UNIVERSIDADE DE SÃO PAULO
ESCOLA SUPERIOR DE AGRICULTURA “LUIZ DE QUEIROZ”
LCE 5801 REGRESSÃO E COVARIÂNCIA – 2016/02
PROF.ª TACIANA VILLELA SAVIAN

MULTICOLINEARIDADE

Mayara Salvian

Piracicaba - SP
Novembro, 2016

Sumário

1. Introdução

A multicolinearidade é definida como a presença de um alto grau de correlação entre as variáveis independentes (FREUND; WILSON; SA, 2006). As variáveis regressoras, localizadas nas colunas da matriz **X**, encontram-se em exata dependência linear, resultando na matriz singular **X'X**, por isso, a presença de dependência não linear têm impactos na estimação dos coeficientes de regressão (MONTGOMERY; PECK; VINING, 2006).

Supondo o seguinte modelo de regressão:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{1}$$

Em que: y é um vetor (n x 1) da variável resposta, X é uma matriz (n x p) das variáveis regressoras, β é um vetor de constantes desconhecidas e ε um vetor (n x 1) de erros aleatórios, cuja distribuição é $\varepsilon_i \sim N(0;\sigma^2)$.

Assumindo que as variáveis regressoras e a variável resposta são centradas, a matriz **X'X** é a matriz de correlação (p x p) entre as variáveis regressoras e **X'y** é o vetor de correlações (p x 1) entre as variáveis regressoras e a variável resposta. E considerando cada coluna da matriz **X** como sendo as variáveis regressoras, é possível definir formalmente multicolinearidade em termos de dependência linear nas colunas de **X**, quando os vetores $\mathbf{X}_1$, $\mathbf{X}_2$,..., $\mathbf{X}_p$ são linearmente dependentes se existe um conjunto de constantes c_1, c_2,...,c_p diferentes de zero, tal que (MONTGOMERY; PECK; VINING, 2006):

$$\sum_{j=1}^{p} c_j \boldsymbol{X}_j = 0 \tag{2}$$

Em outras palavras a multicolinearidade representa o mau condicionamento da matriz **X'X**, além disso, a menos que as variáveis sejam ortogonais, sempre haverá dependência entre as variáveis, especialmente em delineamentos experimentais.

Com base no seguinte modelo de regressão, é possível visualizar graficamente a presença ou não de multicolinearidade em um conjunto de dados (MONTGOMERY; PECK; VINING, 2006).

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \tag{3}$$

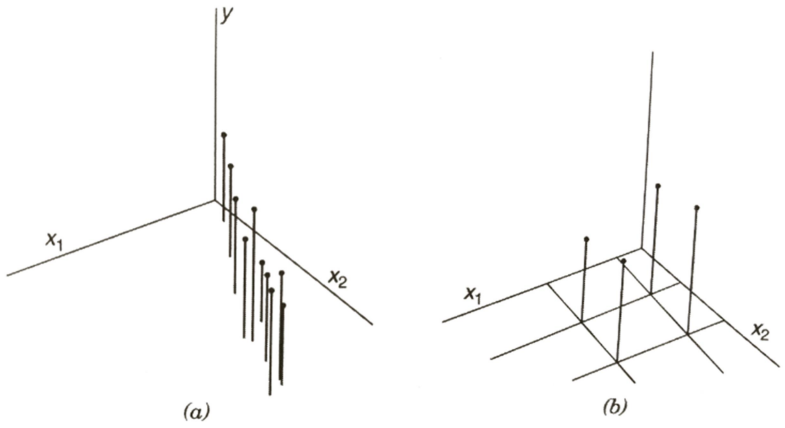

Figura 1. (a) Presença de multicolinearidade; (b) Ausência de multicolinearidade.

