FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

A Correlação Revisitada

Carla Daniela Moreira Bessa

Licenciada em Matemática Aplicada e Licenciada em Matemática Educacional
pela Faculdade de Ciências da Universidade do Porto

Dissertação submetida para satisfação parcial dos requisitos do grau de
mestre em Estatística Aplicada e Modelação

Dissertação realizada sob a orientação de
Professor Doutor Francisco Lage Calheiros,
do Departamento de Engenharia Civil
da Faculdade de Engenharia da Universidade do Porto

e co-orientação de
Professora Doutora Maria Manuela Neves Figueiredo,
do Departamento de Matemática
do Instituto Superior de Agronomia da Universidade Técnica de Lisboa

Porto, Janeiro 2010

Dedico este trabalho ao meu orientador e amigo, Francisco Calheiros,

aos meus pais, Rogério e Rosa,

e a duas pessoas muito especiais, Paulo e Pedro.

Los ríos no llevan agua,
el sol las fuentes secó...

¡Yo sé donde hay una fuente
que no ha de secar el sol!

La fuente que no se agota
es mi propio corazón...

V. Ruiz Aguilera (1862)

Resumo

Através do estudo da correlação entre duas variáveis, determina-se, por simulação, a distribuição amostral do coeficiente de correlação global, η_{**}^2 (proporção de variância que é explicada pela curva de regressão) e por consequência a distribuição dos $\eta_{**}^2 - r^2$, onde r^2 é o quadrado do coeficiente de correlação linear, ficando assim disponível um teste expedito para a não linearidade. O coeficiente global é calculado com os dados agrupados em classes, à excepção de casos em que as classes já estão predefinidas e por isso não há necessidade de se agrupar. As questões de agrupamento são aqui ultrapassadas, pelo menos em alguns casos, por extrapolações para classes de agrupamento de amplitude zero, ou seja, estuda-se uma estatística amostral sem fixação de classes.

É também discutido o que intervém e o que influência os η_{**}^2 e investiga-se o efeito conjunto das características das classes, estudando alguns exemplos clássicos.

Faz-se uma pequena discussão da utilidade de coeficientes de correlação não ligados a normas dos mínimos quadrados, por exemplo, regressão dos Mínimos Desvios Absolutos (MDA) e regressão Quantílica.

Palavras-chave: Correlação; Regressão; Coeficiente de Correlação Global; Linearidade e Não Linearidade.

Abstract

Sampling distribution of the global correlation coefficient, η_{**}^2 (proportion of variance that is explained by the regression curve) is obtained by simulation. Sampling distribution of $\eta_{**}^2 - r^2$ is therefore obtained, where r is the linear correlation coefficient. As a byproduct an easy test for non linearity is obtained. Grouping questions are discussed; some extrapolation for intervals of width zero are worked without setting intervals.

On examples, what change η_{**}^2 is analyzed and grouping effects is evaluated. The central role of global correlation coefficients is highlighted.

A small discussion on other regression methods, like Lest Absolute Deviations (LAD) regression and Quantile regression, is presented.

Keywords: Correlation; Regression; Global Correlation Coefficient; Linearity and Non-linearity.

Résumé

La distribution d'échantillonnage de coefficient de corrélation global, η_{**}^2 (proportion de variance expliquée par la courbe de régression) est obtenus par simulation, et en conséquence la distribution d'échantillonnage de $\eta_{**}^2 - r^2$ où r^2 est le carré du coefficient de corrélation linéaire. Un teste simple de non-linéarité est donc obtenu. Les questions de regroupement sont dépassées, pour le moins dans quelques cas par des extrapolations pour classes de regroupement d'amplitude zéro, c'est-à-dire, on obtenu un statistique échantillon sans fixation de classes.
Dans des exemples on a analyse ce que intervient dans η_{**}^2. L'importance de ces coefficients de corrélation global est souligné.
Un courte discussion sur outres méthodes de corrélation, régression des Minimes Détours Absolus (MDA) et régression Quantile, est présentée.

Mots-clés: Corrélation; Régression, Coefficient de Corrélation Globale; Linéaire et Non Linéaire.

Motivação / Estado da Arte

A correlação linear, pela sua simplicidade, é possivelmente a medida estatística mais usada nas aplicações, Tomassone *et al* (1983). É também a base de inúmeras construções teórico – práticas como, por exemplo, a Análise em Componentes Principais, os estudos de séries cronológicas pelo método de Box-Jenkins, etc., onde tudo o que não é linear não é tido em conta.

