ESTRATÉGIAS PARA MODELAGEM DE DADOS MULTIVARIADOS NA PRESENÇA DE CORRELAÇÃO

Flavio S. Fogliatto
Departamento de Engenharia de Produção & Transportes
Universidade Federal do Rio Grande do Sul
Praça Argentina, 9 – Sala LOPP
Porto Alegre, RS – 90020-000
E-mail: ffogliatto@ppgep.ufrgs.br

Resumo

Dados multivariados ocorrem com freqüência em investigações empíricas. Em estudos de Engenharia, por exemplo, dados multivariados são coletados ao estudar-se o efeito de diferentes condições de processamento sobre características de itens manufaturados. Tais conjuntos de dados podem apresentar variáveis altamente correlacionadas. Neste artigo, investiga-se o efeito da estrutura de correlação de variáveis dependentes em sua modelagem, a partir de regressão linear. Quatro técnicas de regressão são apresentadas e comparadas: regressão de mínimos quadrados ordinários, regressão de mínimos quadrados generalizados, regressão por equações aparentemente não relacionadas e regressão multivariada. Como modelos de regressão são, via de regra, utilizados com fins preditivos, as técnicas de modelagem acima são comparadas com base em sua variância de predição. As diferentes técnicas de regressão são ilustradas em um estudo de caso.

Palavras-chave: técnicas de regressão multivariada, variância de predição, correlação.

1. Introdução

A Análise de Regressão Linear (ARL) é uma das ferramentas estatísticas mais utilizadas na modelagem de dados. A ARL consiste, em sua essência, na determinação de uma equação ou modelo que descreva de maneira eficiente o efeito de um grupo de variáveis independentes sobre uma ou mais variáveis dependentes. A aplicação da técnica de modelagem por

regressão linear a um grupo de dados resulta na determinação de coeficientes lineares que ponderam o efeito de variáveis independentes sobre variáveis dependentes. Modelos com uma única variável dependente são ditos univariados. Modelos com múltiplas variáveis dependentes são ditos multivariados.

Dados multivariados ocorrem com freqüência em investigações empíricas. Em estudos econômicos, por exemplo, pode-se avaliar o efeito de medidas de ajuste tributário sobre indicadores de desempenho econômico (ADELMAN *et al*., 1969). Em estudos de Engenharia, pode-se estudar o efeito de diferentes ajustes nos controles de um equipamento sobre as características de unidades por ele produzidas (FOGLIATTO *et al.*, 1998). Em ambos os casos, deseja-se analisar o efeito de um grupo de variáveis independentes (medidas de ajuste tributário, ajustes nos controles de um equipamento) sobre um grupo de variáveis dependentes (indicadores de desempenho econômico, características do produto). Outros exemplos podem ser encontrados em JOHNSON & WICHERN (1992).

Neste trabalho, analisa-se o efeito da estrutura de correlação de variáveis dependentes na modelagem de grupos de dados multivariados. Para esse fim, quatro técnicas de modelagem via ARL são comparadas relativamente a um critério predeterminado; são elas: (*i*) regressão de mínimos quadrados ordinários (RMQ), (*ii*) regressão de mínimos quadrados generalizados (RMG), (*iii*) regressão por equações aparentemente não relacionadas (SURE – *seemingly unrelated equations regression*), e (*iv*) regressão multivariada (RMV); ver DRAPER & SMITH (1981), MYERS (1986), ZELLNER (1962) e SEBER (1984), respectivamente. Cada técnica considera de maneira distinta a estrutura de correlação entre variáveis dependentes na estimação dos coeficientes a serem utilizados nos modelos de regressão. Cabe ressaltar que a estrutura de correlação entre variáveis independentes é identicamente considerada em todas as estratégias, não servindo, assim, como critério diferenciador.

Modelos de regressão são, via de regra, utilizados para fins de predição, estimação e controle (MONTGOMERY & PECK, 1992). Em todos os casos, desejam-se modelos que possam ser utilizados como estimadores eficientes das variáveis dependentes modeladas. Assim, sugere-se como base para comparação das técnicas de modelagem listadas acima, a variância das predições geradas a partir de cada modelo. A melhor técnica será aquela que, na média, gerar predições com menor variância. As técnicas de modelagem citadas acima são brevemente introduzidas na seqüência.

Considere um grupo de variáveis independentes utilizadas na modelagem de variáveis dependentes. Variáveis dependentes podem ser modeladas individualmente (ou seja, desconsiderando eventuais correlações entre elas) ou em conjunto. Nas regressões RMQ e RMG, supõe-se correlação inexistente e modelam-se variáveis dependentes individualmente. Essas estratégias de regressão são as mais utilizadas na prática, principalmente por encontrarem-se implementadas em pacotes computacionais de análise estatística e serem de fácil compreensão por parte do analista. Nem sempre, todavia, essas técnicas constituem uma escolha adequada. Por exemplo, DERRINGER & SUICH (1980) e RIBEIRO & ELSAYED (1995) utilizam RMQ na modelagem de variáveis dependentes altamente correlacionadas (correlações da ordem de 0,8), resultando em modelos com alta variância de predição.

Na regressão SURE, variáveis dependentes são modeladas simultaneamente e a correlação entre variáveis é considerada na modelagem. Esta técnica de modelagem é bastante comum em estudos de Econometria. A regressão MVR é um caso especial de SURE, onde cada variável dependente é modelada como função de um mesmo grupo de variáveis independentes, porém com diferentes coeficientes de regressão. A regressão SURE produz modelos cujas predições apresentam variância pelo menos tão pequena quanto aquelas obtidas usando as demais técnicas de regressão (SRIVASTAVA & GILES,

1987). Assim, recomenda-se a modelagem de variáveis dependentes correlacionadas utilizando a regressão SURE.

Variáveis dependentes apresentam-se correlacionadas em situações em que a avaliação direta de algum atributo ou propriedade em unidades experimentais é difícil. Assim, o pesquisador seleciona um conjunto de variáveis relacionadas ao atributo em questão as quais, via de regra, apresentam-se correlacionadas. Esta é a situação encontrada no estudo de caso apresentado neste trabalho, cujos dados são utilizados para comparar as diferentes técnicas de regressão descritas acima.