Na figura 1a, existe presença de colinearidade entre as variáveis X_1 e X_2, deste modo, qualquer plano situado ao longo do eixo de dispersão dos dados será instável e resulta na mesma soma de quadrados do erro. Por outro lado, na figura 1b, é possível observar a presença de ortogonalidade entre as variáveis X_1 e X_2 e a dispersão dos pontos fornece um plano de mínimos quadrados bem definido. Deste modo, os parâmetros serão estimados com precisão. Além disso, o intercepto do plano com o eixo y estima o β_0 e as inclinações nas direções de x_1 e x_2 estimam β_1 e β_2.

2. Fontes de Multicolinearidade:

Existem quatro fontes primárias de multicolinearidade (MONTGOMERY; PECK; VINING, 2006):

a. Método de coleta de dados utilizado: Quando o pesquisador coleta amostras somente no subespaço da região dos regressores definidos.

b. Restrições no modelo ou na população: Pode haver alguma razão para que haja a restrição do modelo. Por exemplo, um modelo para avaliar o consumo de energia elétrica em função da renda e do tamanho da casa, há uma restrição física, pois famílias com maiores casas têm maior renda.

c. Escolha do modelo: as variáveis escolhidas para compor o modelo podem ser linearmente dependentes, causando assim multicolinearidade.

d. Modelo com excesso de termos: O modelo possui mais variáveis regressoras do que observações. Estes modelos são comumente encontrados em pesquisas médicas. Neste caso é comum eliminar algumas variáveis regressoras para lidar com a multicolinearidade.

3. Efeitos da multicolinearidade:

Devido à dependência entre as variáveis regressoras, a multicolinearidade apresenta sérios efeitos nas estimativas de mínimos quadrados dos coeficientes de regressão (MONTGOMERY; PECK; VINING, 2006), uma vez que não é possível resolver o sistema de equações normais.

Supondo que existem duas variáveis regressoras x_1 e x_2, e assumindo o seguinte modelo:

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{4}$$

O sistema de equações normal dos mínimos quadrados é:

$$(\boldsymbol{X'X})\widehat{\boldsymbol{\beta}} = \boldsymbol{X'y} \tag{5}$$

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix} \tag{6}$$

Em que: r_{12} é a correlação simples entre x_1 e x_2 e r_{jy} é correlação simples entre x_j e y, j=1,2.

A matriz inversa de **(X'X)** é:

$$\boldsymbol{C} = (\boldsymbol{X'X})^{-1} = \begin{bmatrix} \frac{1}{1-r_{12}^2} & \frac{-r_{12}}{1-r_{12}^2} \\ \frac{-r_{12}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{bmatrix} \tag{7}$$

E as estimativas dos coeficientes de regressão são:

$$\hat{\beta}_1 = \frac{r_{1y} - r_{12} r_{2y}}{1 - r_{12}^2} \quad (8)$$

$$\hat{\beta}_2 = \frac{r_{2y} - r_{12} r_{1y}}{1 - r_{12}^2} \quad (9)$$

Portanto, se existe multicolinearidade entre x_1 e x_2, então o coeficiente de correlação r_{12} será alto, assim como as variâncias e as covariâncias para os estimadores dos mínimos quadrados dos coeficientes de regressão também serão altos. Isto implica que amostras diferentes tomadas nos mesmos níveis de x poderiam levar a estimativas amplamente diferentes dos parâmetros do modelo (MONTGOMERY; PECK; VINING, 2006).

4. Diagnosticando a multicolinearidade:

a. Matriz de correlação:

A matriz de correlação permite avaliar a existência de dependência linear entre par de variáveis. É possível então detectar a existência de multicolinearidade verificando se algum par apresenta correlação alta. Porém, quando mais de dois regressores estão envolvidos na dependência linear, a matriz de correlação não é eficiente. Os autovalores da matriz de correlação podem ser utilizados para diagnosticar a multicolinearidade, neste caso, um autovalor pequeno em relação aos demais indica um mau condicionamento da matriz (MONTGOMERY; PECK; VINING, 2006).

b. Fator de Inflação da Variância (VIF):

Supondo que as variáveis estão centradas e padronizadas, tem-se que $R = (X^T X)^{-1}$ em que os elementos da diagonal dessa matriz são chamados de fatores de inflação de variância (VIF) e representam o incremento da variância devido à presença de multicolinearidade (MONTGOMERY; PECK; VINING, 2006).