Muitas adaptações são usadas para acomodar a não linearidade, muitas vezes sem qualquer verificação de que a não linearidade é significativa (e em que sentido), privilegiando geralmente um melhor ajuste aos dados em detrimento de uma melhor capacidade preditiva (ver, por exemplo, os estudos com Splines de Friedman (1991) e, Durand and Sabatier (1997) e trabalhos com o uso de transformações desvalorizando a presença de misturas, Calheiros (2002)). Muitos trabalhos tentam melhorar os estudos de correlação-regressão procurando métodos alternativos, ver Dodge and Birkes (1993) e, por exemplo, Santos and Neves (2007), Magalhães (2006). Geralmente estes trabalhos não têm em conta os coeficientes de correlação global e portanto não é avaliada a não linearidade.

A não linearidade seguiu portanto outra via, onde a variável predictora não toma uma determinada forma mas é construída com base no conjunto de dados. Uma das técnicas usadas é o método do núcleo Gasser *et al* (1979), através de métodos não paramétricos e outra é o método das Splines que usa métodos semi-paramétricos, Eubank (1999) cuja motivação é a regressão paramétrica, sem avaliação prévia da não linearidade.

Desde os trabalhos pioneiros, por exemplo, Blakeman (1905), Student (1913), Fisher (1925), Kelley (1935), percebeu-se da importância dos coeficientes de correlação global $\eta^2_{X|Y}$ e $\eta^2_{Y|X}$ (designados por razões de correlação sobretudo no contexto das ANOVAS gaussianas e que correspondem, respectivamente, à correlação global de *X* dado *Y* e à correlação global de *Y* dado *X*). A possibilidade de usar na análise estes dois coeficientes simultaneamente parece não ter sido analisada.

Estes coeficientes são referidos mais tarde em textos em francês para disciplinas introdutórias de Estatística no contexto da Estatística Descritiva Clássica, Calot (1969), Grais (1974), Dagnelie (1973). Há por vezes indicação para o uso de $\eta^2_{**} - r^2$ como medida de não linearidade. Pela nossa pesquisa não houve qualquer desenvolvimento para além do trabalho solitário de Dodge and Rousson (1999) com o uso de momentos cruzados de terceira ordem.

A razão para o pouco uso destes coeficientes foi identificada por Kelly (1935) pois as estimativas de η^2_{**} para além das flutuações amostrais têm uma componente sistemática associada à criação *ad-hoc* de classes na variável independente. As correcções do tipo Sheppard para os coeficientes de correlação estão pouco estudadas mesmo no caso linear.

Índice

Índice de Figuras

Índice de Tabelas

Capítulo 1: Introdução

1.1. Objectivo

Dadas duas variáveis, estuda-se a informação que o comportamento da respectiva distribuição conjunta pode fornecer. Tira-se melhor partido destes dados se os estudarmos individualmente ou agrupando em classes? A resposta a esta questão só poderá ser dada depois de algum estudo da correlação e da regressão.

Numa primeira fase, vai-se rever conjuntos de dados, através do percurso clássico (habitual).

1.2. Percurso Clássico

No sentido de haver uma melhor compreensão do que se vai estudar neste trabalho e de que forma estudos mais extensos podem levar a uma obtenção de maior informação e mais útil, foram escolhidos três conjuntos de dados e elaborou-se um breve estudo inicial em torno destes dados. Os exemplos escolhidos foram retirados de Aïvazian (1970), Grais (1974) e Calot (1969) e foram já inseridos no âmbito do ajustamento linear, Bessa (2007).

O exemplo escolhido de Aïvazian (1970) refere-se a determinações da resistência à ruptura de cubos de cimento após 1, 2, 3, 7 e 28 dias de cura (Tabela 1 e Figura 1). Temos de ter em conta que a precisão da data não é absoluta, por exemplo, cada valor é representante de um dia mas não de uma hora fixada são por isso classes.