Este trabalho divide-se em cinco seções, incluindo a presente introdução. Na seção 2, introduz-se a notação e estrutura genérica do modelo de regressão a ser usado nas seções seguintes. A seção 3 é dividida em seis subseções: as quatro subseções iniciais trazem descrições detalhadas das técnicas de regressão contempladas neste trabalho (ou seja, RMQ, RMG, SURE e RMV); um estimador de correlação amostral é apresentado na seqüência; a última subseção traz uma comparação entre técnicas. Na quarta seção, um exemplo numérico com dados obtidos em um estudo de caso ilustra a aplicação das técnicas de regressão. A última seção traz a conclusão do trabalho.

2. Considerações Preliminares

A seguinte notação e definições são utilizadas neste trabalho. Letras maiúsculas em negrito designam matrizes e letras minúsculas em negrito designam vetores. O inverso de uma matriz $\mathbf{A}$ é designado por $\mathbf{A}^{-1}$ e sua transposta por $\mathbf{A}'$; analogamente, $\mathbf{a}'$ designa o transposto de um vetor $\mathbf{a}$. Uma matriz identidade de dimensão N por N é designada por $\mathbf{I}_N$. O operador $\otimes$ designa o produto direto, ou de Kroenecker, de matrizes (seja $\mathbf{A}$ uma matriz com elementos designados por a_{ij}; então $\mathbf{A} \otimes \mathbf{B}$ corresponde a uma matriz de blocos, com blocos dados pelo produto $a_{ij}\mathbf{B}$). A função $tr(\mathbf{A})$ designa o traço de uma matriz $\mathbf{A}$ e é dada pela soma dos elementos de sua diagonal principal. Uma matriz definida positiva apresenta somente elementos positivos em sua diagonal principal.

Considere um grupo de dados multivariados formado por P variáveis dependentes e C variáveis independentes, observadas em T situações ou níveis distintos. Corriqueiramente, as observações das variáveis independentes corresponderiam aos tratamentos em um experimento planejado. O vetor $\mathbf{x} = [x_1,\ldots, x_C]'$ apresenta os valores observados para as C variáveis independentes em uma dada situação. $Y_i(\mathbf{x})$ designa o valor da $i^{ésima}$ variável dependente quando os níveis das variáveis independentes correspondem a $\mathbf{x}$.

As variáveis dependentes são preditas por modelos de regressão desenvolvidos a partir de um grupo de dados com características dadas acima. Supõe-se o seguinte modelo genérico de regressão para a $i^{ésima}$ variável dependente:

$$\mathbf{y}_i = \mathbf{X}_i\beta_i + \mathbf{u}_i \quad , \; i = 1,\ldots,P \tag{1}$$

onde $\mathbf{y}_i = [Y_{1i},\ldots, Y_{Ti}]'$ é um vetor (T x 1) de observações da variável dependente; $\mathbf{X}_i$ é a matriz (T x K_i) de regressores, onde K_i indica o número de regressores no $i^{ésimo}$ modelo; $\boldsymbol{\beta}_i = [\beta_{0i}, \beta_{1i},\ldots, \beta_{(K_i-1)i}]'$ é um vetor (K_i x 1) de coeficientes de regressão; e $\mathbf{u}_i = [u_{1i},\ldots, u_{Ti}]'$ é um vetor (T x 1) formado por resíduos supostamente seguindo uma distribuição Normal, com matriz de covariâncias dada por

$$\mathbf{D}[\mathbf{u}_i] = \mathbf{V}_i \tag{2}$$

onde $\mathbf{V}_i$ é uma matriz (T x T) definida positiva que se supõe conhecida (um caso especial de $\mathbf{V}_i$ é $\mathbf{V}_i = \sigma_i^2.\mathbf{I}_T$, ou seja, resíduos não correlacionados com variância comum dada por σ_i^2). Por exemplo, seja $K_1 = 4$; então o modelo de regressão para a variável dependente Y_1 possui quatro termos. Os termos, também denominados regressores, são (por exemplo) a média, x_1, x_2^2, e $x_1 x_2$. Suponha que a primeira situação ou tratamento experimental observado corresponda a $x_1 = -1$ e $x_2 = -1$. Assim, a primeira linha da matriz $\mathbf{X}_1$ será dada por (1, -1, 1, 1) e o vetor de coeficientes $\boldsymbol{\beta}_1$ por $(\beta_{01}, \beta_{11}, \beta_{221}, \beta_{121})'$.

As P equações em (1) podem ser escritas como uma única equação,

$$\mathbf{y}=\mathbf{X}\beta+\mathbf{u} \tag{3}$$

onde $\mathbf{y} = [\mathbf{y}_1,\ldots, \mathbf{y}_P]'$ é um vetor (TP x 1) de variáveis dependentes, $\mathbf{X}$ é uma matriz de regressores do tipo diagonal em blocos, de dimensão (TP x $\sum_{i=1}^{P} K_i$), com os blocos na diagonal principal dados pelas matrizes $\mathbf{X}_i$, i=1,…,P (os demais blocos da matriz são $\mathbf{0}$), $\boldsymbol{\beta} = [\boldsymbol{\beta}_1,\ldots, \boldsymbol{\beta}_P]'$ é um vetor ($\sum_{i=1}^{P} K_i$ x 1) de coeficientes, e $\mathbf{u} = [\mathbf{u}_1,\ldots, \mathbf{u}_P]'$ é um vetor (TP x 1) de resíduos. Note que os vetores $\mathbf{y}$, $\boldsymbol{\beta}$ e $\mathbf{u}$ são vetores formados por vetores.

A matriz de covariâncias de $\mathbf{u}$, designada por $\mathbf{D}[\mathbf{u}]$, é uma matriz de blocos com o seguinte formato

$$\mathbf{D}[\mathbf{u}]=\begin{bmatrix} \sigma_1^2\mathbf{I}_T & \sigma_{12}\mathbf{I}_T & \hbar & \sigma_{1P}\mathbf{I}_T \\ \sigma_{21}\mathbf{I}_T & \sigma_2^2\mathbf{I}_T & \hbar & \sigma_{2P}\mathbf{I}_T \\ \hbar & \hbar & \hbar & \hbar \\ \sigma_{P1}\mathbf{I}_T & \sigma_{P2}\mathbf{I}_T & \hbar & \sigma_P^2\mathbf{I}_T \end{bmatrix} = \Sigma \otimes \mathbf{I}_T \tag{4}$$

onde $\boldsymbol{\Sigma} = [\sigma_{ij}]$, i,j=1,...,P, e σ_{ij} corresponde à covariância entre variáveis dependentes i e j. Um estimador para σ_i^2 e σ_{ij} é apresentado na próxima seção.