O VIF pode ser calculado pela seguinte equação:

$$VIF_j = \frac{1}{1 - R_j^2} \qquad j = 1, 2, \ldots, p \quad (10)$$

Em que: p é o número das variáveis preditoras; R_j^2 é o coeficiente de correlação múltipla, resultante da regressão de X_j nos outros *p-1* regressores.

A matriz de variâncias e covariâncias para as estimativas dos coeficientes de regressão padronizados é dada por:

$$V(\hat{\beta}) = \sigma^2(X^TX)^{-1} = \sigma^2 VIF = \sigma^2(1 - R_j^2)^{-1} \quad (11)$$

Se o valor de R_j^2 for próximo a um, isto significa que existe uma alta correlação entre a variável X_j e as demais variáveis, então $1 - R_j^2$ estará próximo de zero e consequentemente, o VIF assumirá um valor grande, apontando para o envolvimento dessa covariável em colinearidades. Um VIF máximo acima de 10 indica que a multicolinearidade pode estar influenciando as estimativas de mínimos quadrados.

c. Número de condição:

Fornece informações das potenciais dificuldades a serem encontradas em vários cálculos baseados na matriz **X'X**. Quanto maior o número de condição, maior será o mau condicionamento da matriz, e é calculado da seguinte maneira:

$$\kappa = \frac{\lambda_{máx}}{\lambda_{mín}} \quad (12)$$

Em que: $\lambda_1, \lambda_2, ..., \lambda_p$ são os autovalores da matriz **X'X**.

Geralmente valores de $k<100$ indicam que não há problemas sérios relacionados à multicolinearidade; $100<k<1000$ indicam problemas moderados com multicolinearidade; e $k>1000$ indicam fortes evidências de multicolinearidade.

5. Métodos que reparam a multicolinearidade

Diversas metodologias têm sido propostas para remediar o problema causado pela multicolinearidade. É possível transformar as variáveis explicatórias, coletar informações extras, excluir covariáveis do modelo ou utilizar estimadores viesados, como proposto na regressão de cumeeira (MONTGOMERY; PECK; VINING, 2006).

- **Regressão de Cumeeira:**

Quando o método dos mínimos quadrados é aplicado em dados não ortogonais, a estimação dos coeficientes de regressão obtidos é ruim, pois o $\hat{\beta}$ pode ser um estimador não viesado de β. Isto resulta em valores absolutos das estimativas muito grandes e instáveis (MONTGOMERY; PECK; VINING, 2006).

Proposta por Hoerl e Kennard (1970), a regressão de cumeeira tem o intuito de remediar os problemas de multicolinearidade, alterando o método dos mínimos para permitir que os estimadores dos coeficientes de regressão sejam viesados. Este método adiciona um coeficiente k a diagonal principal da matriz de correlações (**X'X**), resultando num decréscimo na estimativa das variâncias.

O erro quadrático médio do estimador $\hat{\beta}$ é igual à variância do estimador mais o viés ao quadrado, sendo escrito como:

$$EQM(\hat{\beta}) = E\left(\hat{\beta}^* - \beta\right)^2 = V\left(\hat{\beta}^*\right) + [E(\hat{\beta}^*) - \beta]^2 \qquad (13)$$

Note que se o estimador for não viesado, o erro médio quadrático é igual ao estimador da variância.

Deste modo, o estimador da regressão de cumeeira $\left(\hat{\beta}_R\right)$ pode ser obtido da seguinte forma:

$$\hat{\beta}_R = (\boldsymbol{X'X} + k\boldsymbol{I})^{-1}\boldsymbol{X'y} \qquad (14)$$

Em que: k é uma constante maior ou igual à zero. Quando k=0, o $\hat{\beta}_R$ é o estimador dos mínimos quadrados.

