: procedimentos para análise confirmatória de fatores, regressão múltipla, análise de caminhos (*path*), modelos de dados dependentes no tempo, estruturas de covariâncias, modelos recursivos e não recursivos para dados de corte e dados longitudinais. Os modelos de equações estruturais são úteis para resolver problemas em ciências sociais e do comportamento humano, sendo aplicados no marketing e nas tradicionais áreas de sociologia, psicologia, educação e econometria (*cf.* Jöreskog e Sörbom, 1982).

Este modelo normalmente é empregado quando o fenômeno sob estudo está especificado em termos de variáveis de causas e efeitos. Cada equação no modelo representa uma ligação causal ao invés de uma mera associação empírica entre as variáveis. Os parâmetros estruturais representam características do processo (mecanismo) que gera as variáveis observadas. Goldberger (1973) *apud* Jöreskog e Sörbom (1982), menciona três situações que requerem o emprego das equações estruturais ao invés dos modelos de regressão linear: (1) quando as variáveis observadas contêm erros de medidas e quando os interesses estão centrados nos relacionamentos entre as variáveis verdadeiras, (2) quando existe interdependência ou causas simultâneas entre as variáveis de respostas observadas, e (3) quando variáveis explicativas importantes não foram observadas ou omitidas.

Um Modelo de Equações Estruturais com variáveis latentes é definido conforme a equação 4.7 (*cf.* Bollen, 1989):

$$\begin{aligned} \boldsymbol{\eta} &= \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \\ \mathbf{y} &= \boldsymbol{\mu}_y + \boldsymbol{\lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon} \\ \mathbf{x} &= \boldsymbol{\mu}_x + \boldsymbol{\lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta} \end{aligned} \qquad (4.7)$$

A primeira equação é a parte estrutural do modelo e as outras duas são as partes das medidas. Os vetores aleatórios $\boldsymbol{\eta}' = (\eta_1, \ldots, \eta_m)$ e $\boldsymbol{\xi}' = (\xi_1, \ldots, \xi_n)$ não são observáveis e representam as variáveis latentes dependentes (endógenas) e independentes (exógenas) respectivamente. O vetor $\boldsymbol{\alpha}$ é o intercepto da equação estrutural, no entanto ele não aparecerá no modelo se as variáveis latentes e as observadas forem tomadas desviadas de suas respectivas médias. As matrizes $\mathbf{B}(m\times m)$ e $\boldsymbol{\Gamma}(m\times n)$ são as matrizes de coeficientes e $\boldsymbol{\zeta}' = (\zeta_1, \ldots, \zeta_m)$ é o vetor de resíduos ou distúrbios aleatórios (erros na equação estrutural). Os elementos de $\mathbf{B}$ representam os efeitos causais diretos das variáveis η em outras η e os elementos de $\boldsymbol{\Gamma}$ representam os efeitos diretos das variáveis ξ nas variáveis η. As matrizes $\boldsymbol{\Phi}(n\times n)$ e $\boldsymbol{\Psi}(m\times m)$, não representadas na equação, são as matrizes de covariância de $\boldsymbol{\xi}$ e $\boldsymbol{\zeta}$ respectivamente. As hipóteses do modelo são: $\boldsymbol{\zeta}$ é não correlacionado com $\boldsymbol{\xi}$; $\boldsymbol{\zeta_i}$ é homocedástico e não possui autocorrelações; $\mathbf{I-B}$ é não singular.

As outras duas equações contêm os vetores observáveis $\mathbf{y}' = (y_1, \ldots, y_p)$ e $\mathbf{x}' = (x_1, \ldots, x_q)$. Os vetores $\boldsymbol{\mu_y}$ e $\boldsymbol{\mu_x}$ são os interceptos da equação (médias das variáveis). Os vetores $\boldsymbol{\varepsilon}$ e $\boldsymbol{\delta}$ são os erros de medidas de $\mathbf{y}$ e $\mathbf{x}$ respectivamente. As matrizes $\boldsymbol{\lambda_y}(p\times m)$ e $\boldsymbol{\lambda_x}(q\times n)$ são as matrizes de regressão de $\mathbf{y}$ em $\boldsymbol{\eta}$ e de $\mathbf{x}$ em $\boldsymbol{\xi}$ respectivamente. É conveniente chamar $\mathbf{y}$ e $\mathbf{x}$ de variáveis observadas e $\boldsymbol{\eta}$ e $\boldsymbol{\xi}$ de variáveis latentes. As matrizes $\boldsymbol{\Theta_\varepsilon}(p\times p)$ e $\boldsymbol{\Theta_\delta}(q\times q)$, não representadas na equação, são as matrizes de covariâncias de $\boldsymbol{\varepsilon}$ e $\boldsymbol{\delta}$ respectivamente. Por hipótese os erros de medidas são não correlacionados com $\boldsymbol{\eta}$, $\boldsymbol{\xi}$ e $\boldsymbol{\zeta}$, mas podem ser correlacionados entre si.

Conforme mencionado acima, o modelo de equações estruturais engloba outros tipos de modelos. Por exemplo, se as variáveis $\mathbf{y}$ e $\boldsymbol{\eta}$ não forem especificadas, o modelo se resumirá na equação 4.8, ou seja, num modelo clássico de análise fatorial.

$$\mathbf{x} = \boldsymbol{\lambda_x}\boldsymbol{\xi} + \boldsymbol{\delta} \tag{4.8}$$

Se as variáveis $\mathbf{x}$ e $\boldsymbol{\xi}$ não forem especificadas, então o modelo se resumirá na equação 4.9, ou seja, também num modelo de análise fatorial com a vantagem

de se poder manipular as relações entre os fatores através da especificação da estrutura da matriz **B**.

$$\begin{aligned} \mathbf{y} &= \lambda_y \eta + \varepsilon \\ \eta &= \mathbf{B}\eta + \zeta \end{aligned} \tag{4.9}$$

Se a variável **x** não for especificada e a matriz $\mathbf{B} = \mathbf{0}$, então o modelo se transforma num modelo de análise fatorial de segunda ordem, conforme a equação 4.10.

$$\mathbf{y} = \lambda_y (\Gamma\xi + \zeta) + \varepsilon \tag{4.10}$$

Se as matrizes $\lambda_y = \lambda_x = \mathbf{I}$ e as matrizes $\varepsilon = \delta = \mathbf{0}$, então o modelo se transforma num modelo para sistemas independentes, conforme a equação 4.11.

$$\mathbf{y} = \mathbf{B}\mathbf{y} + \Gamma\mathbf{x} + \zeta \tag{4.11}$$

4.2.2.1 Identificação

O primeiro passo na modelagem de equações estruturais é a especificação das hipóteses multivariadas usando diagramas de caminhos e equações simultâneas. Porém, antes da estimação dos parâmetros deve-se demonstrar que o modelo está especificado de tal forma que existe uma solução identificada.

Conforme mencionado acima, um modelo estatístico é identificado quando os seus parâmetros podem ser expressos como funções independentes das informações conhecidas e que estão disponíveis. Se existir apenas uma função para cada um dos parâmetros, então a solução (valores) para os parâmetros será única. Se existir mais de uma função pelo menos para um dos parâmetros, então a solução não será única (embora uma seja escolhida como a solução ótima) e, neste caso, o modelo é considerado sobre-identificado. Normalmente isto ocorre quando existem mais informações conhecidas do que parâmetros a serem estimados. Por último, ser houver mais parâmetros a serem estimados do que informações

conhecidas, existirão infinitas soluções para os parâmetros e, neste caso, o modelo é considerado sub-identificado ou não-identificado. A identificação neste último caso só poderá ser efetuada impondo restrições aos parâmetros do modelo. A menos que o modelo esteja identificado, estimativas coerentes para os parâmetros não poderão ser obtidas, mesmo com um número grande de observações.