Estimando-se $\boldsymbol{\beta}_i$ em (1), pode-se predizer o valor $Y_i(\mathbf{x})$ da $i^{ésima}$ variável dependente em um dado arranjo $\mathbf{x}$ das variáveis independentes. Deseja-se um estimador $\hat{\boldsymbol{\beta}}_i$ de $\boldsymbol{\beta}_i$ que gere predições $\hat{Y}_i$ tão próximas quanto possível dos valores observados Y_i. Quatro estimadores de $\boldsymbol{\beta}_i$ são apresentados na próxima seção, cada um associado a diferentes suposições.

O desempenho dos diferentes estimadores de $\boldsymbol{\beta}_i$ será avaliado por sua variância de predição (ou seja, por sua eficiência; ver MOOD, 1974). Um bom estimador de $\boldsymbol{\beta}_i$ resulta em predições com pequena variância. A variância de predição da $i^{ésima}$ variável dependente, avaliada em um dado arranjo $\mathbf{x}$ das variáveis independentes, é dada por

$$\begin{aligned} V[\hat{Y}_i(\mathbf{x})] &= V[\hat{\beta}_{0i}+\hat{\beta}_{1i}x_1+\hbar+\hat{\beta}_{(K_i-1)i}x_{K_i-1}] = \\ &= V(\hat{\beta}_{0i})+x_1^2V(\hat{\beta}_{1i})+\hbar+x_{K_i-1}^2V(\hat{\beta}_{(K_i-1)i})+ \\ &\quad +2x_1Cov(\hat{\beta}_{0i},\hat{\beta}_{1i})+\hbar+ \\ &\quad +2x_{K_i-2}x_{K_i-1}Cov(\hat{\beta}_{(K_i-2)i},\hat{\beta}_{(K_i-1)i}) \end{aligned} \tag{5}$$

A expressão acima deixa claro que a variância de predição depende do método utilizado para estimação dos coeficientes de regressão.

3. Quatro Técnicas para Modelagem de Dados Multivariados Através de Regressão Linear

As variáveis dependentes em (1) podem ser modeladas individualmente ou simultaneamente como função das variáveis independentes, conforme as suposições feitas acerca do vetor de resíduos $\mathbf{u}_i$. Supondo Cov($\mathbf{u}_i$, $\mathbf{u}_j$) = 0, i, j = 1,…,P, $i \neq j$, a modelagem individual ou simultânea das variáveis dependentes resulta no mesmo conjunto de coeficientes de regressão e, por simplicidade, opta-se pela modelagem individual. Todavia, ao supor-se Cov($\mathbf{u}_i$, $\mathbf{u}_j$) $\neq$ 0, deve-se considerar uma estratégia que permita a modelagem simultânea das variáveis dependentes.

Nesta seção são apresentadas quatro técnicas para estimação dos coeficientes de regressão em (1). Nas primeiras duas técnicas, RMQ e RMG, as variáveis dependentes são modeladas individualmente, supondo inexistência de correlação entre elas. Nas duas últimas técnicas, SURE e RMV, supõe-se variáveis correlacionadas, as quais são modeladas simultaneamente. Um método para estimação da matriz de covariâncias $\boldsymbol{\Sigma}$ em (4) é apresentado na seqüência. A seção é concluída com uma comparação entre as técnicas de regressão abordadas.

3.1 Modelagem Individual de Variáveis Dependentes Através da Regressão de Mínimos Quadrados Ordinários (RMQ)

Este é a técnica de regressão mais comumente encontrada na literatura (ver, por exemplo, CHATTERJEE & PRICE, 1991; STAPLETON,

1995; e DANIEL & WOOD, 1980). Estimativas dos coeficientes $\boldsymbol{\beta}_i$ em (1) são obtidas minimizando a soma dos quadrados dos resíduos em $\mathbf{u}_i$, pela equação:

$$\hat{\beta}_i = \left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\mathbf{y}_i$$

Pode-se demonstrar que $\hat{\beta}_i$ é o melhor estimador linear não tendencioso de $\boldsymbol{\beta}_i$ (STAPLETON, 1995, p.87). A matriz de covariâncias de $\hat{\beta}_i$ é dada por:

$$\mathbf{D}[\hat{\beta}_i] = \sigma_i^2\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1} = \left[\frac{\left(\mathbf{y}_i - \mathbf{X}_i\hat{\beta}_i\right)'\left(\mathbf{y}_i - \mathbf{X}_i\hat{\beta}_i\right)}{T - K_i}\right]\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1} \quad (6)$$

onde $\sigma_i^2 = V(\mathbf{u}_i)$ é normalmente estimado a partir da média do quadrado dos resíduos.

A variância do valor predito $\hat{Y}_i$ para um dado arranjo **x** das variáveis independentes é obtida a partir da expressão em (5), usando a informação em (6); isto é:

$$V\left[\hat{Y}_i(\mathbf{x})\right] = \sigma_i^2\left(\mathbf{x}'\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\mathbf{x}\right)$$

3.2 Modelagem Individual de Variáveis Dependentes Através da Regressão de Mínimos Quadrados Generalizados (RMG)

A regressão RMG generaliza a regressão RMQ apresentada acima. As suposições são: (*i*) $\mathbf{V}_i$ em (2) corresponde a uma matriz definida positiva *qualquer* (ou seja, resíduos em $\mathbf{u}_i$ podem estar correlacionados e apresentar variâncias desiguais) e (*ii*) variáveis dependentes não se apresentam correlacionadas.

Considerando as suposições acima, chega-se aos seguintes estimadores dos coeficientes $\boldsymbol{\beta}_i$ em (1) (MYERS, 1986):

$$\hat{\beta}_i = \left(\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{y}_i, \quad (7)$$

com matriz de covariâncias de $\hat{\beta}_i$ dada por:

$$\mathbf{D}[\hat{\beta}_i] = \left(\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}. \quad (8)$$

A variância do valor predito $\hat{Y}_i$ para um dado arranjo **x** das variáveis independentes é obtida a partir de (5) e (8); isto é:

$$V\left[\hat{Y}_i(\mathbf{x})\right] = \left(\mathbf{x}'\left(\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{x}\right). \quad (9)$$

A matriz de covariâncias $\mathbf{V}_i$ é, via de regra, desconhecida e deve ser estimada a partir de dados experimentais. A estimação de $\mathbf{V}_i$ é restrita a situações em que repetidas observações das variáveis dependentes $\mathbf{y}_i$ estão disponíveis para um mesmo arranjo **x** das variáveis independentes. A partir de múltiplas observações de $Y_i(\mathbf{x})$, calculam-se variâncias e covariâncias amostrais utilizadas na determinação da matriz estimada de covariâncias, $\hat{\mathbf{V}}_i$. A matriz $\hat{\mathbf{V}}_i$ só deve ser usada quando variâncias e covariâncias amostrais forem estimadas a partir de amostras de tamanho considerável (mais de oito observações de Y_i para cada **x**, conforme sugerido por DEATON *et al.*, 1983). Caso contrário, a modelagem das variáveis dependentes via RMQ costuma gerar resultados mais confiáveis.