A matriz de variâncias e covariâncias é estimada pela seguinte fórmula:

$$Var\left(\hat{\beta}_R\right) = \sigma^2(\boldsymbol{X'X} + k\boldsymbol{I})^{-1}\boldsymbol{X'X}(\boldsymbol{X'X} + k\boldsymbol{I})^{-1} \qquad (15)$$

O erro quadrático médio do estimador da regressão de cumeeira $\left(\hat{\beta}_R\right)$ é igual à soma das variâncias dos estimadores dos parâmetros mais o quadrado do viés, sendo escrito como:

$$EQM(\hat{\beta}_R) = \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j+k)^2} + k^2\beta'(X'X + kI)^{-2}\beta \quad (16)$$

Em que: λ_1, λ_2,..., λ_p são os autovalores de X'X. Quando o valor de k é grande, o viés do estimador também será grande, mas a variância tende a ser pequena. Portanto, para que as estimativas do parâmetro $\hat{\beta}_R$ sejam estáveis, é necessário encontrar um valor de k pequeno (MONTGOMERY; PECK; VINING, 2006).

- **Escolha de *k*:**

Diversos autores propuseram diferentes formas de se obter o parâmetro viesado *k*. Mallows (1973) sugeriu selecionar o k com base na estatística C_p, enquanto que Marquardt (1970) propôs utilizar o valor de k como sendo o máximo valor do VIF entre 1 e 10, mas preferencialmente próximo de 1.

Hoerl, Kennard e Baldwin (1975), sugeriram o seguinte cálculo para a escolha de *k*:

$$k = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}} \quad (17)$$

Em que: $\hat{\beta}$ e $\hat{\sigma}^2$ são encontrados a partir da solução dos mínimos quadrados

6. Exemplo

Para exemplificar a presença de multicolinearidade, foi utilizado um banco de dados contendo informações de 192.145 animais da raça Montana Tropical, nascidos entre 1994 e 2008. A raça Montana tropical é uma população multirracial, os indivíduos são formados por raças agrupadas em quatro tipos biológicos, de acordo com o sistema NABC (FERRAZ et al., 1999, MOURÃO et al., 2007; PETRINI et al., 2012). O tipo biológico N inclui as raças *Bos indicus* (Gir, Guzerá, Indubrasil, Nelore, Tabapuã outras raças Zebu). O tipo biológico A inclui as raças *Bos taurus* adaptadas aos trópicos por seleção natural ou artificial (Bonsmara e Belmont Red). O tipo biológico B é formado por raças *Bos taurus* de origem britânica (Angus, Devon e Hereford). O tipo

biológico C possui raças *Bos taurus* da Europa continental (Charolês, Limousin e Simental).

Foi utilizado o seguinte modelo de regressão:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad (18)$$

Em que *y* é um vetor (n x 1) da variável resposta peso ao nascimento, *X* é uma matriz (n x p) das variáveis regressoras, β é um vetor de efeitos fixos e ε um vetor (n x 1) de erros aleatórios, cuja distribuição é $\varepsilon_i \sim N(0;\sigma^2)$. Incluídos no vetor β estão: os efeitos classificatórios de grupo de contemporâneos e classe de idade da mãe ao parto (cimp); efeitos genéticos aditivos diretos associados às composições raciais individuais (A, B e C); efeitos genéticos aditivos maternos (AM, BM e CM); as heterozigoses diretas (NxA, NxB, NxC, AxB, AxC e BxC); e a heterozigose materna total (HM).

- **Resultados:**

A variável resposta peso ao nascimento (PESNAS) possui 192.145 observações com média igual a 32,60 kg, desvio padrão de 4,42 kg, peso mínimo e peso máximo de 20 kg e 44 kg, respectivamente. Os VIFs foram calculados para verificar quais variáveis explicativas da matriz **X** apresentavam multicolinearidade (Tabela 1).