Do ponto de vista matemático, os modelos de equações estruturais são conjuntos de equações lineares simultâneas, onde as equações representam as hipóteses sobre como as covariâncias ou correlações entre as variáveis observadas são produzidas. Logo, para resolver um conjunto de equações simultâneas deve-se ter informações suficientes (valores conhecidos ou restrições impostas ao modelo) para estimar os valores dos parâmetros desconhecidos. Geralmente, as informações conhecidas são as características da distribuição populacional das variáveis observadas que neste caso são as variâncias e covariâncias da distribuição dos dados observados, já que as mesmas contam com estimadores amostrais consistentes.

A identificação poderá ser realizada, de modo algébrico, exprimindo os parâmetros como funções independentes dos elementos da matriz de covariância ou correlação dos dados observados. Infelizmente, este processo resulta num grande esforço, sujeito inclusive a erros, em modelos mais complexos, tornando este tipo de solução inviável. Para contornar este problema, foi criado um conjunto de regras para serem seguidas, a fim de garantir a identificação do modelo. Essas regras estão listadas abaixo (*cf*. Bollen, 1989).

4.2.2.2
Regras de Identificação

Existem regras somente para a identificação da parte estrutural do modelo e somente para a parte das medidas. Esta divisão tem por objetivo simplificar o problema da identificação. Algumas regras são condições necessárias mas não são suficientes para garantir a identificação do modelo. Outras são condições suficientes, porém não são necessárias e, finalmente, outras são condições necessárias e suficientes para a identificação. A identificação da parte estrutural dos modelos que possuem variáveis latentes e observadas é tratada da mesma forma dos modelos que possuem somente variáveis observadas. As regras para

identificação da parte das medidas são para modelos de “Complexidade 1”, isto é, cada indicador está associado a um único fator ou variável latente, e com erros de medidas não correlacionados entre si.

4.2.2.2.1
Regra-t

Esta regra é aplicada na parte estrutural do modelo e é uma condição necessária, mas não suficiente para a identificação. Ela estabelece que o modelo deve ter mais informações (variáveis) conhecidas do que parâmetros a serem estimados, isto é, o número de elementos não redundantes da matriz de covariância ou de correlações das variáveis observadas deve ser maior ou igual ao número de parâmetros livres em $\boldsymbol{\theta}$ a serem estimados. Se esta condição for satisfeita então o modelo poderá ser (mas não necessariamente) identificado. Do contrário, será não-identificado. Este é um teste fácil de ser aplicado e através dele se descobre rapidamente se o modelo é ou não-identificado.

A equação da regra encontra-se na equação abaixo:

$$t \leq \tfrac{1}{2} (p+q) (p+q+1) \qquad (4.12)$$

Onde (p+q) é o número de variáveis observadas e *t* é o número de parâmetros livres em $\boldsymbol{\theta}$ a serem estimados.

4.2.2.2.2
Regra $\mathbf{B = 0}$ (nulo)

Esta regra é aplicada na parte estrutural do modelo e é uma condição suficiente, porém não necessária para a identificação. O modelo será identificado se a matriz $\mathbf{B} = \mathbf{0}$, ou seja, se não existir relações de causas e efeitos entre as variáveis latentes endógenas no modelo. Logo, as matrizes $\boldsymbol{\Phi}$, $\boldsymbol{\Gamma}$ e $\boldsymbol{\Psi}$ poderão ser escritas como funções das matrizes de covariâncias identificadas das variáveis observadas ($\boldsymbol{\Sigma}_{xx}$, $\boldsymbol{\Sigma}_{yy}$, $\boldsymbol{\Sigma}_{xy}$). Se a matriz $\mathbf{B} = \mathbf{0}$ e a matriz $\boldsymbol{\Psi}$ diagonal, isto é, os erros

das variáveis endógenas não são correlacionados entre si, então o modelo poderá ser desagregado em modelos separados para cada variável dependente.

4.2.2.2.3
Regra Recursiva

Esta regra é similar à regra anterior. O modelo será identificado se a matriz **B** é triangular inferior, isto é, não existem causas recíprocas entre as variáveis endógenas, e a matriz **Ψ** diagonal. Logo, as matrizes **B, Φ, Γ e Ψ** poderão ser escritas como funções das matrizes de covariâncias identificadas das variáveis observadas ($\mathbf{\Sigma_{xx}}$, $\mathbf{\Sigma_{yy}}$, $\mathbf{\Sigma_{xy}}$).

4.2.2.2.4
Condições de Posto e de Ordem

Estas regras são aplicadas na parte estrutural do modelo e são condições necessárias e suficientes para a identificação, sendo, portanto, a mais geral delas. Ela pode ser aplicada em qualquer modelo, inclusive nos que falham nas regras anteriores. Estas condições lidam com o modelo equação a equação e informam onde o modelo deve ser modificado para satisfazer a identificação, sendo, portanto, de grande utilidade.

A Ordem é a condição necessária desta regra e é definida do seguinte modo: seja *p* o número de variáveis endógenas do modelo. Para cada equação de uma variável endógena Y, no mínimo (p-1) variáveis não deverão ser causas diretas da variável que está sendo avaliada. Isto é fácil de ser verificado, pois basta contar a quantidade de variáveis Y, observar individualmente cada equação de Y e verificar se ela omite (p-1) variáveis causais, ou seja, devem existir (p-1) variáveis X e Y que não causam efeitos diretos na variável Y em questão. Outra forma de verificar esta regra é através da matriz **C**, que é união (junção) das matrizes (**I**-**B**) e (-1×**Γ**). Cada linha dessa matriz deverá conter pelo menos (p-1) elementos com valor zero.

O Posto é a condição suficiente desta regra e é avaliada observando individualmente cada linha da matriz **C** anterior. Como cada linha dessa matriz

informa sobre uma das variáveis dependentes, então uma nova matriz poderá ser formada a partir de cada linha da matriz **C** original, bastando excluir todas as colunas cujo valor na linha que está sendo avaliada seja igual zero. Daí segue que, para cada matriz criada a partir das linhas da matriz **C,** se o seu posto for (p-1), então a condição de Posto estará satisfeita.

O modelo será identificado se satisfazer ambas as condições de Ordem e de Posto.

4.2.2.2.5
Regra dos Três Indicadores

Esta regra é aplicada na parte das medidas e é uma condição suficiente, porém não necessária para a identificação. O modelo das medidas para um fator (ou variável latente) será identificado se as seguintes condições forem satisfeitas simultaneamente: (1) se existir somente um elemento diferente de zero em cada linha da matriz Λ_x ou Λ_y (assumindo a hipótese de que cada variável observada esteja associada à um único fator); (2) se existirem três ou mais indicadores por fator com coeficientes de Carga (*Loadings*) diferentes de zero; e (3) se a matriz Θ_δ ou Θ_ε for diagonal, isto é, os erros de medidas não são correlacionados entre si. Com mais de três indicadores por fator o modelo será sobre-identificado. Este tipo de modelo de medida normalmente é oriundo de pesquisas que possuem múltiplas questões (perguntas) para medir cada variável latente.

No próximo capítulo, esta regra será utilizada para identificar a parte das medidas do modelo da Satisfação que servirá de base para a avaliação das metodologias de estimação de escores. Logo, cabe apresentar esta regra de forma explícita através de um exemplo de um modelo de medidas:

Seja o modelo Fatorial contendo apenas uma variável latente ξ e 3 indicadores do tipo reflexivos: $\mathbf{x_1}$, $\mathbf{x_2}$ e $\mathbf{x_3}$ (*cf*. Bollen, 1989).

