Solução para multicolinearidade: Regressão do Componente Principal

Rodrigo H. Ozon

02/09/2020

Resumo

Este tutorial replica o conteúdo presente em Maddala, p. 150-153, 2001, com algumas adaptações. *Algumas estimativas podem ser diferentes provavelmente pelas omissões de detalhes a este respeito do próprio autor ou pelos métodos empregados pelo uso de diferentes softwares.

Regressão do Componente Principal

Outra solução frequentemente sugerida para o problema de multicolinearidade é a *regressão do componente principal,* que procede da seguinte forma. Suponha que tenhamos k variáveis explicativas. Então podemos considerar funções lineares dessas variáveis.

$$z_1 = a_1x_1 + a_2x_2 + \ldots + a_kx_k$$

$$z_2 = b_1x_1 + b_2x_2 + \ldots + b_kx_k \quad \text{etc.}$$

Suponha que escolhemos os $a^{'}s$ de forma que variância de z_1 seja maximizada sujeita à condição de que

$$a_1^2 + a_2^2 + \ldots + a_k^2 = 1$$

Isso é chamado condição de normalização. (Isso é necessário pois caso contrário, a variância de z_1 pode aumentar indefinidamente.) z_1 é, então, chamado de primeiro componente principal. É a função linear dos $x^{'}s$ que tem maior variância (sujeito à regra de normalização).

Discutiremos os principais destaques e usos desse método, os quais são fáceis de se compreender sem o uso de álgebra matricial. Além disso, para se usar o método existem programas de computador disponíveis que fornecem os componentes principais ($z^{'}s$), dado qualquer conjunto de variáveis $x_1, x_2, \ldots, x_k$.

Esse processo de maximização da variância da função linear z sujeito à condição de que a soma dos quadrados dos coeficientes dos $x^{'}s$ é igual a 1 produz k soluções. Correspondendo a isso, construímos k funções lineares $z_1, z_2, \ldots, z_2$, denominadas de componentes principais dos $x^{'}s$. Elas podem ser ordenadas de forma que

$$var(z_1) > var(z_2) > \ldots > var(z_k)$$

z_1, que tem a maior variância, é chamada de primeiro componente principal, z_2, que tem a segunda maior variância, é chamada de segundo componente principal e assim por diante. Esses componentes principais têm as seguintes propriedades:

1. $var(z_1) + var(z_2) + \ldots + var(z_k) = var(x_1) + var(x_2) + \ldots + var(x_k)$
2. Diferentemente dos $x^{'}s$, que são correlacionados, os $z^{'}s$ são ortogonais e não-correlacionados. Logo, existe zero multicolinearidade entre os $z^{'}s$.

Por vezes é sugerido que em vez de se regredir y em $x_1, x_2, \ldots, x_k$, devemos regredir y em $z_1, z_2, \ldots, z_k$. Mas isso não é um problema para solução da multicolinearidade. Se regredirmos y nos $z^{'}s$ e, então, substituimos os valores dos $z^{'}s$ nos termos dos $x^{'}s$, acharemos finalmente as mesmas respostas que antes. O fato de os $z^{'}s$ serem não-correlacionados não significa que acharemos melhores estimativas dos coeficientes na equação de regressão original. Dessa forma, faz sentido usar os componentes principais apenas se regredirmos y em um subconjunto dos $z^{'}s$. Mas esse procedimento também tem problemas. São eles:

1. O primeiro componente principal, z_1, embora tenha a maior variância, não precisa ser o mais correlacionado com Y. Na verdade, não há relação necessária entre a ordem dos componentes principais e o grau de correlação com a variável dependente Y.
2. Pode-se pensar em se escolher apenas aqueles componentes principais que têm correlação elevada com Y e em descartar o restante, mas o mesmo procedimento pode ser usado com o conjunto de variáveis original $x_1, x_2, \ldots, x_k$ escolhendo primeiro a variável com a maior correlação com Y, depois a com maior correlação parcial e assim por diante. Isso é o que os "programas de regressão" passo a passo fazem.

Por vezes é sugerido que em vez de se regredir y em $x_1, x_2, \ldots, x_k$, devemos regredir y em $z_1, z_2, \ldots, z_k$. Mas isso não é uma solução para o problema da multicolinearidade. Se regredirmos y nos $z^{'}$ nos termos dos $x^{'}$, acharemos finalmente as mesmas respostas de antes. O fato de os $z^{'}s$ serem não-correlacionados não singifica que acharemos melhores estimativas dos coeficientes na equação de regressão original. Dessa forma, faz sentido usar os componentes principais *apenas* se regredirmos y em um subconjunto dos $z^{'}s$. Mas esse procedimento também tem problemas. São eles:

1. O primeiro componente principal z_1, embora tenha a maior variância, não precisa ser o mais correlacionado com y. Na verdade, não há relação necessária entre a ordem dos componentes principais e o grau de correlação com a variável dependente y.
2. Pode-se pensar em se escolher apenas aqueles componentes principais que têm correlação elevada com y e em se descartar o restante, mas o mesmo procedimento pode ser usado com o conjunto de variáveis original $x_1, x_2, \ldots, x_k$ escolhendo primeiro a variável com maior correlação com y, depois a com maior correlação parcial e assim por diante. Isso é o que os "programas de regressão" passo a passo fazem.
3. As combinações lineares dos $z^{'}s$ com frequência não tem sentido econômico. O que significa, por exemplo, 2(renda) +3(preço) ? Essa é uma das falhas mais importantes do método.
4. Alterar as unidades de medição dos $x^{'}s$ transformará os componentes principais. Esse problema pode ser evitado se todas as variáveis forem padronizadas para terem variância unitária.

No entanto, o método do componente principal tem alguma utilidade nos estágios explicativos da investigação. Suponha, por exemplo, que haja muitas taxas de juros no modelo (como todas são medidas nas mesmas unidades, não há problema de escolha de unidade de medida). Se a análise do componente principal mostrar dois componentes principais respondem por 99% da variação das taxas de juros e se, olhando para os coeficientes, conseguirmos identificá-lo como componentes de curto prazo e de longo prazo poderemos argumentar que existem apenas duas variáveis "latentes" que respondem por todas as variações das taxas de juros. Portanto, o método do componente principal nos oferecerá alguma direção para a pergunta: "quantas fontes independentes de variação existem ?" Além disso, se pudermos dar uma interpretação econômica aos componentes principais, isso será útil.

Ilustramos o método com referência a um conjunto de dados de Mallinvaud (E. Mallinvaud, *Statistical Methods of Econometrics*, 2a. ed., Amsterdã: North-Holland, 1970, p. 19). Escolhemos esses dados porque eles foram usados por Chatterjee Price (*in Regression Analysis,* p. 161) para ilustrar o método do componente principal.

Você pode baixar a planilha Excel com os dados aqui.

Primeiro estimamos uma função de demanda por importação. Ao regredir y em x_1, x_2, x_3 acham-se os seguintes resultados:

	A	B	C	D	E	F	G	H	I	J
1		RESUMO DOS RESULTADOS								
2										
3		*Estatística de regressão*								
4		R múltiplo	0.98642945							
5		R-Quadrado	0.973043059							
6		R-quadrado ajustado	0.967266572							
7		Erro padrão	2.258165584							
8		Observações	18							
9										
10		ANOVA								
11			*gl*	*SQ*	*MQ*	*F*	*F de significação*			
12		Regressão	3	2576.920746	858.9736	168.4489	3.211656E-11			
13		Resíduo	14	71.39036528	5.099312					
14		Total	17	2648.311111						
15										
16			*Coeficientes*	*Erro padrão*	*Stat t*	*valor-P*	*95% inferiores*	*95% superiores*	*Inferior 95,0%*	*Superior 95,0%*
17		Interseção	-19.72510712	4.125252548	-4.78155	0.000293	-28.57289387	-10.87732037	-28.57289387	-10.87732037
18		Produto Doméstico Bruto, X1	0.032204469	0.186884316	0.172323	0.86565	-0.368622524	0.433031462	-0.368622524	0.433031462
19		Formação de estoque, X2	0.414199097	0.322259758	1.285296	0.219545	-0.276979342	1.105377536	-0.276979342	1.105377536
20		Consumo, X3	0.242747006	0.285360659	0.850667	0.409268	-0.369290735	0.854784748	-0.369290735	0.854784748
21										
22										

Dados | Regressao todo periodo | Regr 1949-1959

Figure 1: *Resultados da Regressão com a série completa*

O R^2 é muito elevado e a razão F é altamente significativa, mas todas as razões t individuais são insignificantes. Isso é evidência do problema da multicolinearidade. Chatterjee e Price argumentam que antes que qualquer análise adicional seja feita, devemos olhar para os resíduos dessa equação. Eles acham (veja o gráfico de resíduos) um padrão definido – os resíduos declinam até 1960 e então aumentam.

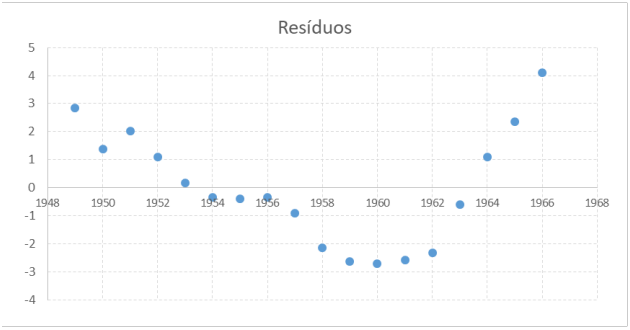


Figure 2: *Padrão de resíduos da regressão com a série completa*

Chaterjee e Price argumentam que a dificuldade desse modelo é que o Mercado Comum Europeu começou a operar em 1960, causando mudanças nas relações entre importação e exportação. Portanto, eles excluem os anos posteriores a 1959 e consideram apenas os 11 anos do período 1949–1959. Agora os resultados da regressão são os seguintes:

Como podemos ver no gráfico de resíduos a seguir, não encontramos um padrão sistemático

de forma que podemos prosseguir.