Tabela 1 – Resistência à ruptura de cubos de cimento (kgf/cm^2) e o seu tempo de cura (dias)

Cubo de cimento / Tempo	1	2	3	4	5
1	13,0	13,3	11,8		
2	21,9	24,5	24,7		
3	29,8	28,0	24,1	24,2	26,2
7	32,4	30,4	34,5	33,1	35,7
28	41,8	42,6	40,3	35,7	37,3

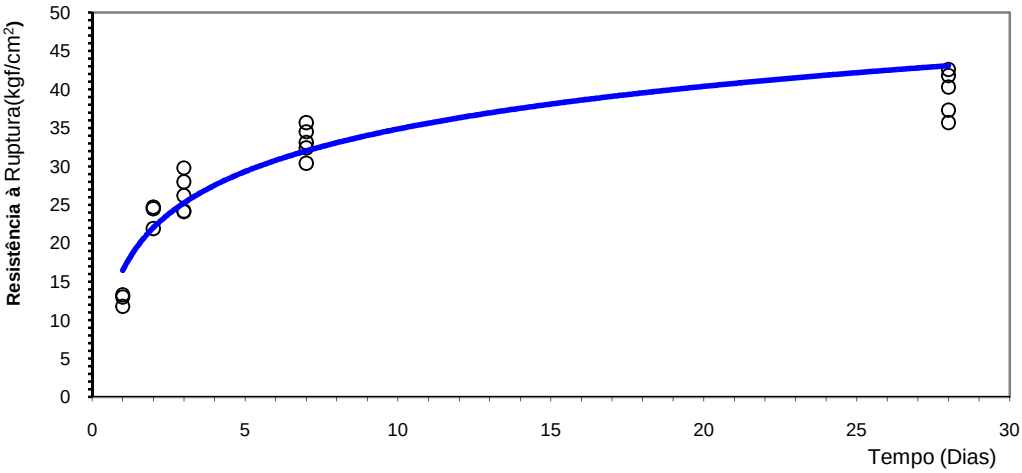


Figura 1 – Curva de regressão da dependência entre a resistência à ruptura dos cubos de cimento e o tempo de cura.

Através da pesquisa feita, conclui-se que para muitos investigadores só se começou a ver a regressão não linear com este tipo de dados. O procedimento normal é não tomar em conta a correlação não linear e obter os valores de η^2_{**} e de r^2 quando aplicada uma transformação ao conjunto. Neste conjunto de dados, a transformação consiste em considerar no eixo dos *xx* o inverso do tempo e fazer-lhe corresponder, no eixo dos *yy*, o logaritmo neperiano da resistência à ruptura (Figura 2).

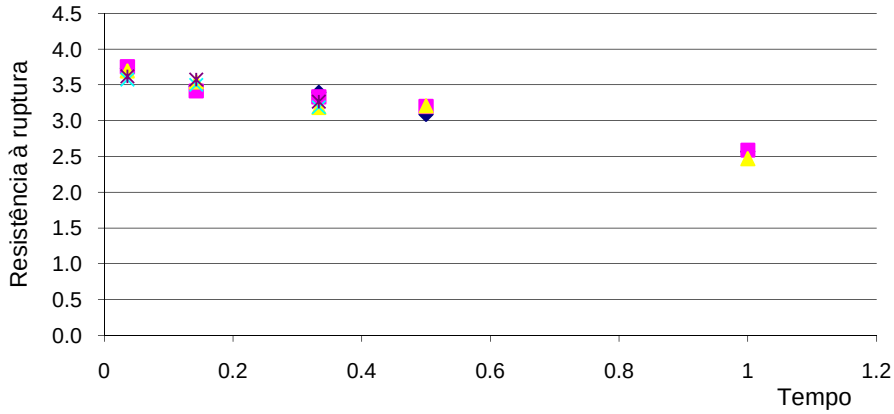


Figura 2 – Dependência entre os valores do inverso do tempo e o logaritmo da resistência à ruptura

Para o cálculo das medidas estatísticas procedeu-se ao agrupamento dos dados, em classes, apenas para a variável *y* – "resistência à ruptura", visto que a variável *x* – "tempo" já se encontra dividida em classes que correspondem aos cinco dias da recolha. Na Tabela 2, encontra-se a descrição deste agrupamento (a tabela dos dados transformados não será aqui apresentada visto ter a mesma estrutura). Para a escolha das classes e da respectiva amplitude, teve-se em consideração a amplitude dos dados e de seguida optou-se por uma divisão em cinco classes. Junto às classes encontram-se as marcas das classes, que correspondem à média dos valores inseridos em cada classe.