De acordo com a tabela 1, é possível observar que as variáveis explanatórias que possuem VIFs maiores que 10 são: os efeitos aditivos diretos associados aos tipos biológicos A (21,75254), B (15,32731) e C (39,67365); os efeitos aditivos maternos BM (21,51806) e CM (17,26589); e as heterozigoses diretas NxB (17,23765), NxC (20,71882) e AxC (16,2998), indiciando as covariáveis envolvidas em uma ou mais colinearidades.

Tabela 1. Fatores de Inflação da Variância (VIF) para as variáveis explicativas consideradas na matriz de delineamento **X** do modelo, para a variável resposta peso ao nascimento.

Variáveis	Estimativas dos parâmetros	VIF
A	0,00381	21,75254
B	-0,02336	15,32731
C	0,00273	39,67365
AM	0,03071	7,75293
BM	0,05834	21,51806
CM	0,03929	17,26589
NXA	0,00533	8,887
NXB	0,01387	17,23765
NXC	0,00321	20,71882
AXB	0,0036	6,85061
AXC	-0,00586	16,2998
BXC	0,00318	6,65902
HM	0,00262	2,95334
cimp	0,64302	1,13265

Outra maneira de verificar o problema de multicolinearidade é calculando os autovalores da matriz **X'X** (Tabela 2), por meio da fórmula do número de condição.

Tabela 2. Autovalores da matriz **X** do modelo para peso ao nascimento.

Variável	Autovalor	Variável	Autovalor
A	7,10081	**NXB**	0,17617
B	2,24082	**NXC**	0,11608
C	1,85282	**AXB**	0,04729
AM	1,18464	**AXC**	0,02152
BM	0,97581	**BXC**	0,01775
CM	0,87214	**HM**	0,01279
NXA	0,27363	**cimp**	0,00648

$$k = \frac{\lambda_{máx}}{\lambda_{mín}} = \frac{7{,}10081}{0{,}00648} = 1.095{,}80$$

O valor do k calculado é maior do que 1000 isto significa que há fortes evidências da presença de multicolinearidade.

Portanto, para lidar com o problema de multicolinearidade neste exemplo, realizou-se a regressão de cumeeira, obtendo novas estimativas para os coeficientes de regressão e de VIF, a partir do cálculo do parâmetro *k*. Com os resultados obtidos, é possível observar que houve uma melhora nas estimativas do VIF, reduzindo o efeito da multicolinearidade para as características estudadas (Tabela 3) e obtendo estimativas dos coeficientes de regressão mais precisas.

Tabela 3. Valores dos coeficientes de regressão e do VIF, estimados pelo método dos mínimos quadrados ($\hat{\beta}\ e\ VIF\_QM$), por regressão de cumeeira ($\hat{\beta}^*\ e\ VIF\_RC$) e valores de $k_{i.}$

Variável	$\widehat{\boldsymbol{\beta}}$	VIF_QM	$\widehat{\boldsymbol{\beta}}^*$	VIF_RC	$\boldsymbol{k_i}$
A	0,00381	21,75254	0,00955	9,99875	0,01511
B	-0,02336	15,32731	-0,00752	6,38165	0,01064
C	0,00273	39,67365	0,00918	6,36544	0,02755
AM	0,00533	8,88700	0,00060	5,22124	0,00617
BM	0,01387	17,23765	0,00446	7,09049	0,01197
CM	0,00321	20,71882	-0,00093	4,20952	0,01439
NXA	0,00360	6,85061	0,00718	5,20592	0,00476
NXB	-0,00586	16,29980	0,00026	6,63517	0,01132
NXC	0,00318	6,65902	0,00600	2,87569	0,00462
AXB	0,03071	7,75293	0,02243	3,68883	0,00538
AXC	0,05834	21,51806	0,04179	7,29267	0,01494
BXC	0,03930	17,26589	0,02470	7,79040	0,01199
HM	0,00262	2,95334	0,00345	2,77566	0,00205
cimp	0,64302	1,13265	0,64241	1,12946	0,00079