Embora o R^2 seja muito elevado, o coeficiente de x_1 não é significante ($t = -0,731305765$ e valor$-p = 0,488344308$). Existe, por conseguinte, um problema de multicolinearidade.

Para ver o que deve ser feito sobre isso, primeiro olhamos para as correlações simple entre as variáveis explicativas:

Suspeitamos que a alta correlação presente nos dados de Consumo ($X3$) com Produto Doméstico Bruto ($X1$), possa ser a causa do problema.

RESUMO DOS RESULTADOS									
Estatística de regressão									
R múltiplo	0.99594								
R-Quadrado	0.991897								
R-quadrado ajustado	0.988424								
Erro padrão	0.488869								
Observações	11								
ANOVA									
	gl	*SQ*	*MQ*	*F*	*e significação*				
Regressão	3	204.7761	68.25871	285.6099	1.11E-07				
Resíduo	7	1.672949	0.238993						
Total	10	206.4491							
	Coeficiente	*Erro padrão*	*Stat t*	*valor-P*	*% inferior*	*% superior*	*ferior 95,0*	*perior 95,0%*	
Interseção	-10.128	1.21216	-8.35532	6.9E-05	-12.9943	-7.26169	-12.9943	-7.26169	
Produto Doméstico Bru	-0.0514	0.07028	-0.73131	0.488344	-0.21758	0.11479	-0.21758	0.11479	
Formação de estoque,	0.586949	0.094618	6.203327	0.000444	0.363212	0.810686	0.363212	0.810686	
Consumo, X3	0.286849	0.102208	2.806516	0.026277	0.045165	0.528532	0.045165	0.528532	

Dados | Regressao todo periodo | **Regr 1949-1959**

Figure 3: *Regressão para 1949-1959*

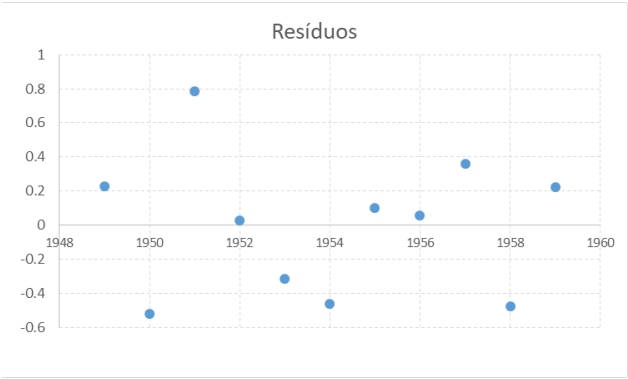


Figure 4: *Padrão residual da Regressão para 1949-1959*

Matriz de correlação das explicativas				
		Produto Doméstico Bruto, X1	Formação de estoque, X2	Consumo, X3
	Produto Doméstico Bru	1	0.025850673	0.997260693
	Formação de estoque,	0.025850673	1	0.035673223
	Consumo, X3	0.997260693	0.035673223	1

Figure 5: *Matriz de correlação das explicativas para a Regressão para 1949-1959*

A análise de componentes principais ajuda ? Vamos para o primeiro passo, rumando para a obtenção dos valores dos componentes principais no pacote estatístico R:

```
library(readxl)

url<-"https://github.com/rhozon/Introdu-o-Econometria-com-Excel/blob/master/Maddala,%20p.%20151.xlsx?ra
dados <- tempfile()
download.file(url, dados, mode="wb")
dados<-read_excel(path = dados, sheet = 1)

dados
```

```
## # A tibble: 11 x 9
##      Ano     Y    X1    X2    X3  Ypadr X1padr X2padr  X3padr
##    <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>  <dbl>  <dbl>   <dbl>
##  1  1949  15.9  149.   4.2  108. -1.32  -1.51   0.546 -1.53
##  2  1950  16.4  161.   4.1  115. -1.21  -1.11   0.485 -1.21
##  3  1951  19    172.   3.1  123. -0.636 -0.770 -0.121 -0.801
##  4  1952  19.1  176.   3.1  127. -0.614 -0.636 -0.121 -0.622
##  5  1953  18.8  181.   1.1  132. -0.680 -0.460 -1.33  -0.370
##  6  1954  20.4  191.   2.2  138. -0.328 -0.130 -0.667 -0.0987
##  7  1955  22.7  202.   2.1  146   0.178  0.250 -0.728  0.304
##  8  1956  26.5  212.   5.6  154.  1.01   0.594  1.39   0.696
##  9  1957  28.1  226.   5    162.  1.37   1.05   1.03   1.09
## 10  1958  27.6  232.   5.1  164.  1.26   1.24   1.09   1.19
## 11  1959  26.3  239    0.7  168.  0.970  1.48  -1.58   1.35
```