Então, a matriz de covariância das variáveis observadas **x** em função dos parâmetros θ, representada como $\Sigma(\theta)$, pode ser obtida do seguinte modo:

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathrm{E}(\mathbf{x}\mathbf{x}') = \mathrm{E}(\left[\left(\boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta}\right)\left(\boldsymbol{\xi}'\boldsymbol{\Lambda}_x' + \boldsymbol{\delta}'\right)\right] = \boldsymbol{\Lambda}_x\boldsymbol{\Phi}\boldsymbol{\Lambda}_x' + \boldsymbol{\Theta}_\delta$$

Cujas matrizes estão definidas abaixo:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \qquad \mathbf{\Lambda_x} = \begin{bmatrix} \lambda_{11} \\ \lambda_{21} \\ \lambda_{31} \end{bmatrix}, \quad \boldsymbol{\delta} = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix}, \qquad \boldsymbol{\xi} = [\xi_1],$$

$$\mathbf{\Phi} = [\Phi_{11}]$$

$$\mathbf{\Theta}_\delta = \begin{bmatrix} var(\delta_1) & & \\ 0 & var(\delta_2) & \\ 0 & 0 & var(\delta_3) \end{bmatrix}$$

Onde:

x é o vetor de dados observados, $\mathbf{\Lambda_x}$ = é o vetor de cargas, $\boldsymbol{\xi}$ é o vetor de variáveis latentes, $\mathbf{\Phi}$ é a matriz de covariância de $\boldsymbol{\xi}$, $\boldsymbol{\delta}$ é a matriz de erros de medidas e $\mathbf{\Theta_\delta}$ é a matriz de covariância dos erros de medidas.

Este modelo de medidas também pode ser representado através das seguintes equações:

$$\mathbf{x_1} = \boldsymbol{\lambda_{11}\xi_1} + \boldsymbol{\delta_1}$$
$$\mathbf{x_2} = \boldsymbol{\lambda_{21}\xi_1} + \boldsymbol{\delta_2}$$
$$\mathbf{x_3} = \boldsymbol{\lambda_{31}\xi_1} + \boldsymbol{\delta_3}$$
$$E(\boldsymbol{\delta_i}) = \mathbf{0}, \quad \text{cov}(\boldsymbol{\xi_1}, \boldsymbol{\delta_i}) = \mathbf{0}, \quad \text{cov}(\boldsymbol{\delta_j}, \boldsymbol{\delta_i}) = \mathbf{0}, \text{ para } i \neq j$$

Onde:

$\mathbf{\Sigma} = \mathbf{\Sigma}(\boldsymbol{\theta})$ é determinado do seguinte modo:

$$\mathbf{\Sigma} = \begin{bmatrix} var(x_1) & & \\ cov(x_2, x_1) & var(x_2) & \\ cov(x_3, x_1) & cov(x_3, x_2) & var(x_3) \end{bmatrix}$$

$$\mathbf{\Sigma}(\boldsymbol{\theta}) = \begin{bmatrix} \lambda_{11}^2\phi_{11} + var(\delta_1) & & \\ \lambda_{21}\lambda_{11}\phi_{11} & \lambda_{21}^2\phi_{11} + var(\delta_2) & \\ \lambda_{31}\lambda_{11}\phi_{11} & \lambda_{31}\lambda_{21}\phi_{11} & \lambda_{31}^2\phi_{11} + var(\delta_3) \end{bmatrix}$$

Na igualdade $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$, percebe-se claramente que existem mais parâmetros a serem determinados na matriz $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ (7 parâmetros) do que dados fornecidos pela matriz $\boldsymbol{\Sigma}$ (6 informações). Logo será necessário impor restrições aos parâmetros da matriz $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ para garantir a identificação do modelo. Uma restrição normalmente utilizada é: $\lambda_{11} = 1$. Isto equivale a igualar a escala da variável latente ξ com a escala de um dos seus indicadores, neste caso, com a de $\mathbf{x_1}$. Outra forma de garantir a identificação é assumir que a variância da variável latente ξ seja igual a 1, ou seja, $\phi_{11} = 1$. Neste caso as variáveis observadas deverão estar padronizadas.

4.2.2.2.6
Regra dos Dois Indicadores

Esta regra é similar à regra anterior. O modelo das medidas será identificado se as seguintes condições forem satisfeitas simultaneamente: (1) se existir mais de uma variável latente; (2) se cada variável latente é correlacionada no mínimo com outra latente; (3) se existir somente um elemento diferente de zero em cada linha da matriz $\boldsymbol{\Lambda_x}$ ou $\boldsymbol{\Lambda_y}$; (4) se existirem dois ou mais indicadores por fator; e (3) se a matriz $\boldsymbol{\Theta_\delta}$ ou $\boldsymbol{\Theta_\varepsilon}$ for diagonal. Este tipo de modelo de medida normalmente é oriundo da Análise Confirmatória de Fatores ou de estudos onde existe uma quantidade de variáveis exógenas correlacionadas.

Estas duas regras não se aplicam caso $\boldsymbol{\Theta_x}$ ou $\boldsymbol{\Theta_y}$ não é diagonal. Esta situação freqüentemente ocorre quando os mesmos indicadores são utilizados em estudos de Painel.

4.2.2.3
Estimação

Na estimação dos parâmetros, assume-se que a distribuição das variáveis observadas pode ser descrita através do seu vetor de médias e da sua matriz de covariância, ignorando-se os momentos de ordem superior. Como não há restrições para o vetor de médias, a estimação dos parâmetros pode ser realizada

através do ajuste da matriz de covariância imposta pelo modelo (**Σ**), para a matriz de covariância amostral (**S**).

Assumindo que a distribuição dos dados seja Normal Multivariada, estimadores de Máxima Verossimilhança dos parâmetros poderão ser obtidos. Outros métodos de estimação dos parâmetros tais como: Mínimos Quadrados Generalizados e Mínimos Quadrados não Ponderados, não fazem uso da hipótese de normalidade dos dados (*cf.* Bollen, 1989; Jöreskog e Sörbom, 1982). A estimação, definida de acordo com a verossimilhança, é efetuada minimizando a função F abaixo: (*cf.* Bollen, 1989).

$$F_{ML} = \log|\mathbf{\Sigma}| + tr(\mathbf{S}\mathbf{\Sigma}^{-1}) - \log|\mathbf{S}| - (p+q) \qquad (4.13)$$

Onde:

Σ é a matriz de covariância do modelo e **S** é a matriz de covariância amostral.

|**A**| representa o determinante de uma matriz **A**.

tr(**A**) é o traço da matriz **A**.

O ajuste perfeito se dá quando o valor de F = 0, ou seja, quando **Σ** = **S**.

O método dos Mínimos Quadrados Generalizados é definido conforme a equação 4.14.

$$F_{GLS} = \tfrac{1}{2}\, tr\,(\{[\mathbf{S} - \mathbf{\Sigma}(\boldsymbol{\theta})]\mathbf{W}^{-1}\}^2\} \qquad (4.14)$$

Onde $\mathbf{W}^{-1}$ é a matriz de pesos para a matriz residual. O método dos Mínimos Quadrados não Ponderados é um caso especial deste método, onde $\mathbf{W}^{-1} = \mathbf{I}$.

As equações 4.13 e 4.14 acima são consideradas funções independentes dos parâmetros $\boldsymbol{\theta}$, isto é, dos parâmetros livres e limitados em $\mathbf{\Lambda_x}$, $\mathbf{\Lambda_y}$, **B**, **Γ**, **Φ**, **Ψ**, $\mathbf{\Theta_\delta}$ e $\mathbf{\Theta_\varepsilon}$, que serão minimizadas em relação a eles.

Uma exigência para o método da máxima verossimilhança é que a matriz **S** seja positivo-definida e os valores iniciais para as estimativas dos parâmetros sejam dados de tal forma que a matriz **Σ** seja também positivo-definida. Esta exigência não é necessária para os métodos GLS e ULS.

Nos métodos de estimação ULS, GLS e ML, a função de ajuste $F(\theta)$ será minimizada através de um processo iterativo, tomando-se uma estimativa inicial para os parâmetros $\theta^{(1)}$, e gerando sucessivamente novos pontos $\theta^{(2)}$, $\theta^{(3)}$ ..., dentro do espaço paramétrico permitido de tal forma que: $F(\theta^{(n+1)}) < F(\theta^{(n)})$. Este processo continua até que a convergência seja obtida, ou seja, $|F(\theta^{(n+1)}) - F(\theta^{(n)})| < \varepsilon$, onde ε é um valor positivo dado e normalmente menor do que zero.