3.3 Modelagem Simultânea de Variáveis Dependentes Através da Regressão SURE (Regressão por Equações Aparentemente Não Relacionadas)

Considere resíduos em (3) com matriz de covariâncias conforme apresentado em (4). RMQ e RMG não podem ser aplicados na estimação dos coeficientes $\boldsymbol{\beta}_i$ em (1), já que os vetores $\mathbf{u}_i$ não são independentes. Neste caso, procede-se com a seguinte transformação visando eliminar a dependência entre vetores $\mathbf{u}_i$, $i = 1,\ldots,P$ (SEBER, 1977): considere a matriz de covariâncias em (4) e suponha que $\mathbf{D}[\mathbf{u}] = \mathbf{V}$, onde **V** é uma matriz definida positiva. Assim, pode-se determinar uma matriz não singular **K** tal que

$$\mathbf{V}=\mathbf{K}\mathbf{K}'.$$

Na seqüência, os elementos em (3) são transformados e renomeados, tal que $\mathbf{Z} = \mathbf{K}^{-1}\mathbf{y}$, $\mathbf{B} = \mathbf{K}^{-1}\mathbf{X}$, e $\boldsymbol{\eta} = \mathbf{K}^{-1}\mathbf{u}$. A expressão em (3) pode ser reescrita como:

$$\mathbf{Z}=\mathbf{B}\beta+\eta \qquad (10)$$

onde **B** é uma matriz ($PT \times \sum_i K_i$) com *rank* dado por $\sum_i K_i$ ($\mathbf{K^{-1}}$ é não singular por definição, o que implica nas seguintes igualdades: *rank* [$\mathbf{K^{-1}X}$] = *rank* [**X**] = $\sum_i K_i$). Pode-se demonstrar que $\mathbf{D}[\eta] = \sigma^2 \mathbf{I}_{PT}$. Um estimador dos coeficientes em (10), obtido ao minimizar-se a soma dos quadrados de η, vem dado por:

$$\hat{\beta} = \left(\mathbf{X'V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X'V}^{-1}\mathbf{y}\,.$$

Note que $\mathbf{V^{-1}} = (\mathbf{D}[\mathbf{u}])^{-1}$. Assim, reescrevendo a expressão acima, obtém-se:

$$\hat{\beta} = \left(\mathbf{X'}\left(\mathbf{D}[\mathbf{u}]\right)^{-1}\mathbf{X}\right)^{-1}\mathbf{X'}\left(\mathbf{D}[\mathbf{u}]\right)^{-1}\mathbf{y} \qquad (11)$$

Pode-se demonstrar que $E\left(\hat{\beta}\right) = \beta$ (SRIVASTAVA & GILES, 1987). A matriz de covariâncias dos coeficientes em (11) é dada por:

$$\mathbf{D}\left[\hat{\beta}\right] = \left(\mathbf{B'B}\right)^{-1} = \left(\mathbf{X'V}^{-1}\mathbf{X}\right)^{-1} = \left(\mathbf{X'}\left(\mathbf{D}[\mathbf{u}]\right)^{-1}\mathbf{X}\right)^{-1} \qquad (12)$$

Após estimar $\hat{\beta}$ usando a expressão em (11), os modelos para cada variável dependente podem ser separados e a variância do valor predito $\hat{Y}_i$ para um dado arranjo **x** das variáveis independentes determinada usando a expressão em (5) e a informação em (12).

O método apresentado acima foi originalmente concebido por AITKEN (1935) para lidar com casos de regressão múltipla univariada (ou seja, $P = 1$) apresentando resíduos correlacionados. ZELLNER (1962) estendeu o método para contemplar casos de regressão multivariada, chegando ao resultado em (11). Modelos obtidos utilizando a expressão em (11) são denominados "equações de regressão aparentemente não relacionadas" (*Seemingly Unrelated Regression* (SURE) *Models*).

A regressão SURE é operacionalizada modelando-se inicialmente as variáveis dependentes via RMQ. Uma vez conhecidas as matrizes de regressores $\mathbf{X}_i$, i = 1,…, P, elas podem ser arranjadas em uma matriz **X**, a qual é então usada em (11). Por essa razão, os modelos de regressão obtidos por RMQ e SURE compartilham dos mesmos regressores.

3.4 Modelagem Simultânea de Variáveis Dependentes Através de Regressão Multivariada (RMV)

Esta regressão corresponde ao análogo multivariado de RMQ, descrito na seção 3.1. Supõe-se que as P variáveis dependentes são modeladas pelos mesmos regressores; isto é, $\mathbf{X}_1 = \mathbf{X}_2 = \ldots = \mathbf{X}_P = \mathbf{Z}$ e $K_1 = K_2 = \ldots = K_P = K$. Após substituições, o modelo em (3) assume a seguinte forma:

$$\mathbf{y}=\left(\mathbf{I}_P \otimes \mathbf{Z}\right)\beta+\mathbf{u} \qquad (13)$$

onde **y** é um vetor (TP x 1) de variáveis dependentes, com elemento $\mathbf{y}_i$, **Z** é uma matriz (T x K) de regressores, β é um vetor (KP x 1) de coeficientes, com elemento β_i, e **u** é um vetor (TP x 1) de resíduos, com elemento $\mathbf{u}_i$.

As suposições acerca do vetor **u** em (13) vêm dadas em (4). Os vetores $\mathbf{u}_i$ podem apresentar-se correlacionados. Todavia, quando todas as variáveis dependentes são modeladas pelos mesmos regressores, considerando ou não a correlação entre vetores de resíduos leva aos mesmos estimadores de β, como demonstrado a seguir.