Tabela 2 – Dados agrupados da resistência à ruptura de cubos de cimento (kgf/cm^2) e o seu tempo de cura (dias)

Tempo \ Resistência à ruptura	[9,16[	[16,23[	[23,30[	[30,37[	[37,44[	Total
	12,70	21,90	25,9286	33,633	40,5	
1	3					3
2		1	2			3
3			5			5
7				5		5
28				1	4	5
Total	3	1	7	6	4	21

Os valores das medidas do conjunto de dados e da sua transformação podem ser observados na Tabela 3. Estes valores foram obtidos através do software *Microsoft Office Excell 2003*.

Tabela 3 – Valores descritivos da resistência à ruptura de cubos de cimento e o seu tempo de cura

	Variáveis iniciais	Variáveis transformadas
$\bar{x}$	9,4762	0,3362
$\bar{y}$	28,8238	3,3027
s_x	10,5543	0,3117
s_y	8,6483	0,3590
$\eta^2_{Y\|X}$	0,9691	0,9630
$\eta^2_{X\|Y}$	0,8423	0,9759
r	0,7792	-0,9847
r^2	0,6071	0,9647
γ_{1x}	1,1044	1,1303
γ_{1y}	-0,4448	-1,0631
γ_{2x}	-0,6109	0,1890
γ_{2y}	-0,5900	0,2308

Analisando o coeficiente de assimetria na variável y, verifica-se que a transformação afastou-nos da normalidade, o que não era o desejado, visto que uma transformação "linearizante" nos deveria aproximar da normalidade. Em termos de "achatamento" esta foi neutra.

Pode-se concluir ainda, que o coeficiente de correlação global de *Y* dado *X* não muda muito com a transformação. No entanto, a transformação nos dados leva a um claro aumento do coeficiente de correlação global de *X* dado *Y*. Quanto ao coeficiente de correlação linear, este também é claramente alterado pois o seu valor absoluto aumenta de 0.7792 para 0.9847, o que indica que a transformação é "linearizante", isto é, lineariza fortemente os dados. Sendo assim, não se tomaria em atenção a correlação não linear. Este é o procedimento clássico de quem estuda este tipo de dados. No entanto, se analisarmos o conjunto de dados antes da transformação, este possui uma grande

componente de não linearidade, $\eta^2_{Y|X} - r^2 = 0.3620$ e $\eta^2_{X|Y} - r^2 = 0{,}2352$. Com este resultado somos levados a pensar, de que forma esta componente não linear afecta as conclusões que retiramos ao analisar o conjunto de dados resultante da transformação feita? Esta transformação resultou bem nos dados e linearizou-os, mas é de ter em atenção que nem sempre é possível encontrar uma transformação que tenha este efeito. Sendo assim, se não houvesse ou não se conhecesse uma transformação linearizante como procederíamos? Existem teorias, métodos, processos que nos permitam tirar conclusões válidas?

De seguida, procedeu-se ao cálculo de mais algumas medidas estatísticas, mas condicionadas a cada uma das classes das variáveis X e Y (Tabelas 4 e 5).

Tabela 4 – Valores descritivos condicionais da resistência à ruptura de cubos de cimento dado o tempo de cura – $Y \mid x_i$

Condicionais			Valores das Medidas	
Variáveis iniciais	Variáveis transformadas		Variáveis iniciais	Variáveis transformadas
$y\|_{x=1}$	$y\|_{x=1}$	$\bar{y}$	12,700	2,540
		s_y	0	0
$y\|_{x=2}$	$y\|_{x=0.5}$	$\bar{y}$	24,586	3,197
		s_y	1,899	0,00612
$y\|_{x=3}$	$y\|_{x=1/3}$	$\bar{y}$	25,929	3,252
		s_y	0	0
$y\|_{x=7}$	$y\|_{x=1/7}$	$\bar{y}$	33,662	3,514
		s_y	0	0
$y\|_{x=28}$	$y\|_{x=1/3}$	$\bar{y}$	39,127	3,663
		s_y	2,747	0,00554

Tabela 5 – Valores descritivos condicionais do tempo de cura dada a resistência à ruptura de cubos de cimento – $X \mid y_j$