Então seleciono somente as variáveis explicativas e em seguida calculo as variâncias de cada uma das três variáveis

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
explicativas<-dados%>%
  select(X1,X2,X3)

vars <- apply(explicativas, 2, var)
vars
```

```
##       X1       X2       X3
## 899.9709   2.7200 425.7785
```

A maior fração da variância explicada entre essas variáveis é 70%, e a menor é quase 0% (X2). Também podemos calcular essas frações para subconjuntos de variáveis. Por exemplo, as variáveis 1 e 3 juntas explicam 99,94% da variância total, e as variáveis 1 e 2 explicam 70,02%.

A análise de componentes principais calcula um novo conjunto de variáveis ("componentes principais") e expressa os dados em termos dessas novas variáveis. Consideradas em conjunto, as novas variáveis representam

a mesma quantidade de informação que as variáveis originais, no sentido de que podemos restaurar o conjunto de dados original do transformado.

Além disso, a variância total permanece a mesma. No entanto, é redistribuído entre as novas variáveis da forma mais "desigual": a primeira variável não apenas explica a maior variância entre as novas variáveis, mas também a maior variância que uma única variável pode possivelmente explicar.

```
vars/sum(vars)
```

```
##          X1          X2          X3
## 0.677449456 0.002047469 0.320503075
```

De forma mais geral, os primeiros componentes principais (onde k pode ser 1, 2, 3 etc.) explicam a maior variância que qualquer k variável pode explicar, e as últimas k variáveis explicam a menor variância que qualquer k variável pode explicar, sob algumas restrições gerais. (As restrições garantem, por exemplo, que não podemos ajustar a variância explicada de uma variável simplesmente escalando-a.)

```
pca <- prcomp(explicativas, retx=T)
expl_transformada <- pca$x
expl_transformada
```

```
##              PC1        PC2        PC3
##  [1,] -55.243186  0.8526477 -0.6319214
##  [2,] -41.641198  0.4642629 -1.7916913
##  [3,] -28.396312 -0.2823642 -0.5011266
##  [4,] -23.004179 -0.1131437  0.2645521
##  [5,] -15.693755 -1.7829447  1.9673323
##  [6,]  -4.361502 -0.9482986  0.7576149
##  [7,]   9.734530 -0.9774781  1.1588999
##  [8,]  22.815447  2.6055542  1.1976004
##  [9,]  38.749791  1.7756696  0.3621647
## [10,]  44.662808  1.4979238 -1.2531020
## [11,]  52.377555 -3.0918290 -1.5303231
```

Estas são as variâncias de amostra das novas variáveis.

```
vars_transformadas <- apply(expl_transformada, 2, var)
# ou: pca$sdev^2
vars_transformadas
```

```
##         PC1         PC2         PC3
## 1324.168516    2.781380    1.519559
```

Observe que sua soma, a variância total, é a mesma das variáveis originais: 2,99.

E essas são as mesmas variâncias divididas pela variância total, ou seja, quanto da variância total cada nova variável explica:

```
vars_transformadas/sum(vars_transformadas)
```

```
##         PC1         PC2         PC3
## 0.996762486 0.002093673 0.001143842
```

os componentes principais (obtidos pelo pacote Eviews) são:

Obtemos assim os mesmos valores encontrados por Maddala, p. 152, 2001:

$$z_1 = 0,706330X_1 + 0,043501X_2 + 0,706544X_3$$
$$z_2 = -0,035689X_1 + 0,999029X_2 - 0,025830X_3$$

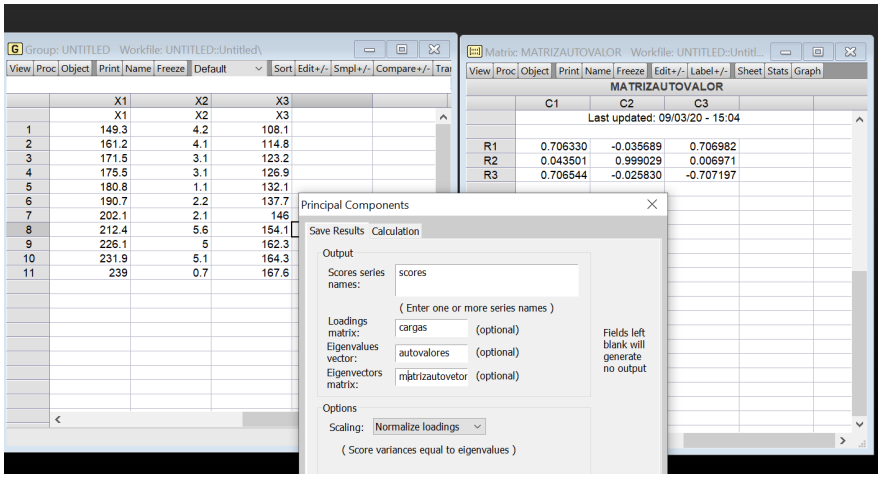


Figure 6: *Matriz de Autovalores para os dados 1949-1959*

$$z_3 = 0,706982X_1 + 0,006971X_2 - 0,707197X_3$$

onde X_1, X_2, X_3 são os valores normalizados de x_1, x_2, x_2. Isto é, $X_1 = (x_{1i} - \overline{x_1})/\sigma(x_1)$ e $X_2 = (x_{2i} - \overline{x_2})/\sigma(x_2)$ e $X_3 = (x_{3i} - \overline{x_3})/\sigma(x_3)$ onde $\overline{x_1}, \overline{x_2}$ *e* $\overline{x_3}$ são as médias aritméticas de x_1, x_2, x_3 respectivamente. Logo

$$var(X_1) = var(X_2) = var(X_3) = 1$$

As variâncias dos componentes principais são

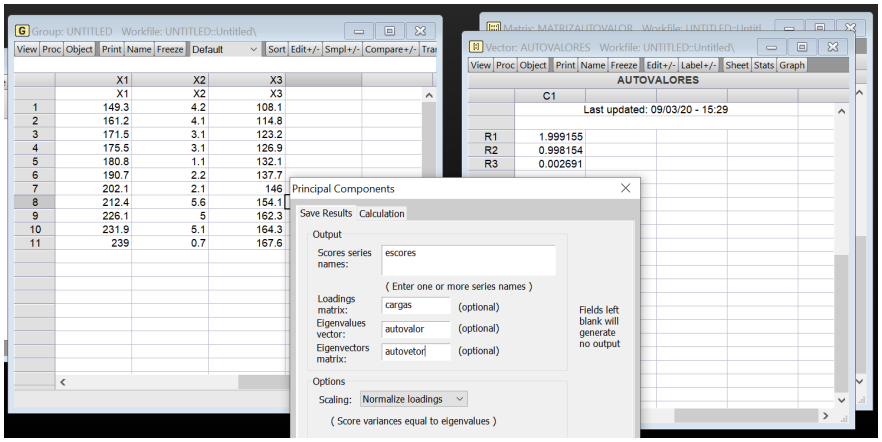


Figure 7: *Matriz de Autovetores para os dados 1949-1959*

então

$$var(z_1) = 1,99915493449973 \quad var(z_2) = 0,9981541760451606 \quad var(z_3) = 0,002690889455111145$$

Note que $\sum var(z_i) = \sum var(X_i) = 3$.O fato de que $var(z_3) = 0$ identifica aquela função linear como causa da multicolinearidade. Nesse exemplo, há apenas uma função linear desse tipo. Em alguns exemplos pode haver mais. Como $E(X_1) = E(X_2) = E(X_3) = 0$ por causa da normalização, os $z^{'}s$ têm média zero. Assim, z_3 tem média zero e sua variância também está perto de zero. Portanto, podemos afirmar que $z \simeq 0$. Olhando para os coeficientes $X^{'}s$, podemos dizer que (ignorando os coeficientes muito pequenos)

$$z_1 \simeq 0,7063(X_1 + X_3)$$

$$z_2 \simeq X_2$$

$$z_3 \simeq 0,707(X_3 - X_1)$$

$$z_3 \simeq 0 \quad \text{logo} \quad X_1 \simeq X_3$$

Na verdade, teríamos achado os mesmos resultados da regressão de X_3 em X_1. O coeficiente de regressão é $r_{13} = 0,99726$. (Observe que X_1 e X_3 estão em forma padronizada e, por consequência, o coeficiente de regressão é r_{13}.)

Em termos de variáveis originais (não normalizadas), a regressão de x_3 em x_1 é: (erro padrão entre parêntesis)

$$x_3 = 6,258606642 + 0,685940354x_1 \quad R^2 = 0,99452889 \quad (0,0169588)$$

	A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS								
2									
3	*Estatística de regressão*								
4	R múltiplo	0.997261							
5	R-Quadrado	0.994529							
6	R-quadrado ajust	0.993921							
7	Erro padrão	1.608823							
8	Observações	11							
9									
10	ANOVA								
11		*gl*	*SQ*	*MQ*	*F*	*e significação*			