Considere o vetor de estimadores $\hat{\beta}^{(1)}$ de β em (13), obtido pela expressão para $\hat{\beta}$ de RMQ, na seção 3.1 (com vetores $\mathbf{u}_i$ considerados independentes):

$$\hat{\beta}^{(1)} = \left[\left(\mathbf{I}'_P \otimes \mathbf{Z}'\right)\left(\mathbf{I}_P \otimes \mathbf{Z}\right)\right]^{-1}\left(\mathbf{I}'_P \otimes \mathbf{Z}'\right)\mathbf{y}\,.$$

Como $\left(\mathbf{A}\otimes\mathbf{B}\right)'\left(\mathbf{C}\otimes\mathbf{D}\right) = \mathbf{A'C}\otimes\mathbf{B'D}$, a equação acima equivale a:

$$\hat{\beta}^{(1)} = \left(\mathbf{I}'_P \otimes \left(\mathbf{Z'Z}\right)^{-1}\mathbf{Z'}\right)\mathbf{y}$$

Modelando as variáveis dependentes usando a regressão SURE (observando que MVR é um

caso especial de SURE e que, em ambos os casos, a correlação entre vetores $\mathbf{u}_i$ é considerada), chega-se a um vetor de estimadores $\hat{\beta}^{(2)}$ de β dado por:

$$\hat{\beta}^{(2)} = \left(\mathbf{X}'\left(\mathbf{D}[\mathbf{u}]\right)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}\left(\mathbf{D}[\mathbf{u}]\right)^{-1}\mathbf{y} =$$
$$= \left(\mathbf{X}'[\Sigma \otimes \mathbf{I}_T]\mathbf{X}\right)^{-1}\mathbf{X}[\Sigma \otimes \mathbf{I}_T]\mathbf{y} =$$
$$= \left[(\mathbf{I}_P \otimes \mathbf{Z}')(\Sigma^{-1} \otimes \mathbf{I}_T)(\mathbf{I}_P \otimes \mathbf{Z})\right]^{-1}[\mathbf{I}_P \otimes \mathbf{Z}'][\Sigma^{-1} \otimes \mathbf{I}_T]\mathbf{y},$$

o qual, dado que $(\mathbf{A} \otimes \mathbf{B})'(\mathbf{C} \otimes \mathbf{D}) = \mathbf{A}'\mathbf{C} \otimes \mathbf{B}'\mathbf{D}$, pode ser reescrito como:

$$\hat{\beta}^{(2)} = \left[(\mathbf{I}'_P\Sigma^{-1} \otimes \mathbf{Z}'\mathbf{I}_T)(\mathbf{I}_P \otimes \mathbf{Z})\right]^{-1}\left[\mathbf{I}'_P\Sigma^{-1} \otimes \mathbf{Z}'\mathbf{I}_T\right]\mathbf{y} =$$
$$= \left[(\Sigma\Sigma^{-1} \otimes (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}')\mathbf{Z}'\right]\mathbf{y} = \left(\mathbf{I}'_P \otimes (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\right)\mathbf{y}$$

Os estimadores $\hat{\beta}^{(1)}$ e $\hat{\beta}^{(2)}$ são idênticos (ou seja, considerando-se ou não a correlação entre vetores $\mathbf{u}_i$ na estimação dos coeficientes de regressão chega-se ao mesmo resultado). Como esses estimadores independem de $\mathbf{D}[\mathbf{u}]$, pode-se modelar as variáveis dependentes individualmente, simplificando o trabalho algébrico.

3.5 Estimação da Matriz de Covariâncias Σ

A matriz Σ em (4) é, via de regra, desconhecida, sendo estimada a partir de dados amostrais. Assim, para fins práticos, Σ pode ser substituída por uma estimativa $\hat{\Sigma} = [\hat{\sigma}_{ij}]$. ZELLNER (1962) propõe um estimador de $\hat{\sigma}_{ij}$ baseado nos resíduos obtidos pela modelagem de variáveis dependentes via RMQ e dado por:

$$\hat{\sigma}_{ij} = \frac{\mathbf{y}'_i\left[\mathbf{I}_T - \mathbf{X}_i(\mathbf{X}'_i\mathbf{X}_i)^{-1}\mathbf{X}'_i\right]\left[\mathbf{I}_T - \mathbf{X}_j(\mathbf{X}'_j\mathbf{X}_j)^{-1}\mathbf{X}'_j\right]\mathbf{y}_j}{D}, \quad i,j = 1,\ldots,P,$$

onde $D = T - K$, se as variáveis dependentes i e j forem modeladas pelos mesmos regressores (como descritos na seção 3.4). Caso contrário, D será dado por:

$$D = T - K_i - K_j + tr\left[(\mathbf{X}'_i\mathbf{X}_i)^{-1}\mathbf{X}'_i\mathbf{X}_j(\mathbf{X}'_j\mathbf{X}_j)^{-1}\mathbf{X}'_j\mathbf{X}_i\right].$$

3.6 Comparação das Variâncias dos Estimadores Obtidos a Partir das Diferentes Técnicas de Regressão e Comentários Gerais

Como mencionado anteriormente, modelos de regressão são normalmente utilizados para fins preditivos. Assim, atenção especial deve ser dada à variância dos estimadores dos coeficientes de regressão usado nesses modelos. A variância dos estimadores $\hat{\beta}$, $V(\hat{\beta})$, determina a variância de valores preditos, conforme apresentado em (5). Desta forma, deve-se sempre optar por uma modelagem que resulte nos menores valores possíveis para $V(\hat{\beta})$.