Condicionais			Valores das Medidas	
Variáveis iniciais	Variáveis transformadas		Variáveis iniciais	Variáveis transformadas
$\boldsymbol{x}\|_{\boldsymbol{y\epsilon[9,16[}}$	$\boldsymbol{x}\|_{\boldsymbol{y\epsilon[\ln 9,\ln 16[}}$	$\bar{x}$	1	1
		s_x	0	0
$\boldsymbol{x}\|_{\boldsymbol{y\epsilon[16,23[}}$	$\boldsymbol{x}\|_{\boldsymbol{y\epsilon[\ln 16,\ln 23[}}$	$\bar{x}$	2	0,5
		s_x	0	0
$x\|_{y\epsilon[23,30[}$	$x\|_{y\epsilon[\ln 23,\ln 30[}$	$\bar{x}$	2,714	0,381
		s_x	0,452	0,0753
$x\|_{y\epsilon[30,37[}$	$x\|_{y\epsilon[\ln 30,\ln 37[}$	$\bar{x}$	10,5	0,125
		s_x	7,826	0,0399
$x\|_{y\epsilon[37,44[}$	$x\|_{y\epsilon[\ln 37,\ln 44[}$	$\bar{x}$	28	0,0367
		s_x	0	0

É de salientar que, neste exemplo os dados não são suficientes para se avaliar a assimetria e o "achatamento", pelo que estas medidas não foram calculadas.

O segundo exemplo é retirado de Grais (1974) e trata da relação entre o "salário mensal", variável *y* e a "idade dos trabalhadores", variável *x*. Este conjunto de dados encontra-se descrito na Tabela 6, onde estão também indicadas as marcas das classes junto às respectivas classes.

Tabela 6 – Número de trabalhadores distribuídos de acordo com a idade e salário mensal

Idade \ Salário		< 800	[800,900[	[900,1000[	[1000,1200[	[1200,1500[	[1500,2000[	>2000	Total
		700	850	950	1100	1350	1750	2200	
< 25	20	207	302	18					527
[25,30[	27,5	121	461	526	111	1			1220
[30,35[	32,5	38	513	682	342	3			1578
[35,40[	37,5	17	103	567	298	182	18	1	1186
[40,45[	42,5	10	86	613	416	227	22	14	1388
[45,50[	47,5	2	6	431	480	263	13	6	1201
[50,55[	52,5	7	10	105	226	98	12	7	465
> 55	60	3	2	60	37	18	5	5	130
	Total	405	1483	3002	1910	792	70	33	7695

O último exemplo foi retirado de Calot (1969) e descreve a "resistência à ruptura de fibras de linho", variável *y* e o seu "comprimento", variável *x*. Este conjunto de dados encontra-se descrito na Tabela 7 onde, tal como no exemplo anterior, estão indicadas as marcas das classes junto às respectivas classes.

Tabela 7 – Resistência à ruptura de fibras de linho e respectivo comprimento

Comprimento \ Resistência à ruptura		< 80	[80,85[	[85,90[	[90,95[	[95,100[	[100,105[	[105,110[	[110,115[	[115,120[	[120,125[	[125,130[	> 130	Total
		77,5	82,5	87,5	92,5	97,5	102,5	107,5	112,5	117,5	122,5	127,5	132,5	
< 60	57,5	1			1									2
[60,65[	62,5	1	2	1	2	1		1						8
[65,70[	67,5	1	3	4	4	12	4	1						29
[70,75[	72,5	2	2	5	12	10	8	1						40
[75,80[	77,5		1		5	5	9	9	4					33
[80,85[	82,5		1	2	3	7	11	7	2	2		2		37
[85,90[	87,5			1	1	7	3	1	2	1	1	1	2	20
[90,95[	92,5					1	1	1		2	3	1	1	10
[95,100[	97,5											1	2	3
> 100	102,5											1		1
Total		5	9	13	28	43	36	21	8	5	4	6	5	183

Em cada um destes exemplos foram determinadas as medidas usuais de estudo através do software *Matlab 7.7.0*, sendo que as marcas das classes aqui descritas foram as consideradas pelos respectivos autores. Esta escolha deve-se ao facto de não se possuir o conjunto de dados individuais, e assim sendo, considera-se que os autores usaram os melhores valores representativos de cada classe. Os valores das medidas estatísticas associadas a cada uma das variáveis estão indicados na Tabela 8.

Tabela 8 – Valores descritivos para os conjuntos de dados, Salários e Fibras.