12	Regressão	1	4234.491	4234.491	1636.004	1.72E-11			
13	Resíduo	9	23.29481	2.588313					
14	Total	10	4257.785						
15									
16		*Coeficientes*	*rro padrão*	*Stat t*	*valor-P*	*% inferior*	*% superior*	*ferior 95.0*	*perior 95.0%*
17	Interseção	6.258607	3.335482	1.876373	0.093345	-1.28678	13.80399	-1.28678	13.80399
18	X1	0.68594	0.016959	40.44755	1.72E-11	0.647577	0.724304	0.647577	0.724304
19									
20									
21									
22	RESULTADOS DE RESÍDUOS								

regr x3 em x1 | Dados pro R pca | Dados para R | Dados | Regressao todo periodo | Rec

Figure 8: *Regressão de X_3 em X_1 para os dados 1949-1959*

Não obtivemos, nesse exemplo, mais informação a partir da análise do componente principal do que do estudo da correlação simples. De qualquer forma, qual é a solução agora ? Dado que existe uma relação quase exata entre x_3 em x_1, não podemos esperar estimar os coeficientes de x_1 e x_3 separadamente. Se a equação original for

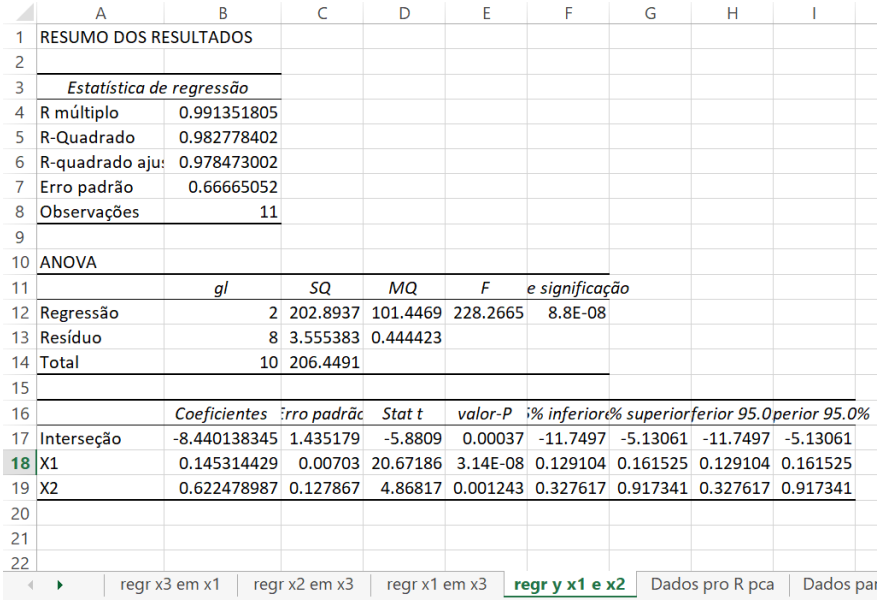
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

então, substituindo por x_3 os termos de x_1, temos

$$y = (\beta_0 + 6,258\beta_3) + (\beta_1 + 0,686\beta_3)x_1 + \beta_3 x_3 + u$$

Isso fornece as funções lineares dos $\beta^{'}s$ que são estimáveis. São eles $(\beta_0 + 6,258\beta_3)$,$(\beta_1 + 0,686\beta_3)$ e β_2. Da regressão de y em x_1 e x_2, acham-se os seguintes resultados:

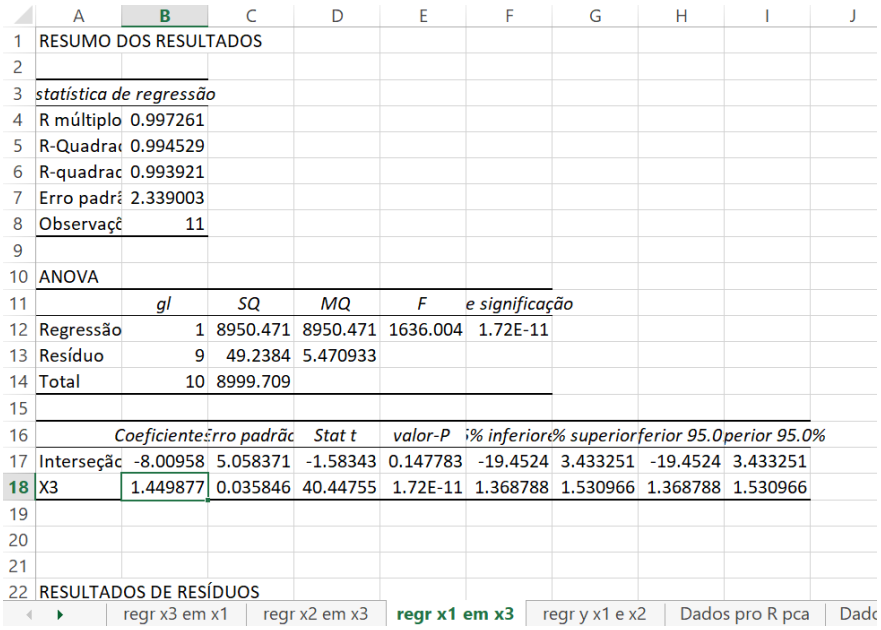
Obviamente, podemos estimar uma regressão de x_1 e x_3.



	A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS								
2									
3	*Estatística de regressão*								
4	R múltiplo	0.991351805							
5	R-Quadrado	0.982778402							
6	R-quadrado ajus	0.978473002							
7	Erro padrão	0.66665052							
8	Observações	11							
9									
10	ANOVA								
11		*gl*	*SQ*	*MQ*	*F*	*e significação*			
12	Regressão	2	202.8937	101.4469	228.2665	8.8E-08			
13	Resíduo	8	3.555383	0.444423					
14	Total	10	206.4491						
15									
16		*Coeficientes*	*rro padrãc*	*Stat t*	*valor-P*	*% inferior*	*% superior*	*ferior 95.0*	*perior 95.0%*
17	Interseção	-8.440138345	1.435179	-5.8809	0.00037	-11.7497	-5.13061	-11.7497	-5.13061
18	X1	0.145314429	0.00703	20.67186	3.14E-08	0.129104	0.161525	0.129104	0.161525
19	X2	0.622478987	0.127867	4.86817	0.001243	0.327617	0.917341	0.327617	0.917341
20									
21									
22									

regr x3 em x1 | regr x2 em x3 | regr x1 em x3 | **regr y x1 e x2** | Dados pro R pca | Dados par

Figure 9: *Regressão de y em X_1 e X_2 para os dados 1949-1959*



	A	B	C	D	E	F	G	H	I	J
1	RESUMO DOS RESULTADOS									
2										
3	*statística de regressão*									
4	R múltiplo	0.997261								