Os métodos de otimização fazem uso das derivadas de primeira ordem da função F e das aproximações (limites em probabilidades - *plim*) das derivadas de segunda ordem de F. No caso do método ML, a aproximação das derivadas de segunda ordem é também conhecida como Matriz de Informação, que é sempre positivo-definida em modelos identificados. A Matriz de Informação também poderá ser calculada e utilizada para computar os erros padrões de todos os parâmetros do modelo. A estimativa da matriz de covariância ou correlação de todos os parâmetros estimados também poderá ser obtida.

Vários mínimos locais poderão ser encontrados durante o processo de otimização da função de ajuste. De acordo com Jöreskog & Sörbom, (1982), a única forma de evitar este problema é ter um modelo que seja apropriado para os dados além de uma amostra aleatória de tamanho grande. No entanto, experiências indicam que múltiplas soluções poderão ser encontradas, mas normalmente elas se encontrarão na fronteira ou fora do espaço paramétrico admissível.

4.2.2.4
Validação

Nesta etapa procura-se avaliar se os coeficientes e as magnitudes dos efeitos estimados estão de acordo com as hipóteses previamente levantadas sobre o modelo. Para esta finalidade, foi criada uma série de medidas com o objetivo de avaliar o ajuste do modelo aos dados sendo categorizadas em medidas de avaliação geral e medidas de avaliação individual.

A primeira e mais óbvia forma de verificar o ajuste do modelo é examinar os resultados das seguintes quantidades: (1) estimativas dos parâmetros; (2) erros padrões das estimativas (apenas para o método de máxima verossimilhança); (3) correlações múltiplas quadráticas; (4) coeficientes de determinação e (5)

estimativas das correlações dos parâmetros (também apenas para o método de máxima verossimilhança).

Valores “duvidosos” para essas quantidades indicam que o modelo está fundamentalmente errado e não está bem ajustado aos dados. Por exemplo, variâncias negativas, correlações maiores do que 1 (um) e matrizes de covariância e correlação que não são positivo-definidas. Outros indicadores de ajustes ruins são: correlações múltiplas quadráticas e coeficientes de determinação com valores negativos, erros padrões de grande magnitude ou estimativas dos parâmetros altamente correlacionadas entre si. Esses valores significam que o modelo é não-identificado e que alguns parâmetros não poderão ser determinados a partir dos dados.

Estas medidas são definidas do seguinte modo:

a) A Correlação múltipla quadrática da i-ésima variável observada é dada pela equação abaixo:

$$1 - \frac{\hat{\theta}_{ii}}{S_{ii}} \tag{4.15}$$

Onde: $\hat{\theta}_{ii}$ é a variância do erro e s_{ii} é a variância observada da i-ésima variável.

b) O coeficiente de determinação é definido conforme a equação abaixo:

$$1 - \frac{|\boldsymbol{\theta}|}{|\mathbf{S}|} \tag{4.16}$$

Onde: $|\boldsymbol{\theta}|$ é o determinante de $\boldsymbol{\theta}$ e $|\mathbf{S}|$ é o determinante da matriz de covariância das variáveis observadas.

Estas medidas mostram como as variáveis observadas servem, separadamente ou conjuntamente, como instrumentos de medidas das variáveis latentes. Estes coeficientes variam de zero a um e, quanto mais próximo de um, melhor o ajuste do modelo.

A correlação múltipla quadrática e o coeficiente de determinação da equação estrutural são definidos de acordo com as equações abaixo:

$$1 - \mathrm{Var}(\zeta_i) / \mathrm{Var}(\eta_i) \quad (4.17)$$

$$1 - |\psi| / \mathrm{Cov}\,(\eta) \quad (4.18)$$

Para uma avaliação geral do ajuste do modelo aos dados, as seguintes estatísticas poderão ser utilizadas: χ^2, GFI, AGFI e RMR (*cf*. Jöreskog & Sörbom, 1982).

O valor da estatística χ^2 pode ser obtido multiplicando-se (N-1) (tamanho da amostra menos 1) pelo valor resultante da função de ajuste do modelo estimado (F_{ML} ou F_{GLS}). Se o modelo está correto e a amostra possui tamanho "suficientemente grande", então esta medida equivale ao teste estatístico da razão da verossimilhança, normalmente empregado para testar a hipótese de que a matriz $\boldsymbol{\Sigma}$ é da forma insinuada pelo modelo contra a hipótese alternativa que considera a matriz $\boldsymbol{\Sigma}$ não restringida. A quantidade de graus de liberdade da estatística χ^2 é calculada através da equação abaixo:

$$\text{d.f.} = \tfrac{1}{2}\, k\,(k+1) - t \quad (4.19)$$

Onde: *k* é o número de variáveis observadas analisadas e *t* é o total de parâmetros independentes estimados. O valor da estatística χ^2 equivale a probabilidade de se obter um valor χ^2 maior do que o valor realmente obtido, dado que modelo está correto.

Jöreskog & Sörbom (1982), enfatizam a limitação do uso da estatística χ^2 por várias razões, dentre elas, as condições que devem ser atendidas na prática para a validade do teste: (1) todas as variáveis observadas devem ter distribuição normal multivariada; (2) a análise deve ser baseada na matriz de covariância amostral (padronizações não são permitidas); e (3) o tamanho da amostra deve ser grande.

As outras medidas mencionadas acima para a avaliação geral do ajuste do modelo são definidas conforme as equações 4.20, 4.21 e 4.22.

$$GFI_{ML} = 1 - \frac{tr(\hat{\Sigma}^{-1}\mathbf{S} - \mathbf{I})^2}{tr(\hat{\Sigma}^{-1}\mathbf{S})} \quad (4.20)$$

$$AGFI_{ML} = 1 - [k(k+1)/2df](1\text{-}GFI) \quad (4.21)$$

$$RMR = \left[2\sum_{i=1}^{q}\sum_{j=1}^{i} \frac{(s_{ij} - \hat{\sigma}_{ij})^2}{k(k+1)} \right]^{1/2} \quad (4.22)$$

Onde $\hat{\mathbf{\Sigma}}$ é a matriz ajustada, *k* é o número de variáveis observadas e *df* é a quantidade de graus de liberdade.

As medidas GFI (*goodness-of-fit index*) e AGFI (*Adjusted goodness-of-fit index*) variam de 0 a 1 e medem o total relativo da variância e covariância em **S** que é explicada por $\hat{\mathbf{\Sigma}}$. A medida AGFI é ajustada para a quantidade de graus de liberdade de um modelo relativo à quantidade de variáveis observadas. Para um dado valor de GFI e *k*, a medida AGFI simplesmente recompensa os modelos com poucos parâmetros. Ambas as medidas atingem o seu valor máximo quando **S** = $\hat{\mathbf{\Sigma}}$. Ao contrário da medida χ^2, as medidas GFI e AGFI são independentes do tamanho da amostra e relativamente robustas em relação a desvios da normalidade dos dados, embora suas distribuições estatísticas não sejam conhecidas. Ambas as medidas podem ser utilizadas tanto para comparar o ajuste de modelos diferentes para os mesmos dados quanto para comparar o ajuste de modelos para diferentes dados. Para os métodos de estimação ULS e GLS, essas equações são ligeiramente modificadas (*cf.* Bollen, 1989).

A medida RMR (*root mean square residual*) refere-se a uma média dos resíduos e pode ser interpretada somente em relação aos tamanhos das variâncias e covariâncias observadas em **S**. Se a hipótese mantida pelo modelo, ou seja, $\mathbf{\Sigma} = \mathbf{\Sigma}(\boldsymbol{\theta})$ (chamada de H_0) for verdadeira, então a matriz de covariância residual da população será igual a zero. Qualquer resíduo diferente de zero para a matriz de covariância da população significa que existe um erro na especificação do modelo, embora os resíduos sejam estimados com base na matriz de covariância amostral **S**, já que não se tem a matriz de covariância da população **Σ**. Quanto mais próximos de zero estiverem os valores dos resíduos, melhor o ajuste do

modelo e, resíduos positivos significam que o modelo está subestimando as covariâncias entre duas variáveis, enquanto valores negativos significam que o modelo está superestimando essas covariâncias. Esta medida pode ser utilizada inclusive para comparar o ajuste de diferentes modelos para os mesmos dados.