	Salários	Fibras
$\bar{x}$	37,4299	77,3361
$\bar{y}$	1008,6095	100,5601
s_x	9,1770	8,5874
s_y	190,6751	11,7199
$\eta^2_{Y\|X}$	0,3609	0,4533
$\eta^2_{X\|Y}$	0,2706	0,4162
r	0,5012	0,6321
r^2	0,2512	0,3996
γ_{1x}	0,0504	0,2385
γ_{1y}	1,8535	0,7077
γ_{2x}	-0,5740	-0,3312
γ_{2y}	7,1434	0,7032

Na Tabela 8, se compararmos os coeficientes de correlação global e o quadrado do coeficiente de correlação linear, verificámos que em ambos os casos (ou seja, se considerarmos a variável y condicionada à *x* e vice-versa) vamos ter uma componente não linear, sendo esta muito menor no último caso. É no exemplo das Fibras que existe uma componente não linear mais visível, tomando esta o valor $\eta^2_{Y|X} - r^2 = 0.1097$. Se analisarmos a assimetria e o "achatamento" da variável *y*, no exemplo dos Salários, verificamos que esta parece ser normal, sendo que a última medida, γ_{2y}, encontra-se muito distante dos valores normais. No que diz respeito à variável *x*, a mesma parece ser simétrica e não possui um elevado valor do "achatamento". No exemplo das Fibras, verifica-se que a variável *y* é mais assimétrica, mas os valores não são demasiado elevados.

Será legítimo tirar conclusões para este conjunto de dados usando os valores de *x* condicionados aos de *y*? Ou seja, no caso da recta de regressão, deve-se tirar conclusões a partir da recta de ajustamento de x dado y?

De seguida, tal como no exemplo do Betão, fez-se um estudo mais pormenorizado das medidas condicionadas a cada classe. Também aqui temos classes com poucos elementos, pelo que o cálculo da assimetria só foi determinado nas classes que possuíam no mínimo 3 elementos e para o cálculo do "achatamento", classes com pelo menos 4 elementos

Tabela 9 – Valores descritivos condicionais para o conjunto de dados Salários

Condicionais – $X\mid_{Y\in]Y_0,Y_0+h[}$		**Valores das medidas**	**Condicionais –** $Y\mid_{X\in]X_0,X_0+h[}$		**Valores das medidas**
$\boldsymbol{x\mid_{y<800}}$	$\bar{x}$	25,698	$\boldsymbol{y\mid_{x<25}}$	$\bar{y}$	794,497
	s_x	7,629		s_y	78,095
	γ_1	1,815		γ_1	-0,221
	γ_2	4,031		γ_2	-1,475
$\boldsymbol{x\mid_{y\epsilon[800,900[}}$	$\bar{x}$	29,560	$\boldsymbol{y\mid_{x\epsilon[25,30[}}$	$\bar{y}$	901,393
	s_x	6,566		s_y	98,960
	γ_1	0,464		γ_1	-0,026
	γ_2	0,809		γ_2	0,576
$x\mid_{y\epsilon[900,1000[}$	$\bar{x}$	37,938	$y\mid_{x\epsilon[30,35[}$	$\bar{y}$	944,740
	s_x	7,893		s_y	99,544
	γ_1	0,401		γ_1	0,313
	γ_2	-0,295		γ_2	-0,048
$x\mid_{y\epsilon[1000,1200[}$	$\bar{x}$	41,836	$y\mid_{x\epsilon[35,40[}$	$\bar{y}$	1050,000
	s_x	7,649		s_y	181,664
	γ_1	-0,055		γ_1	1,480
	γ_2	-0,729		γ_2	3,212
$x\mid_{y\epsilon[1200,1500[}$	$\bar{x}$	44,590	$y\mid_{x\epsilon[40,45[}$	$\bar{y}$	1077,666
	s_x	5,448		s_y	208,773
	γ_1	0,341		γ_1	2,236
	γ_2	-0,088		γ_2	7,829
$x\mid_{y\epsilon[1500,2000[}$	$\bar{x}$	45,107	$y\mid_{x\epsilon[45,50[}$	$\bar{y}$	1111,532
	s_x	6,580		s_y	181,177
	γ_1	0,663		γ_1	1,767
	γ_2	-0,381		γ_2	6,128
$x\mid_{y>2000}$	$\bar{x}$	48,030	$y\mid_{x\epsilon[50,55[}$	$\bar{y}$	1140,753
	s_x	6,506		s_y	222,403
	γ_1	0,623		γ_1	2,022
	γ_2	-0,788		γ_2	6,782
			$y\mid_{x>55}$	$\bar{y}$	1119,61
				s_y	293,421
				γ_1	2,200
				γ_2	5,100

Na Tabela 9, podemos observar que a média de *x* dado *y* cresce com o aumento do valor de *y*, o mesmo não acontece na média de *y* dado *x*, basta olharmos para a última classe.