5	R-Quadrac	0.994529								
6	R-quadrac	0.993921								
7	Erro padrá	2.339003								
8	Observaçõ	11								
9										
10	ANOVA									
11		*gl*	*SQ*	*MQ*	*F*	*e significação*				
12	Regressão	1	8950.471	8950.471	1636.004	1.72E-11				
13	Resíduo	9	49.2384	5.470933						
14	Total	10	8999.709							
15										
16		*Coeficientes*	*rro padrãc*	*Stat t*	*valor-P*	*% inferior*	*% superior*	*ferior 95.0*	*perior 95.0%*	
17	Interseçã	-8.00958	5.058371	-1.58343	0.147783	-19.4524	3.433251	-19.4524	3.433251	
18	X3	1.449877	0.035846	40.44755	1.72E-11	1.368788	1.530966	1.368788	1.530966	
19										
20										
21										
22	RESULTADOS DE RESÍDUOS									

regr x3 em x1 | regr x2 em x3 | **regr x1 em x3** | regr y x1 e x2 | Dados pro R pca | Dadc

Figure 10: *Regressão de X_1 em X_3 para os dados 1949-1959*

	A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS								
2									
3	*statística de regressão*								
4	R múltiplo	0.995629							
5	R-Quadrac	0.991277							
6	R-quadrac	0.989097							
7	Erro padrã	0.474442							
8	Observaçõ	11							
9									
10	ANOVA								
11		*gl*	*SQ*	*MQ*	*F*	*e significação*			
12	Regressão	2	204.6483	102.3242	454.5809	5.79E-09			
13	Resíduo	8	1.800765	0.225096					
14	Total	10	206.4491						
15									
16		*Coeficientes*	*Erro padrão*	*Stat t*	*valor-P*	*% inferior*	*% superior*	*ferior 95.0*	*perior 95.0%*
17	Interseção	-9.74274	1.059489	-9.1957	1.58E-05	-12.1859	-7.29956	-12.1859	-7.29956
18	X2	0.596052	0.091028	6.548002	0.000179	0.386141	0.805963	0.386141	0.805963
19	X3	0.212305	0.007276	29.18041	2.06E-09	0.195527	0.229082	0.195527	0.229082
20									

Figure 11: *Regressão de Y em X_2 e X_3 para os dados 1949-1959*

O coeficiente de regressão é 1,45. Agora, substituímos por x_1 e estimamos uma regressão de y em x_2 e x_3. Os resultados que achamos são ligeiramente melhores (temos um R^3 maior). Os resultados são:

O coeficiente de x_3 agora é (β_3+1,45).

Podemos achar estimativas separadas de β_1 e β_3 apenas se tivermos alguma informação a priori. Como indica esse exemplo, a multicolinearidade implica que não podemos estimar coeficientes individuais com boa precisão, mas podemos estimar algumas funções lineares dos parâmetros com boas precisão. Se quisermos estimar os parâmetros individuais, precisaremos de alguma informação a priori. Mostraremos que o uso dos componentes principais implica a utilização de alguma informação a priori sobre as restrições dos parâmetros.

Suponha que consideremos regredir y nos componentes principais z_1 e z_2 (z_3 é omitido porque ele é quase zero). Vimos que $z_1 = 0,7(X_1 + X_3)$ e $z_2 = X_2$. Precisamos transformá-los nas variáveis originais. Temos