4.3
Escores de Variáveis Latentes

4.3.1
Escores de Fatores

Na Análise de Fatores, os interesses estão centrados geralmente nos parâmetros do modelo de fatores, no entanto, os valores estimados dos fatores comuns, chamados de Escores Fatoriais, também podem ser utilizados para fins de diagnósticos e também como dados de entrada em análises subseqüentes.

Escores Fatoriais não são estimativas dos parâmetros desconhecidos no sentido habitual, pelo contrário, eles são as estimativas dos valores dos Fatores aleatórios não observados (*cf*. Johnson & Wichern, 1998).

O modelo de análise fatorial é apresentado, genericamente, em forma matricial através da equação 4.23:

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\varepsilon} \tag{4.23}$$

Onde: $\mathbf{X'} = (X_1,...,X_p)$ é um vetor transposto de variáveis aleatórias observáveis, $\boldsymbol{\mu}' = (\mu_1,...,\mu_p)$ é um vetor transposto de médias da variável observada $\mathbf{X}$, $\mathbf{F'} = (F_1,...,F_r)$ é um vetor transposto ($r < p$) de variáveis não observáveis ou fatores; $\mathbf{L}$ uma matriz ($p \times r$) de coeficientes fixos ou cargas fatoriais (*loadings*) e $\boldsymbol{\varepsilon}' = (\varepsilon_1, \ldots, \varepsilon_p)$ um vetor transposto de erros aleatórios.

A Análise Fatorial tem propriedades importantes. A primeira é que $E(\boldsymbol{\varepsilon}) = E(\mathbf{F}) = \mathbf{0}$ e a segunda refere-se aos fatores, que devem ser ortogonais. Nem sempre a estrutura inicial das estimativas das cargas fatoriais é definitiva, empregando-se normalmente as cargas estimadas rotacionadas.

A estimação dos Escores Fatoriais, após a rotação ortogonal da estrutura fatorial inicial, situa cada observação no espaço dos fatores comuns. Assim, para cada fator F_i, o *i-ésimo* Escore Fatorial a ser extraído é definido por f_i expresso por:

$$f_i = \sum_{i=1}^{n} b_j X_{ij}, \text{ com } j = 1,...,p \tag{4.24}$$

Onde: b_j são os coeficientes de regressão e X_{ij} as p variáveis observáveis.

Para estimar a variável f_i, que não é observável, utiliza-se a técnica de análise fatorial por meio da matrix **X** de variáveis observáveis. A forma matricial empregada é a equação (4.25), devidamente reestruturada:

$$\mathbf{f}_{(nxq)} = \mathbf{X}_{(nxp)}\mathbf{B}_{(pxq)} \tag{4.25}$$

Os Escores Fatoriais são afetados pelas unidades em que as variáveis X_i são medidas, tornando-se conveniente trabalhar com variáveis normalizadas. Desta forma, substitui-se a variável X_i pela normalizada $Z_{ij,}$ expressando em desvios padrões os desvios das observações originais em relação à sua média, conforme a equação 4.26:

$$Z_{ij} = \left[\left(X_{ij} - \mu_{xi} \middle/ \sigma_{xi}\right)\right] \tag{4.26}$$

Onde: μ_{xi} é a média de X_i e o σ_{xi} o seu desvio padrão.

A equação 4.25 é então modificada, sendo reescrita da seguinte forma:

$$\mathbf{f}_{(nxq)} = \mathbf{Z}_{(nxp)}\boldsymbol{\beta}_{(pxq)} \tag{4.27}$$

Como as variáveis estão normalizadas em ambos os lados da equação o vetor dos coeficientes da regressão **B** é substituído pelo vetor **β**. Multiplicando-se os dois lados da equação 4.27 por (1/n)**Z'**, obtém-se a equação 4.28:

$$(1/n)\mathbf{Z}'\mathbf{f} = (1/n)\mathbf{Z}'\mathbf{Z}\boldsymbol{\beta} \tag{4.28}$$

Onde: n é o número de observações e **Z'** a matriz transposta de **Z**.

O segundo membro da equação 4.28 é a matriz de correlação entre os termos de X_i, que, a partir de agora, será representada por **R**. Já o primeiro membro da equação representa a correlação entre os escores fatoriais e os próprios fatores e será identificada por $\boldsymbol{\Lambda}$. Assim, pode-se reescrever a equação 4.28 da seguinte forma:

$$\boldsymbol{\Lambda} = \mathbf{R}\boldsymbol{\beta} \tag{4.29}$$

Supondo que a matriz **R** seja não-singular, em que $|\mathbf{R}| \neq 0$, então multiplicando ambos os lados da equação 4.29 por $\mathbf{R}^{-1}$, que é a inversa de **R**, tem-se:

$$\boldsymbol{\beta} = \mathbf{R}^{-1}\boldsymbol{\Lambda} \tag{4.30}$$

Finalmente, estimado o vetor $\boldsymbol{\beta}$, pode-se substituí-lo na equação 4.25, para obtermos os escores fatoriais de cada observação.

4.3.2
Escores de Variáveis Latentes dos Modelos CSM

Os escores das variáveis latentes podem ser obtidos somente para as variáveis exógenas, tal como na Análise Confirmatória de Fatores, ou para as variáveis exógenas e endógenas simultaneamente.

No primeiro caso, os escores são obtidos através de uma função ponderada das variáveis observadas, como uma regressão "hipotética" de $\boldsymbol{\xi}$ em **x** (método derivado da Análise Fatorial para estimar os Escores Fatoriais - conforme descrito no item anterior). Isso conduz ao seguinte resultado (*cf.* Bollen, 1989):

$$\hat{\boldsymbol{\xi}} = \hat{\boldsymbol{\Phi}}\hat{\boldsymbol{\Lambda}}_x'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{x} \tag{4.31}$$

Onde $\hat{\boldsymbol{\xi}}$ é a estimativa de $\boldsymbol{\xi}$. O "peso" que pré-multiplica ***x*** é o estimador de mínimos quadrados ordinários dos coeficientes da regressão "hipotética" de $\boldsymbol{\xi}$ em

x. As matrizes $\hat{\Phi}, \hat{\Lambda}_x^{'}$ e $\hat{\Sigma}^{-1}$ são respectivamente as estimativas da matriz de covariância das variáveis latentes exógenas, da matriz de cargas e da matriz de covariância imposta pelo modelo às variáveis observadas.

No segundo caso, os escores das variáveis latentes exógenas e endógenas podem ser obtidos através do método de Jöreskog (*cf*. Jöreskog, 2000).

Da equação 4.7, sejam as seguintes matrizes:

k é o vetor de médias de ξ;

$\mathbf{\Phi}$ e $\mathbf{\Psi}$ são as matrizes de covariância de ξ e de ζ;

$\mathbf{\Theta_\varepsilon}$ e $\mathbf{\Theta_\delta}$ são as matrizes de covariância de $\boldsymbol{\varepsilon}$ e $\boldsymbol{\delta}$;

$\mathbf{\Theta_{\varepsilon\delta}}$ é a matriz de covariância de $\boldsymbol{\varepsilon}$ e $\boldsymbol{\delta}$.

Mantendo-se as hipóteses assumidas para o modelo de equações estruturais, define-se o vetor $\xi^* = (\eta', \xi')'$, englobando todas as variáveis latentes do modelo. Para este vetor, seja **k*** o seu vetor de médias e **Φ*** a sua matriz de covariância, definidas conforme as equações 4.32 e 4.33 abaixo:

$$\mathbf{k}^* = \begin{pmatrix} \mathbf{(I-B)^{-1}(\alpha+\Gamma k)} \\ \mathbf{k} \end{pmatrix} \tag{4.32}$$

$$\mathbf{\Phi}^* = \begin{pmatrix} \mathbf{A(\Gamma\Phi\Gamma'+\Psi)A'} & \mathbf{A\Gamma\Gamma} \\ \mathbf{\Phi\Gamma' A'} & \mathbf{\Phi} \end{pmatrix} \tag{4.33}$$

Onde: $\mathbf{A} = (\mathbf{I}-\mathbf{B})^{-1}$.