Na seqüência, apresenta-se uma comparação entre as variâncias das estimativas de β conforme obtidas pelos métodos RMQ e SURE. Os estimadores RMQ e SURE de β são denominados $\hat{\beta}^{(1)}$ e $\hat{\beta}^{(2)}$, respectivamente, seguindo a notação introduzida na seção 3.4. $V(\hat{\beta}^{(1)})$ e $V(\hat{\beta}^{(2)})$ estão apresentadas em (6) e (12), respectivamente. Seja $\mathbf{D}[\mathbf{u}] = \mathbf{V}$. Considere a seguinte quantidade (SRIVASTAVA & GILES, 1987):

$$V\left(\hat{\beta}^{(1)}\right) - V\left(\hat{\beta}^{(2)}\right) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} - \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}$$

Seja $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}$; assim, pode-se reescrever a expressão acima da seguinte maneira:

$$V\left(\hat{\beta}^{(1)}\right) - V\left(\hat{\beta}^{(2)}\right) = \mathbf{A}\mathbf{V}\mathbf{A}'.$$

Sendo $\mathbf{V}$ uma matriz definida positiva por construção, $\mathbf{A}\mathbf{V}\mathbf{A}'$ é pelo menos semidefinida positiva (elementos na diagonal principal são ≥ 0; ver SEBER, 1977, p.385), o que implica em $V\left(\hat{\beta}^{(1)}\right) \geq V\left(\hat{\beta}^{(2)}\right)$. Desta forma, $\hat{\beta}^{(2)}$ será pelo menos tão eficiente quanto $\hat{\beta}^{(1)}$. Sugere-se, assim, que $\hat{\beta}^{(2)}$ seja sempre preferido como estimador dos coeficientes dos modelos em (1). Observe que $V\left(\hat{\beta}^{(1)}\right) = V\left(\hat{\beta}^{(2)}\right)$ quando $\mathrm{Cov}(\mathbf{u}_i, \mathbf{u}_j) = 0$.

Conforme apresentado nas seções 3.1 a 3.4, cada técnica de regressão implica em suposições

acerca dos vetores de resíduos e da relação entre variáveis dependentes. Essas suposições são comentadas na seqüência.

- Variáveis dependentes não correlacionadas com variâncias não homogêneas podem ser modeladas através de RMG, disponível no pacote estatístico SAS (1990).
- RMV é adequada na estimação de coeficientes de regressão de variáveis correlacionadas e modeladas pelos mesmos regressores; RMV é um rotina disponível no pacote estatístico SAS (1990).
- SURE é adequada quando as variáveis dependentes apresentam-se correlacionadas; SURE é uma rotina disponível no pacote estatístico SAS (1990).

SURE sempre resulta em estimativas dos coeficientes de regressão com variância menor ou igual às estimativas resultantes nos demais métodos, conforme demonstrado acima.

4. Exemplo

Considere os dados multivariados apresentados na Tabela 1, oriundos de um experimento industrial realizado na Food Manufacturing Technology Facility, um laboratório financiado pelo Exército Americano e localizado na Rutgers University, USA. O objetivo do experimento é determinar as condições de autoclavagem adequadas para produção de cubos de carne acondicionados em *pouches* (embalagens com estrutura composta por lâminas de alumínio, filmes plásticos e resinas) para uso militar.

Três variáveis de autoclavagem (variáveis independentes) são consideradas (ver Tabela 1): *Tempo* – tempo de processamento em minutos, onde (–1) = 25min. e (+1) = 45min.; *Temp* – temperatura de processamento em °F; e *Tipo* – tipo de carne utilizada nos cubos, onde (-1) = natural e (+1) = moída e prensada. Três variáveis de desempenho (variáveis dependentes), avaliadas por oito especialistas em um painel de avaliação sensorial, são consideradas: Y_1 – Firmeza dos cubos de carne, Y_2 – Dureza da carne; e Y_3 – Desfibramento da carne. As avaliações forem realizadas seguindo o *Spectrum Method* (uma técnica de Análise Descritiva Quantitativa proposta por MEILGAARD *et al.*, 1991), e medidas usando uma escala contínua de 15 pontos. Cada avaliação foi replicada quatro vezes. Nesta análise, utilizam-se avaliações feitas por um dos painelistas.

As variáveis dependentes apresentam-se altamente correlacionadas, com valores de correlação amostral dados na Tabela 2. Num primeiro momento, determinam-se modelos RMQ para as variáveis dependentes. As variáveis independentes incluídas nos modelos apresentam um nível de significância ≥ 95%. O desempenho dos modelos em termos de ajuste aos dados foi monitorado pelo coeficiente de determinação R^2. Os modelos RMQ vêm apresentados a seguir (todos os modelos apresentam $R^2 \geq 0{,}90$):

$$\begin{aligned} Y_1 &= -5{,}4504 + 0{,}0493755\,Temp + \\ &\quad + 0{,}780138\,Tempo - 3{,}80832\,Tipo \end{aligned}$$

$$\begin{aligned} Y_2 &= -5{,}81922 + 0{,}0513834\,Temp + \\ &\quad + 0{,}924852\,Tempo - 3{,}77144\,Tipo \end{aligned}$$

$$Y_3 = 6{,}025 - 3{,}35625\,Tipo \qquad (14)$$

Os resíduos resultantes dos modelos em (14) são analisados quanto a homogeneidade das variâncias e correlação entre resíduos dentro de um mesmo vetor $\mathbf{u}_i$ (isto é, resíduos $u_{i1}, \ldots, u_{iT}$ em $\mathbf{u}_i$). Para tanto, (*i*) plotaram-se resíduos contra valores preditos pelos modelos (ver MONTGOMERY & PECK, 1992, p.74) e (*ii*) calculou-se a estatística de Durbin-Watson (ver DRAPER & SMITH, 1981, p.162). Os resultados obtidos em (*i*) para o primeiro modelo em (14) vêm apresentados na Figura 1. Os resíduos apresentam-se homogêneos quanto a variância. A estatística D-W resultou em um valor 2,28. Como o valor crítico d_U para este teste é 1,40, a hipótese de resíduos não correlacionados não pode ser rejeitada. Situação similar foi encontrada ao analisar-se resíduos resultantes dos demais modelos em (14).

Na seqüência, determinaram-se modelos SURE e RMV para as variáveis dependentes

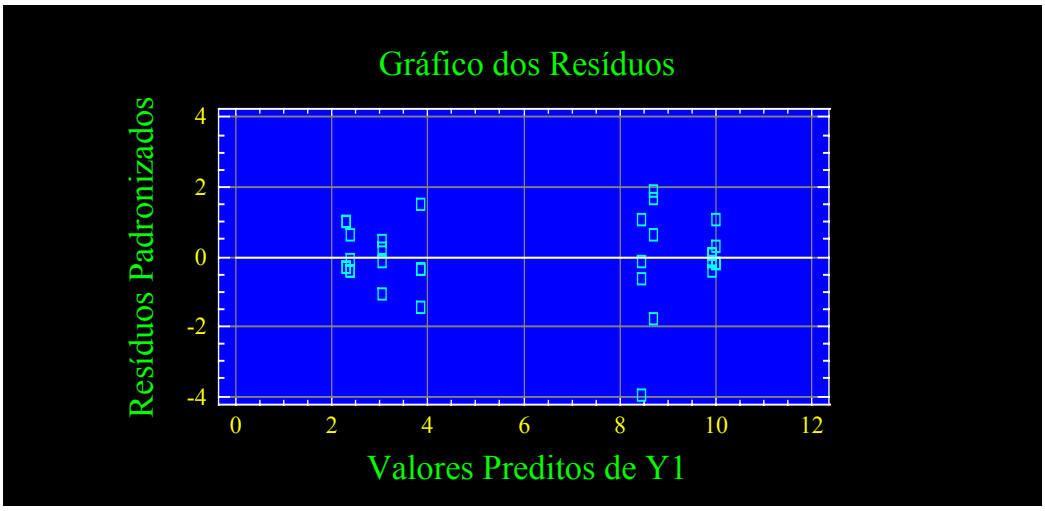

Figura 1 – Gráfico dos resíduos versus valores preditos de Y1.