Tabela 10 – Valores descritivos condicionais para o conjunto de dados Fibras

Condicionais – $X\ \vert_{Y\in]Y_0,Y_0+h[}$		**Valores das medidas**	**Condicionais –** $Y\ \vert_{X\in]X_0,X_0+h[}$		**Valores das medidas**
$\boldsymbol{x\vert_{y<80}}$	$\bar{x}$	65,500	$\boldsymbol{y\vert_{x<60}}$	$\bar{y}$	85,000
	s_x	5,831		$\boldsymbol{s_y}$	7,500
	γ_1	-0,363		γ_1	--
	γ_2	-1,372		γ_2	--
$\boldsymbol{x\vert_{y\in[80,85[}}$	$\bar{x}$	79,444	$\boldsymbol{y\vert_{x\in[60,65[}}$	$\bar{y}$	90,000
	s_x	11,115		$\boldsymbol{s_y}$	9,1014
	γ_1	-1,254		γ_1	0,512
	γ_2	-1,320		γ_2	-0,580
$x\vert_{y\in[85,90[}$	$\bar{x}$	85,577	$y\vert_{x\in[65,70[}$	$\bar{y}$	94,224
	s_x	14,458		$\boldsymbol{s_y}$	7,105
	γ_1	-1,200		γ_1	-0,550
	γ_2	-1,491		γ_2	-0,370
$x\vert_{y\in[90,95[}$	$\bar{x}$	81,875	$y\vert_{x\in[70,75[}$	$\bar{y}$	94,250
	s_x	10,946		$\boldsymbol{s_y}$	6,942
	γ_1	-1,390		γ_1	-0,529
	γ_2	-0,758		γ_2	-0,039
$x\vert_{y\in[95,100[}$	$\bar{x}$	89,419	$y\vert_{x\in[75,80[}$	$\bar{y}$	102,197
	s_x	15,576		$\boldsymbol{s_y}$	7,065
	γ_1	-1,262		γ_1	-0,602
	γ_2	-1,328		γ_2	0,034
$x\vert_{y\in[100,105[}$	$\bar{x}$	103,264	$y\vert_{x\in[80,85[}$	$\bar{y}$	103,041
	s_x	25,962		$\boldsymbol{s_y}$	9,571
	γ_1	-1,081		γ_1	0,516
	γ_2	-1,782		γ_2	0,738
$x\vert_{y\in[105,110[}$	$\bar{x}$	106,429	$y\vert_{x\in[85,90[}$	$\bar{y}$	106,750
	s_x	28,190		$\boldsymbol{s_y}$	13,065
	γ_1	-1,076		γ_1	0,736
	γ_2	-1,785		γ_2	-0,644
$x\vert_{y\in[110,115[}$	$\bar{x}$	101,875	$y\vert_{x\in[90,95[}$	$\bar{y}$	117,000
	s_x	21,038		$\boldsymbol{s_y}$	10,595
	γ_1	-1,053		γ_1	-0,499
	γ_2	-1,865		γ_2	-0,861
$x\vert_{y\in[115,120[}$	$\bar{x}$	120,500	$y\vert_{x\in[95,100[}$	$\bar{y}$	130,833
	s_x	33,302		$\boldsymbol{s_y}$	2,357
	γ_1	-1,027		γ_1	-0,707
	γ_2	-1,929		γ_2	--
$x\vert_{y\in[120,125[}$	$\bar{x}$	91,250	$y\vert_{x>100}$	$\bar{y}$	127,500
	s_x	2,165		$\boldsymbol{s_y}$	0,000
	γ_1	-1,155		γ_1	--
	γ_2	-0,667		γ_2	--
$x\vert_{y\in[125,130[}$	$\bar{x}$	118,333			
	s_x	28,492			
	γ_1	-1,092			
	γ_2	-1,761			
$x\vert_{y>130}$	$\bar{x}$	92,500			
	s_x	4,472			
	γ_1	0,000			
	γ_2	-1,750			

Na Tabela 10, as médias condicionais de *x* apresentam oscilações com o aumento de *y*, a média de *y* dado *x* apresenta um crescimento com a excepção da última classe. Os valores da assimetria não variam muito entre as classes.

Os valores das condicionadas, neste exemplo, não indicam se determinadas classes não sugerem normalidade, mas verifica-se que estas possuem piores resultados do que as obtidas com todo o conjunto de dados.

Para tentar obter mais algumas respostas pode-se recorrer às respectivas representações gráficas.