O objetivo do método é obter escores individuais para as variáveis latentes η e ξ, de forma que o vetor de médias e a matriz de covariância dos escores sejam respectivamente **k*** e **Φ***. O modelo pode envolver todas as matrizes de parâmetros: $\boldsymbol{\kappa}$, $\boldsymbol{\alpha}$, τ_y, τ_x, Λ_y, Λ_x, **B**, **Γ**, **Φ**, **Ψ**, Θ_ε, Θ_δ, $\Theta_{\varepsilon\delta}$, cujos elementos podem ser de três espécies:

Parâmetros fixos, cujos valores são especificados;

Parâmetros restringidos, cujos valores são desconhecidos, porém são funções lineares e não-lineares de um ou mais parâmetros;

Parâmetros livres, cujos valores são desconhecidos e não-restringidos.

O modelo assume que todas essas matrizes são conhecidas, isto é, todos os parâmetros já foram estimados.

Na seqüência, os modelos das medidas na equação 4.7 podem ser combinados de acordo com a equação 4.34 ou, de forma análoga, com a equação 4.35. Aplicando-se as seguintes transformações: $\boldsymbol{\xi} = \boldsymbol{\xi}^* - \boldsymbol{\kappa}^*$, $\boldsymbol{x} = \boldsymbol{x}^* - \boldsymbol{\tau} - \boldsymbol{\Lambda\kappa}^*$ e $\boldsymbol{\delta} = \boldsymbol{\delta}^*$, a equação 4.35 transforma-se na equação 4.36 (fazendo uso da mesma notação). Deste modo, estimar $\boldsymbol{\xi}^*$ é equivalente a estimar $\boldsymbol{\xi}$.

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\tau}_y \\ \boldsymbol{\tau}_x \end{pmatrix} + \begin{pmatrix} \boldsymbol{\Lambda}_y & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_x \end{pmatrix} \begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{\xi} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon} \\ \boldsymbol{\delta} \end{pmatrix} \tag{4.34}$$

$$\mathbf{x}^* = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\xi}^* + \boldsymbol{\delta}^* \tag{4.35}$$

$$\mathbf{x} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta} \tag{4.36}$$

Então, dado uma amostra de observações denotada por $\boldsymbol{x} = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, onde N é o tamanho da amostra, os escores poderão ser obtidos através da minimização da equação 4.37 sujeito à restrição 4.38 (*cf*. Anderson e Rubin 1956, *apud* Jöreskog, 2000).

$$\sum_{n=1}^{N} (\mathbf{x}_a - \boldsymbol{\Lambda}\boldsymbol{\xi}_a)' \boldsymbol{\Theta}^{-1} (\mathbf{x}_a - \boldsymbol{\Lambda}\boldsymbol{\xi}_a) \tag{4.37}$$

$$\left(1/N\right) \sum_{a=1}^{n} \boldsymbol{\xi}_a \boldsymbol{\xi}_a' = \boldsymbol{\Phi} \tag{4.38}$$

Onde: a matriz $\boldsymbol{\Theta}$, definida conforme a equação 4.39, é a matriz de covariância de $\boldsymbol{\delta}$.

$$\boldsymbol{\Theta} = \begin{pmatrix} \boldsymbol{\Theta}_{\varepsilon} & \boldsymbol{\Theta}'_{\delta\varepsilon} \\ \boldsymbol{\Theta}_{\delta\varepsilon} & \boldsymbol{\Theta}_{\delta} \end{pmatrix} \tag{4.39}$$

A solução deste problema é a equação 4.40, onde *n* representa o tamanho da amostra e as matrizes **UDU'** e $\mathbf{VL^{-1/2}V'}$ são decomposições em valores singulares das matrizes $\mathbf{\Phi}$ e $\mathbf{D^{1/2}U'\Lambda'\Theta^{-1}}\sum_{i=1}^{N}\left(\mathbf{x}_i\mathbf{x}_i^{'}\right)\mathbf{\Theta^{-1}\Lambda UD^{1/2}}$ respectivamente.

$$\hat{\xi}_i = \mathbf{UD^{1/2}VL^{-1/2}V'D^{1/2}U'\Lambda'\Theta^{-1}x_i} \quad \forall i = 1...n \tag{4.40}$$

Em termos de ***x**** e ξ* a solução é a equação 4.41.

$$\hat{\xi}_i^* = \mathbf{k}^* + \mathbf{UD^{1/2}Z^{-1}D^{1/2}U'\Lambda'\Theta^{-1}(x_i^* - \mu - \Lambda k^*)} \quad \forall i = 1...n \tag{4.41}$$

Onde: $Z^{-1} = \mathbf{VL^{-1/2}V'}$

Finalmente, através da equação 4.41, pode ser verificado que:

$$E(\hat{\xi}_i^*) = \mathbf{k}^* \quad \text{e} \quad E(\hat{\xi}_i^*\hat{\xi}_i^{*'}) = \mathbf{\Phi}^* \tag{4.34}$$

Ou seja, os escores preservam as relações entre as variáveis latentes no modelo.

4.4
Modelos de Equações Estruturais Baseados no PLS

A metodologia do PLS foi desenvolvida principalmente por Herman Wold (*cf.* Wold, 1985). Jan-Bernd Lohmöller (1989) aprimorou os aspectos computacionais dessa metodologia e adicionou desenvolvimentos teóricos. Wynne W. Chin (1998) desenvolveu um novo *software* com interface gráfica além de técnicas avançadas de validação dos modelos PLS.

Este modelo é descrito por um modelo de medidas, também chamado de modelo exterior, relacionando as variáveis observadas nas variáveis latentes correspondentes, e por um modelo estrutural, também chamado de modelo interior, relacionando as variáveis latentes endógenas em outras variáveis latentes, que podem ser endógenas e exógenas.

O modelo para tratar causalidades entre variáveis latentes (parte estrutural) é descrito conforme a equação 4.35.

$$\xi_j = \beta_{j0} + \sum_{i=1,i\neq j}^{J} \beta_{ji}\xi_i + \zeta_j, \quad \forall j = 1...J \tag{4.35}$$

Onde: J é a quantidade de variáveis latentes, ξ_j e ξ_i são as variáveis latentes, β_{j0} é o termo constante, β_{ji} são os coeficientes da regressão e ζ_j é o termo residual.

A variável latente que nunca aparece como uma variável dependente é chamada de exógena, as demais de endógenas. O modelo do PLS é do tipo recursivo, isto é, não permite relacionamentos recíprocos entre variáveis latentes.

O modelo de medidas admite duas formas de relacionamentos entre as variáveis observadas e as latentes: o modo Reflexivo, onde as variáveis latentes se manifestam através das variáveis observadas (as setas no diagrama de caminhos apontam na direção das variáveis observadas), e o modo Formativo, onde as variáveis latentes são definidas como uma combinação linear exata dos seus indicadores empíricos (as setas no diagrama de caminhos apontam na direção das variáveis latentes). Neste caso, as variáveis latentes são consideradas como se fossem índices (indicadores) produzidos pelas variáveis observadas. O modo Reflexivo é uma formação típica da Teoria Clássica de Testes e da Análise Fatorial, cuja tentativa é estimar as variâncias e/ou covariâncias observadas. O modo Formativo não é designado para as variáveis observadas, pelo contrário, ele é utilizado para minimizar os resíduos nos relacionamentos estruturais. As equações 4.36 e 4.37 mostram as equações do modelo de medidas para o modo Reflexivo e Formativo respectivamente.

$$x_{jh} = \pi_{jh0} + \pi_{jh}\xi_j + \varepsilon_{jh}, \quad \forall j = 1...J \tag{4.36}$$

$$\xi_j = \sum_{h=1}^{k_j} \varpi_{jh} x_{jh} + \delta_j \tag{4.37}$$

Onde: na primeira equação, J é a quantidade de variáveis latentes, h é a quantidade de indicadores vinculados à variável latente, π_{jh0} é o termo constante, π_{jh} são os coeficientes da regressão, ε_{jh} é o termo residual. Na segunda equação,

ϖ_{jh} são os coeficientes da regressão múltipla e δj é o termo residual. Os vetores $\boldsymbol{\pi}$ e $\boldsymbol{\omega}$ são chamados respectivamente de Cargas e de Pesos.