$$z_1 = 0,7\left(\frac{x_{1i} - \overline{x_1}}{\sigma_1} + \frac{x_{3i} - \overline{x}}{\sigma_3}\right)$$

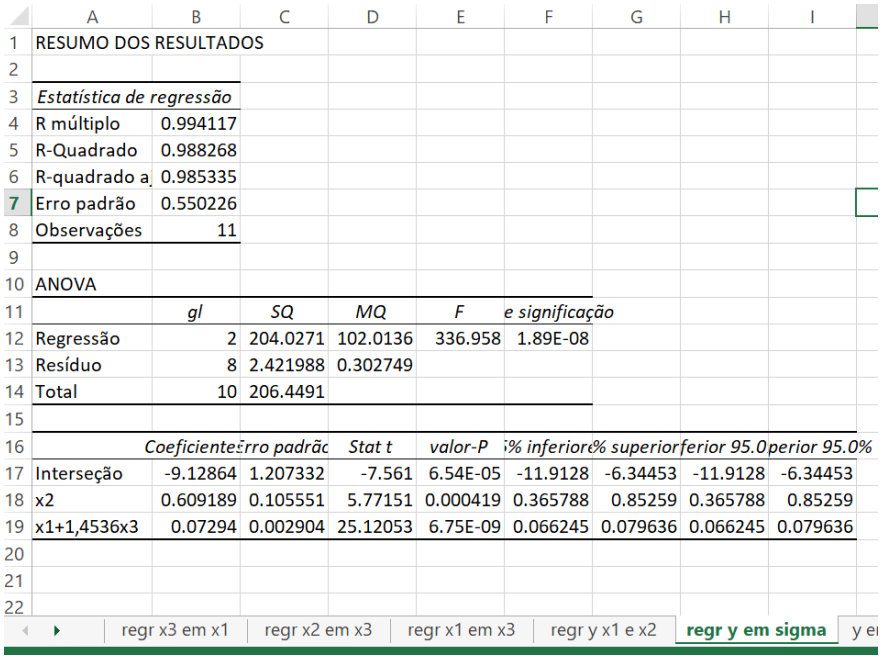
$$= \frac{0,7}{\sigma_1}\left(x_1 + \frac{\sigma_1}{\sigma_3}x_3\right) + \text{uma constante}$$

$$z_2 = \frac{1}{\sigma_2}(x_2 - \overline{x_2})$$

Portanto, usar z_2 como um regressor é equivalente a usar x_2, e usar z_1 é equivalente a usar $(x_1 + (\sigma_1/\sigma_3)x_3)$. Logo, a regressão do componente principal equivale a regredir y em $(x_1 + (\sigma_1/\sigma_3)x_3)$ e x_2. Em nosso exemplo, $\sigma_1/\sigma_3 = 1,453859199$. Esses resultados são

Essa é a equação de regressão que *teríamos estimado* se assumíssemos que $(\beta_3 = (\sigma_1/\sigma_3)\beta_1) = 1,4536\beta_1$. Assim, a regressão do componente principal equivale, nesse mesmo exemplo, ao uso da informação a priori $\beta_3 = 1,4536\beta_1$.

Se todos os componentes principais forem usados, isso é exatamente equivalente a usar o conjunto de variáveis explicativas original. Se alguns componentes principais forem omitidos, isso equivale a usar alguma informação a priori nos $\beta^{'}s$. Em nosso exemplo, a pergunta é se o pressuposto $\beta_3 = 1,45\beta_1$ tem sentido econômico. Sem mais dados desagregados que dividam as importações em bens de consumo e bens de produção, não podemos dizer nada. De qualquer forma, com 11 observações não podemos esperar responder a mais perguntas. O



	A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS								
2									
3	*Estatística de regressão*								
4	R múltiplo	0.994117							
5	R-Quadrado	0.988268							
6	R-quadrado a	0.985335							
7	Erro padrão	0.550226							
8	Observações	11							
9									
10	ANOVA								
11		*gl*	*SQ*	*MQ*	*F*	*e significação*			
12	Regressão	2	204.0271	102.0136	336.958	1.89E-08			
13	Resíduo	8	2.421988	0.302749					
14	Total	10	206.4491						
15									
16		*Coeficientes*	*rro padrão*	*Stat t*	*valor-P*	*% inferior*	*% superior*	*ferior 95.0*	*perior 95.0%*
17	Interseção	-9.12864	1.207332	-7.561	6.54E-05	-11.9128	-6.34453	-11.9128	-6.34453
18	x2	0.609189	0.105551	5.77151	0.000419	0.365788	0.85259	0.365788	0.85259
19	x1+1,4536x3	0.07294	0.002904	25.12053	6.75E-09	0.066245	0.079636	0.066245	0.079636
20									
21									
22									

regr x3 em x1 | regr x2 em x3 | regr x1 em x3 | regr y x1 e x2 | **regr y em sigma** | y e

Figure 12: *Regressão de* Y *em* $X_1 + 1,4536X_3$ *e* X_2 *para os dados 1949-1959*

propósito de nossa análise tem sido meramente mostrar o que é a regressão do componente principal e que ela implica alguma informação a priori.