7. Conclusão

A presença de multicolinearidade resulta em estimativas duvidosas dos parâmetros de regressão, quando estimados pelo método dos mínimos quadrados. Existem diversas metodologias, disponíveis na literatura, para verificar se há ou não a presença de multicolinearidade, assim como, para corrigi-la, como o fator de Inflação da Variância (VIF) e a regressão de cumeeira, respectivamente.

Referências

FERRAZ, J. B. S.; ELER, J. P.; GOLDEN, B. L. Análise e genética do composto Montana Tropical. **Revista Brasileira de Reprodução Animal**, v.23, p.111-113, 1999.

FREUND, R. J.; WILSON, W. J.; SA, P. Regression analysis – Statistical Modeling of a response variable. **Elsevier**, Inc., San Diego, 459p, 2006.

HOERL, A. E.; KENNARD, R. W. Ridge regression: biased estimation for nonorthogonal problems. **Technometrics**, Alexandria, v. 12, n. 1, p. 55-67, 1970.

HOERL, A. E.; KENNARD, R. W.; BALDWIN, K. F. Ridge regression: Some simulations. **Communications in statistics**, v. 4, n. 1, p. 105-123, 1975.

MALLOWS, C. L. Some comments on C_p. **Technometrics**, v. 15, p. 661-675, 1973.

MARQUARDT, D. W. Generalized inverse, ridge regression, biased linear estimation, and nonlinear estimation. **Technometrics**, v. 12, p. 591-612, 1970.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. Introduction to linear regression analysis. **John, Wiley and Sons**, Inc., New York, 612p, 2006.

MOURÃO, G. B.; FERRAZ, J. B. S.; ELER, J. P.; BALIEIRO, J. C. C.; BUENO, R. S.; MATTOS, E. C.; FIGUEIREDO, L. G. G. Genetic parameters for growth traits of a Brazilian *Bos taurus* x *Bos indicus* beef composite. **Genetics and Molecular Research**, v.6, p.1190-1200, 2007.

PETRINI, J.; DIAS, R. A. P.; PERTILE, S. F. N.; ELER, J. P.; FERRAZ, J. B. S.; MOURÃO, G. B. Degree of multicollinearity and variables involved in linear dependence in additive-dominant models. **Pesquisa Agropecuária Brasileira**, Brasília, v. 47, n. 12, p. 1743-1750, 2012.

Anexo

Comandos utilizados no software SAS® para resolver o exemplo citado.

Amostra do banco de dados utilizado:

PESNAS	A	B	C	NXA	NXB	NXC	AXB	AXC	BXC	AM	BM	CM	HM	cimp
29	0	50	31,3	0	37,5	0	0	0	62,5	0	0	62,5	75	2
29	0	38	43,8	0	28,13	9,38	0	0	46,88	0	0	62,5	75	4
30	0	38	43,8	0	28,13	9,38	0	0	46,88	0	0	62,5	75	4
30	0	50	0	0	100	0	0	0	0	0	0	0	0	3
31	0	50	30	0	40	0	0	0	60	0	0	60	48	2
31	0	38	43,8	0	28,13	9,38	0	0	46,88	0	0	62,5	75	5
31	0	38	43,8	0	28,13	9,38	0	0	46,88	0	0	62,5	75	5
31	0	50	31,3	0	37,5	0	0	0	62,5	0	0	62,5	75	2
32	0	50	30	0	37,5	0	0	0	62,5	0	0	62,5	80	2
32	0	50	31,3	0	37,5	0	0	0	62,5	0	0	62,5	46,875	2
32	0	50	0	0	100	0	0	0	0	0	0	0	0	3
33	0	50	31,3	0	37,5	0	0	0	62,5	0	0	62,5	46,875	2
34	0	38	43,8	0	28,13	9,38	0	0	46,88	0	0	62,5	75	5
34	0	50	31,3	0	37,5	0	0	0	62,5	0	0	62,5	46,875	2
34	0	100	0	0	0	0	0	0	0	0	100	0	0	5

```
*Importando arquivo;
data a; set 'C:\Users\User\Regressão\sbz.sas7bdat'; run;

*Convertendo cimp em numérico;
data b; set a;
cim=input(cimp, 2.); *converte caracter em numérico;
drop cimp; *brinco;
rename cim=cimp;
run;

*Selecionando as variáveis do banco de dados;
data dados; set b;
keep produto pesnas cimp a b c am bm cm NxA NxB NxC AxB AxC BxC HM;
if pesnas = . then delete;
run;
```

```
*Varificando a presença de multicolinearidade;
proc reg data=dados;
model pesnas = a b c am bm cm NxA NxB NxC AxB AxC BxC HM cimp
/corrb vif collin;
run;

*Ridgde Regression;
*outest=cria um arquivo com as estimativas do modelo;
*outvif=cria uma coluna de vif no arquivo criado pelo outest;
proc reg data=dados outest=ridge outvif outstb ridge=0.01 to 0.05 by .001; *k começa
em 0.01 e vai até 0.05 com incremento de 0.001;
model pesnas = pesnas = a b c am bm cm NxA NxB NxC AxB AxC BxC HM /vif;
plot / ridgeplot; data new; set ridge;
if _type_='RIDGESTB' or _type_='RIDGEVIF';
proc sort; by _type_;
proc gplot; by _type_;
plot (a b c am bm cm nxa nxb nxc axb axc bxc pesnas)*_RIDGE_/overlay vref= 10
cvref=blue lvref=1 ; run;

*Obtendo os estimadores;
proc iml;
use dados; read all into dados;
* Matriz X e vetor Y;
X = dados[,3:16];
y = dados[,2];
produto=dados[,1];
n=nrow(X);
meanX=X[+,]/n; * Cria um vetor linha com as médias para cada coluna;
Xc=X-repeat(meanX,n,1); * Centralizar X;
cssX=Xc[##,];* Soma dos quadrados dos elementos de cada coluna;
sX=Xc*diag(1/(sqrt(cssX))); * Padronizar Xc;
XX=T(sX)*sX;
meany = y[+,1]/n;
yc=y-repeat(meany,n,1);
```

```
Xy = t(sX)*yc;

*Obtendo soluções por QM;
betar=inv(XX)*Xy;
beta_QM=diag(1/(sqrt(cssX)))*betar;
FIV_QM=vecdiag(inv(XX));
mFIV_QM=FIV_QM/max(FIV_QM);
sFIV_QM=sum(FIV_QM)/ncol(XX);
maxFIV_QM=max(FIV_QM);
print beta_QM FIV_QM;
print sFIV_QM maxFIV_QM;
*finish reg;

* Obtendo soluções por RR;
do teta=.00001 to 1 by .00001;
K=diag(teta*mFIV_QM);
XXIk=XX+K;
iXXIk=inv(XXIk);
H=iXXIK*XX;
H2=H##2;
iBIASp=(1-((sqrt(sum(H2))/sqrt(ncol(XX)))))*100;
FIV_RC=vecdiag(iXXIk*XX*iXXIk);
sFIV_RC=(sum(FIV_RC))/ncol(XX);
maxFIV_RC=max(FIV_RC);
if maxFIV_RC<=10 then do;
solr=iXXIk*Xy;
betaRR=diag(1/(sqrt(cssX)))*solr;
vK=vecdiag(K);
print iBIASp;
print betaRR FIV_RC vK;
print sFIV_RC maxFIV_RC ;
teta=1;
end;
end;
```

quit;