No PLS as variáveis observadas devem ser construídas de tal forma que cada variável x_{jh} fique positivamente correlacionada com a sua variável latente ξ_j. Isto implica que os sinais das Cargas π_{jh} ou dos Pesos ϖ_{jh} serão positivos. No entanto, como não existem restrições para estes sinais nos algorítmos do PLS, então sinais não esperados para as Cargas ou para os Pesos podem indicar problemas nos dados sugerindo ações corretivas, como por exemplo, a remoção destas variáveis dos dados.

As variáveis latentes no PLS devem ser normalizadas. A normalização adotada por Wold (1985) e por Lohmöller (1984) assume que ξ_j têm desvio padrão igual a 1 (um). Fornell (1992) adotou um outro tipo de normalização para estimar a Satisfação do Consumidor, porém as variáveis latentes no modelo de Fornell e no de Wold são co-lineares. Detalhes sobre a metodologia de Fornell podem ser vistos em Bayol *et. al.* (2000).

4.4.1
Estimação

Conforme mencionado no capítulo anterior, o problema da identificação não ocorre no PLS, haja vista que as variáveis latentes são tratadas como combinações lineares ponderadas das variáveis observadas.

Os modelos baseados no PLS são estimados de acordo com os seguintes procedimentos: (1) padronização das variáveis observadas; (2) estimação externa, isto é, as variáveis latentes padronizadas são estimadas como combinação linear de suas variáveis observadas, com os pesos externos (ligação entre as observadas e as latentes) previamente inicializados; (3) estimação interna, isto é, cada variável latente é novamente estimada como uma combinação linear das variáveis latentes que estão conectadas a ela – nesta etapa são estimados os pesos ou coeficientes internos; (4) estimação dos pesos externos, isto é, os pesos previamente inicializados no primeiro passo são atualizados utilizando as variáveis latentes estimadas no passo anterior; (5) todo este procedimento é iterativo até a convergência dos pesos externos (não garantida, porém sempre encontrada),

resultando nas variáveis latentes estimadas; (6) no último passo são estimadas as relações estruturais do modelo através de regressões lineares com as variáveis latentes, isto é, com os seus escores.

Segue abaixo uma descrição mais detalhada de cada um dos passos da estimação.

4.4.1.1
Padronização das variáveis observadas

Em geral, as variáveis observadas podem ser padronizadas de quatro formas distintas. A escolha de uma delas dependerá de três condições que deverão ser verificadas através dos dados:

a) Condição 1: as escalas das variáveis observadas devem ser comparáveis. Por exemplo, no modelo da Satisfação do Consumidor, as escalas de todos os itens de avaliações variam de 1 a 10 (*cf.* capítulo 2), sendo, portanto, comparáveis. Por outro lado, pesos medidos em toneladas e velocidades medidas em km/h são exemplos de escalas não comparáveis.

b) Condição 2: as médias das variáveis observadas devem ser interpretáveis. Por exemplo, se a diferença entre duas variáveis observadas não for interpretável, então os seus respectivos parâmetros de locação não têm significado.

c) Condição 3: as variâncias das variáveis observadas refletem a sua importância.

Se a condição 1 não se verificar, então as variáveis observadas deverão ser padronizadas (média 0 e variância 1).

Se a condição 1 se verificar, então os resultados deverão ser obtidos com base nos dados brutos. Entretanto, os cálculos dos parâmetros do modelo dependerão da validade de outras condições:

a) As condições 2 e 3 não se verificam e as variáveis observadas estão padronizadas (média 0 e variância 1) para a fase de estimação dos parâmetros. Neste caso deve-se redefinir (retornar) as escalas das variáveis observadas para as suas médias e variâncias originais para a expressão final dos Pesos e das Cargas.

b) A condição 2 se verifica, porém a 3 não, e as variáveis observadas estão padronizadas com variância 1 (um) mas não estão centradas na média para a fase de estimação dos parâmetros. Neste caso se redefine as escalas das variáveis observadas para as suas variâncias originais para a expressão final dos Pesos e das Cargas.

c) As condições 2 e 3 se verificam: Neste caso deve-se utilizar as variáveis originais.

As condições acima podem ser resumidas na Tabela 4.1.

Tabela 4.1 – Condições para Padronização das Variáveis Observadas

Escalas das Variáveis são Comparáveis?	Médias das Variáveis são Interpretáveis?	Importância Relacionada c/ Variância	Média	Variância	Redefinir Escala?
Não	-	-	0	1	Não
Sim	Não	Não	0	1	Sim
Sim	Sim	Não	Original	1	Sim
Sim	Sim	Sim	Original	Original	-

4.4.1.2
Estimação das Variáveis Latentes

As variáveis latentes são estimadas de acordo com os seguintes procedimentos:

4.4.1.2.1

Estimação Externa Y_j das Variáveis Latentes Padronizadas ($\xi_j - m_j$)

As variáveis latentes padronizadas (média = 0 e variância = 1) são estimadas como combinações lineares de seus indicadores centrados na média, conforme a equação 4.38.

$$Y_j \propto \sum_{h=1}^{k_j} \varpi_{jh} \left(x_{jh} - \overline{x}_{jh} \right) \tag{4.38}$$

Onde: o símbolo "$\propto$" significa que a variável à esquerda representa a variável direita padronizada. A variável latente padronizada é finalmente escrita conforme a equação 4.39.

$$Y_j = \sum_{h=1}^{k_j} \tilde{\varpi}_{jh} \left(x_{jh} - \overline{x}_{jh} \right) \tag{4.39}$$

A média m_j é estimada pela equação 4.40.

$$\hat{m}_j = \sum_{h=1}^{k_j} \tilde{\varpi}_{jh} \overline{x}_{jh} \tag{4.40}$$

E a variável latente ξ_j pela equação 4.41.

$$\hat{\xi}_j = \sum_{h=1}^{k_j} \tilde{\varpi}_{jh} \overline{x}_{jh} = Y_j + \hat{m}_j \tag{4.41}$$

Onde: $\tilde{\varpi}_{jh}$ são chamados de Pesos Exteriores.

4.4.1.2.2

Estimação Interna Z_j das Variáveis Latentes Padronizadas ($\xi_j - m_j$)

De acordo com o algoritmo original do PLS de Wold (1985) e de Lohmöller (1989), a estimação interna de Z_j das variáveis latentes padronizadas ($\xi_j - m_j$) é definida de acordo com a equação 4.42.

$$Z_j = \sum_{i\,:\,\xi_i \text{ é conectado em } \xi_j}^{k_j} e_{ji} Y_i \tag{4.42}$$

Onde: os pesos internos e_{ji} podem ser escolhidos entre três esquemas: (1) ligação ponderada; (2) centróide; e (3) fator ponderado. Duas variáveis latentes são conectadas se existe uma ligação entre elas, isto é, existe uma seta que aponta de uma para a outra no diagrama de caminhos que descreve o modelo causal. Estes três esquemas são definidos da seguinte forma:

a) Esquema da ligação ponderada: as variáveis latentes conectadas em ξ_j são divididas em dois grupos: as predecessoras de ξ_j (que explicam ξ_j) e as seguidoras (que são explicadas por ξ_j). Para uma predecessora ξ_i da variável latente ξ_j, o peso interno e_{ji} será igual ao coeficiente da regressão de Y_i na regressão múltipla de Y_j em todos os Y_i's relacionados às predecessoras de ξ_j. Se ξ_i é uma sucessora da variável ξ_j então o peso interno e_{ji} será igual à correlação entre Y_i e Y_j.

b) Esquema do centróide: os pesos internos e_{ji} serão iguais aos sinais da correlação entre Y_i e Y_j. Esta foi a escolha original de Wold, entretanto ela apresenta uma desvantagem no caso da correlação ser aproximadamente zero, pois o sinal pode modificar em função de pequenas variações nas variáveis.

c) Esquema do fator ponderado: os pesos internos e_{ji} serão iguais às correlações entre Y_i e Y_j.