(modelos RMG são idênticos aos modelos RMQ, dada a homogeneidade da variância dos resíduos em $\mathbf{u}_i$, i = 1,2,3). Os modelos SURE, com os mesmos regressores e coeficientes de determinação apresentados em (14), são:

$$Y_1 = 0{,}89792 + 0{,}0221544\,Temp + \\ + 0{,}313321\,Tempo - 3{,}45343\,Tipo$$

$$Y_2 = 1{,}64904 + 0{,}01936\,Temp + \\ + 0{,}375680\,Tempo - 3{,}353945\,Tipo$$

$$Y_3 = 6{,}025 - 3{,}35625\,Tipo \qquad (15)$$

Os modelos RMV, com $R^2 \geq 0{,}90$, são dados por:

$$Y_1 = -5{,}4504 + 0{,}0493755\,Temp + \\ + 0{,}780138\,Tempo - 3{,}80832\,Tipo$$

$$Y_2 = -5{,}81922 + 0{,}0513834\,Temp + \\ + 0{,}924852\,Tempo - 3{,}77144\,Tipo$$

$$Y_3 = -3{,}75242 + 0{,}041925\,Temp + \\ + 0{,}7189723\,Tempo - 3{,}90284\,Tipo$$

Os três grupos de modelos geram estimativas similares de $Y_i(\mathbf{x})$, i = 1,2,3, para um dado arranjo $\mathbf{x}$ das variáveis independentes. As variâncias são associadas a essas predições, porém, variam consideravelmente. Os valores de $V\left(\hat{Y}_1(\mathbf{x})\right)$, $V\left(\hat{Y}_2(\mathbf{x})\right)$ e $V\left(\hat{Y}_3(\mathbf{x})\right)$, calculados para arranjos $\mathbf{x}$ das variáveis independentes correspondendo às rodadas experimentais, estão apresentados na Tabela 3. A comparação entre variâncias de predição limita-se aos modelos obtidos via RMQ e SURE, já que estes correspondem aos valores extremos de variâncias (modelos RMV apresentam valores de $V\left(\hat{Y}_i(\mathbf{x})\right)$ intermediários).

Observe que $V\left(\hat{Y}_i(\mathbf{x})\right)_{\text{RMQ}} > V\left(\hat{Y}_i(\mathbf{x})\right)_{\text{SURE}}$, i = 1,2, em todas as rodadas. A alta correlação existente entre Y_1 e Y_2 torna a modelagem via RMQ inadequada, e isto pode ser constatado comparando as variâncias de predição obtidas em cada método. Por outro lado, $V\left(\hat{Y}_3(\mathbf{x})\right)_{\text{RMQ}} = V\left(\hat{Y}_3(\mathbf{x})\right)_{\text{SURE}}$ em todas as rodadas, apesar da existência de correlação entre Y_1 e Y_3, e Y_2 e Y_3. Isso se deve, em grande parte, ao pequeno número de termos na equação de regressão de Y_3. No caso desta variável dependente, a modelagem via RMQ ou SURE é igualmente satisfatória.

5. Conclusão

Neste trabalho, investiga-se a adequação de um grupo de técnicas de regressão linear na

Tabela 1 – Descrição dos níveis das variáveis independentes em cada rodada experimental e resultados da avaliação sensorial das rodadas.

Variáveis Independentes				Y_1	Y_2	Y_3
Rodada	Temp	Tipo	Tempo	Firmeza	Dureza	Desfibramento
1	225	-1	-1	7, 10,5, 9,3, 10,3	6,8, 10,8, 9,5, 10,5	6,8, 10,8, 9,3, 10,8
2	220	-1	1	9,8, 11, 9,8, 10,3	9,8, 11,5, 9,8, 11	9,8, 11,5, 9,8, 11
3	265	1	-1	3,3, 2,9, 2, 3,5	3,5, 3, 2, 3,5	3,5, 2,5, 3,3, 2,5
4	250	1	1	2,5, 5,3, 3,5, 3,5	2,5, 6, 3,8, 3,8	2,3, 3,5, 3, 3
5	220	-1	-1	7,8, 5,3, 9,5, 8,3	6,5, 5,5, 9,3, 8,5	6,8, 6, 9,5, 8,5
6	220	1	1	3, 2, 2, 2,3	3, 2,5, 2, 2,3	3, 2,5, 2, 2,3
7	250	1	-1	3,3, 3,3, 2, 2	3,3, 3,8, 2, 2	3,3, 2, 2, 2
8	250	-1	-1	10, 10, 9,5, 9,8	10, 10, 8,8, 9,6	10, 10, 9,5, 10

Tabela 2 – Matriz de covariâncias (e *correlações*) amostrais das variáveis dependentes.

Matriz de Covariâncias	Y_1	Y_2	Y_3
Y_1	1,083	1,206 (0,9597)	0,928 (0,7458)
Y_2		1,458	1,091 (0,7561)
Y_3			1,429

Tabela 3 – Valores preditos e variâncias de predição, usando modelagem RMQ e SURE.