4.4.1.2.3
Estimação dos Pesos w_{jh}´s

Existem três formas de estimação dos pesos w_{jh}, chamados de modo A, B e C.

No modo A, o peso w_{jh} é o coeficiente de regressão de Z_i em uma simples regressão de x_{jh} na estimação interna Z_j, conforme a equação 4.43.

$$\hat{w}_{jh} = \text{cov}(x_{jh}, Z_j) / \text{Var}(Z_j) \tag{4.43}$$

No modo B, o vetor $\mathbf{w_j}$ de pesos w_{jh} é o vetor coeficiente de regressão na regressão múltipla de Z_j nas variáveis observadas x_{jh}, relacionadas com a mesma variável latente ξ_j , conforme a equação 4.44.

$$\hat{\mathbf{w}}_j = \left(\mathbf{X}_j'\mathbf{X}_j\right)^{-1}\mathbf{X}_j'\mathbf{Z}_j \tag{4.43}$$

Onde: $\mathbf{X_j}$ é a matriz com as colunas definidas pelas variáveis observadas x_{jh} relacionadas com a j-ésima variável latente ξ_j.

O modo C é um caso especial do modo B e refere-se principalmente aos indicadores do tipo formativo.

No algoritmo do PLS de Lohmöller, o processamento é iniciado atribuindo-se em cada bloco, o valor 1 (um) para o vetor de pesos w_{jh} em todas as variáveis observadas, exceto na última variável, cujo valor atribuído é –1. Esta escolha é a principal razão para eventuais estimativas negativas dos pesos exteriores, especialmente no caso quando existem poucas variáveis observadas num bloco (duas ou uma variável). Estes pesos são padronizados para se obter variáveis latentes com variância unitária.

A inicialização do vetor de pesos com valores iguais a 1 (um) em todas as variáveis observadas deveria ser a opção mais razoável, na hipótese de correlação positiva entre elas. Alternativamente, o vetor de pesos inicializado com o valor igual a 1 (um) na primeira variável observada e 0 (zero) nas demais deveria ser a forma mais fácil do ponto de vista computacional.

Os demais passos descritos acima, dependendo do modo selecionado, serão iterados até a convergência do modelo (não garantida, porém sempre encontrada na prática). Após o último passo do cálculo se obtém os pesos internos estimados $\tilde{w}_{jh}$, as variáveis latentes padronizadas Y_i, a média estimada $\hat{m}_j$ da variável latente ξ_j e a estimativa final $\hat{\xi}_j$ de ξ_j.

4.4.1.2.4
Estimação das Equações Estruturais

As equações estruturais são estimadas através de regressões lineares individuais utilizando o método dos mínimos quadrados ordinários, onde as variáveis latentes ξ_j são substituídas por $\hat{\xi}_j$. Tanto os coeficientes da regressão quanto o valor da estatística R^2 são resultados padrões dos programas de Lohmöller e de Chin.

4.4.2
Validação

O cálculo das estatística R^2 e Q^2 de Stone-Geisser entre cada variável latente endógena e suas respectivas variáveis observadas está disponível nos *softwares* do PLS. O nível de significância dos coeficientes da regressão pode ser calculado através de métodos de validação cruzada, tais como *Jackknife* ou *Boostrap*, ou através da estatística *t*-Student, bastando exportar as estimativas das variáveis latentes e as respectivas variáveis observadas para qualquer *software* estatístico estimar as regressões lineares.

No *software* LVPLS, por exemplo, a metodologia de validação do modelo segue os seguintes procedimentos:

a) A matriz de dados é dividida em G grupos. O valor G = 7 é recomendado por Wold.

b) Grupos de células são removidos em torno dos dados, no sentido de eliminar alguns dados.

c) O PLS processa o modelo G vezes excluindo um dos grupos de cada vez.

d) Uma forma de avaliar a qualidade do modelo consiste em medir a sua capacidade de prever variáveis observadas relacionadas com as variáveis latentes endógenas. Dois índices são utilizados para isto: comunalidade e redundância.

e) Na opção da comunalidade, os valores previstos das variáveis observadas centradas na média que não são utilizadas na análise, serão obtidos através da equação 4.44.

$$\mathrm{Pred}\left(x_{jhi} - \bar{x}_{jh}\right) = \hat{\pi}_{jh} Y_{ji} \tag{4.44}$$

Onde $\hat{\pi}_{jh}$ e Y_{ji} são calculados nos dados onde o i-ésimo valor da variável x_{jh} está faltando.

O valor da comunalidade é calculado conforme a equação 4.45.

$$\begin{aligned}
SSO_{jh} &= \sum_i \left(x_{jhi} - \bar{x}_{jh}\right)^2 \\
SSE_{jh} &= \sum_i \left(x_{jhi} - \bar{x}_{jh} - \hat{\pi}_{jh} Y_{ji}\right)^2 \\
SSO_j &= \sum_h SSO_{jh} \\
SSE_j &= \sum_h SSE_{jh} \\
H_j^2 &= 1 - \frac{SSE_j}{SSO_j}
\end{aligned} \tag{4.44}$$

Onde SSO_{jh} é a soma dos quadrados das observações de uma variável latente, SSE_{jh} é a soma dos quadrados dos erros de previsão de uma variável latente, SSO_j é a soma dos quadrados das observações para o bloco j, SSE_j é a soma dos quadrados dos erros de previsão para o bloco j e H^2 é a medida de comunalidade para o bloco j.

f) Na opção de redundância, os valores previstos das variáveis observadas centradas na média que não são utilizadas na análise serão obtidos através da equação 4.45.

$$\mathrm{Pred}\left(x_{jhi} - \bar{x}_{jh}\right) = \hat{\pi}_{jh}\ \mathrm{Pred}\left(Y_{ji}\right) \tag{4.45}$$

Onde $\hat{\pi}_{jh}$ é o mesmo do caso anterior e Pred (Y_{ji}) é a previsão para a i-ésima observação da variável latente endógena Yj, usando o modelo de regressão conforme a equação 4.46 e calculado nos dados onde o i-ésimo valor da variável x_{jh} está faltando.

$$\mathrm{Pred}\left(Y_i\right) = \sum_{j' : \xi_{j'}\ \text{explicando}\ \xi_j} \hat{\beta}_{j'} Y_{j'} \tag{4.46}$$

O valor da redundância é calculado conforme a equação 4.47.

$$\begin{aligned} SSE'_{jh} &= \sum_i \left(x_{jhi} - \bar{x}_{jh} - \hat{\pi}_{jh}\ \mathrm{Pred}\left(Y_{ji}\right)\right)^2 \\ SSE'_j &= \sum_h SSE'_{jh} \\ F_j^2 &= 1 - \frac{SSE'_j}{SSO_j} \end{aligned} \tag{4.47}$$

g) No método *Jackknife*, as médias e os desvios padrões dos parâmetros do modelo (pesos, cargas, parâmetros de ligação, correlações entre variáveis latentes) são calculados utilizando os resultados da análise anterior. Entretanto o valor G = 7 é considerado pequeno, sugerindo-se um valor maior, por exemplo, G = 30.

O PLS-Graph (*software* desenvolvido por Chin), contempla técnicas adicionais para o calculo da significância dos parâmetros nos modelos PLS e, além do método anterior de validação, ele também provê opções de amostragem por *Boostrap* e *Jackknife*.