Resultados para Modelos RMQ							**Resultados para Modelos SURE**						
Rod. (t)	$\hat{Y}_{1t}$	$\hat{Y}_{2t}$	$\hat{Y}_{3t}$	$V(\hat{Y}_{it})$	$V(\hat{Y}_{2t})$	$V(\hat{Y}_{3t})$	*Rod. (t)*	$\hat{Y}_{1t}$	$\hat{Y}_{2t}$	$\hat{Y}_{3t}$	$V(\hat{Y}_{it})$	$V(\hat{Y}_{2t})$	$V(\hat{Y}_{3t})$
1	8,69	8,59	9,38	0,09	0,36	0,09	1	9,02	8,98	9,38	0,08	0,11	0,09
2	10,0	10,2	9,38	0,16	0,44	0,09	2	9,54	9,64	9,38	0,11	0,14	0,09
3	3,05	3,10	2,67	0,13	0,50	0,09	3	3,00	3,05	2,67	0,10	0,13	0,09
4	3,87	4,18	2,67	0,14	0,48	0,09	4	3,30	3,51	2,67	0,10	0,14	0,09
5	8,44	8,33	9,38	0,12	0,38	0,09	5	8,91	8,89	9,38	0,09	0,12	0,09
6	2,38	2,64	2,67	0,18	0,46	0,09	6	2,63	2,93	2,67	0,12	0,15	0,09
7	2,31	2,33	2,67	0,11	0,44	0,09	7	2,67	2,76	2,67	0,09	0,12	0,09
8	9,92	9,87	9,38	0,14	0,48	0,09	8	9,58	9,47	9,38	0,10	0,14	0,09

modelagem de dados multivariados, na presença de correlação entre variáveis dependentes. Quatro técnicas de regressão são examinadas: regressão de mínimos quadrados ordinários (RMQ), regressão de mínimos quadrados generalizados (RMG), regressão por equações aparentemente não relacionadas (SURE – *seemingly unrelated equations regression*) e regressão multivariada (RMV). Cada técnica considera a estrutura de correlação das variáveis dependentes de maneira distinta na estimativa dos coeficientes a serem utilizados nos modelos

de regressão. Essas técnicas de regressão são comparadas tendo como base a variância de predições geradas a partir de seus modelos.

Demonstra-se analiticamente que, na presença de variáveis dependentes correlacionadas, a modelagem de dados pelos métodos RMQ e RMG resulta subótima em termos de variância de predição. Nesses casos, a regressão SURE deve ser a estratégia de modelagem utilizada, apresentando variância de predição pelo menos tão pequena quanto aquela resultante a partir dos demais métodos.

A comparação entre estratégias de modelagem é ilustrada por um estudo de caso. No estudo, três variáveis dependentes altamente correlacionadas são modeladas como função de três variáveis independentes. O exemplo ilustra a superioridade da regressão SURE para casos em que as variáveis dependentes apresentam-se correlacionadas.

Referências Bibliográficas

ADELMAN, I.; GREER, M. & MORRIS, C.T.: "Instruments and Goals in Economic Development". *American Economic Review*, **59**(2), 409-426, 1969.

AITKEN, A.C.: "On Least-Squares and Linear Combination of Observations". *Proceedings of the Royal Society of Edinburgh*, **55**, 42-48, 1935.

CHATTERJEE, S. & PRICE, B.: *Regression Analysis by Example*. 2nd Ed., John Wiley, New York, 1991.

DANIEL, C. & WOOD, F.S.: *Fitting Equations to Data – Computer Analysis of multifactor data*. John Wiley, New York, 1980.

DEATON, M.L.; REYNOLDS, Jr., M.R. & MYERS, R.H.: "Estimation and hypothesis testing in regression in the presence of non-homogeneous error variances". *Communications in Statistics*, **B12**(1), p.45-66, 1983.

DERRINGER, G. & SUICH, R.: "Simultaneous Optimization of Several Response Variables". *Journal of Quality Technology*, **12**(4), 214-219, 1980.

DRAPER, N. & SMITH, H.: *Applied Regression Analysis*. 2nd Ed. John Wiley, New York, 1981.

FOGLIATTO, F.S.; ALBIN, S.L. & TEPPER, B.J.: "A Hierarchical Approach to Optimizing Descriptive Analysis Multiresponse Experiments". *Journal of Sensory Studies* Vol.14(4), Oct-Dec 1999, *forthcoming*.

JOHNSON, R.A. & WICHERN, D.W.: *Applied Multivariate Statistical Analysis*. 3rd Ed., Prentice Hall, New Jersey, 1992.

MEILGAARD, M.; CIVILLE, G.V. & CARR, B.T.: *Sensory Evaluation Techniques*. Second Ed., CRC Press, Boca Raton, 1991.

MONTGOMERY, D.C. & PECK, E.A.: *Introduction to Linear Regression Analysis*. 2nd Ed., John Wiley, New York, 1992.

MOOD, A.M., GRAYBILL, F.A. & BOES, D.C.: *Introduction to the Theory of Statistics*. 3rd Ed., McGraw-Hill, New York, 1974.

MYERS, R.H.: *Classical and Modern Regression with Applications*. Duxbury Press, Boston, 1986.

RIBEIRO, J.L. & ELSAYED, E.A.: "A case Study on Process Optimization Using the Gradient Loss Function". *International Journal of Production Research*, **33**(12), 3233-3248, 1995.

SAS INSTITUTE: *SAS Version 6.0*. SAS Institute, Cary, North Carolina, 1990.

SEBER, G.A.F.: *Linear Regression Analysis*. John Wiley, New York, 1977.

SEBER, G.A.F.: *Multivariate Observations*. John Wiley, New York, 1984.

SRIVASTAVA, V.K. & GILES, D.E.A.: *Seemingly Unrelated Regression Equations Models – Estimation and Inference*. Marcel Dekker, New York, 1987.

STAPLETON, J.H.: *Linear Statistical Models*. John Wiley, New York, 1995.

STATGRAPHICS: *User's Manual Version 1.0*. Manugistics, 1995.

ZELLNER, A.: "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias". *Journal of the American Statistical Association*, 57, 348-368, 1962.

REGRESSION TECHNIQUES FOR MODELING MULTIVARIATE DATA UNDER CORRELATION

Abstract

Multivariate data often arise in empirical investigation. In Engineering studies, for example, multivariate data may be collected on the effect of different processing conditions on the characteristics of a machine output. Such data sets may present highly correlated variables. In this paper, we investigate the effect of correlation among dependent variables on their regression modeling. Four regression techniques are discussed and compared: ordinary least squares regression, generalized least squares regression, seemingly unrelated equations regression, and multivariate regression. Since regression models are most frequently used for prediction purposes, we compare modeling strategies using the prediction variance as a performance measure. The paper contains a case study from the food processing industry.

Key words: multivariate regression techniques, prediction variance, correlation.