A representação gráfica dos conjuntos de dados foi elaborada com a ajuda do software *Winplot*, bem como as respectivas curvas de regressão condicionadas e as rectas de regressão.

Para cada conjunto de dados foram elaborados alguns cálculos para a obtenção da recta de regressão, cálculos que serão aqui omitidos e serão apenas apresentadas as equações das rectas de regressão. Estes cálculos tem por base a teoria que será desenvolvida nas secções 2.2 e 2.3.

Para o exemplo dos Salários, a recta de regressão de X dado Y é dada por $Y = -544{,}3095 + 41{,}4555X$ e a recta de regressão de Y dado X é dada por $Y = 618{,}5289 + 10{,}4216X$. Para o exemplo das Fibras, a recta de regressão de X dado Y é dada por $Y = -66{,}4177 + 2{,}1591X$ e a recta de regressão de Y dado X é dada por $Y = 33{,}8441 + 0{,}86227X$.

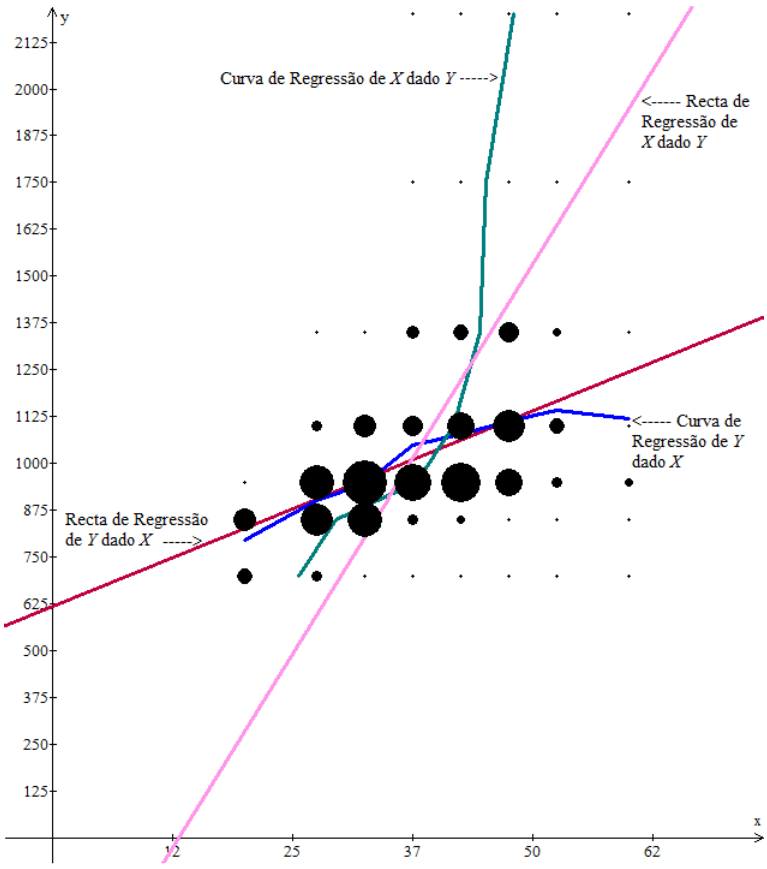


Figura 3 – Distribuição do número de trabalhadores de acordo com a idade e salário mensal. Curvas de regressão condicionadas e recta de regressão.

As curvas de regressão são obtidas unindo os valores médios de cada classe quando se fixa uma das variáveis. Estes valores encontram-se descritos nas Tabelas 9 e 10.

Analisando a Figura 3, parece-nos que a curva de regressão de Y dado X aproxima-se mais da respectiva recta de regressão do que a curva de regressão de X dado Y, no entanto existe uma maior componente não linear quando fixamos as classes em X. Porque será que isto acontece? Será que é pelo facto de $\eta^2_{Y|X}$ ser maior do que $\eta^2_{X|Y}$? Será porque a correlação linear existente não é muito expressiva?

Na Figura 4 ambas as curvas de regressão aproximam-se da recta de regressão. Neste exemplo tem-se que o coeficiente de correlação linear é maior do que no exemplo anterior e a componente não linear é menor. Verifica-se ainda que a curva de regressão de X dado Y aproxima-se melhor da respectiva curva de regressão. De que forma estes coeficientes influenciam então as curvas de regressão? Se em vez de classes, fossem valores individuais, estas conclusões mudariam?

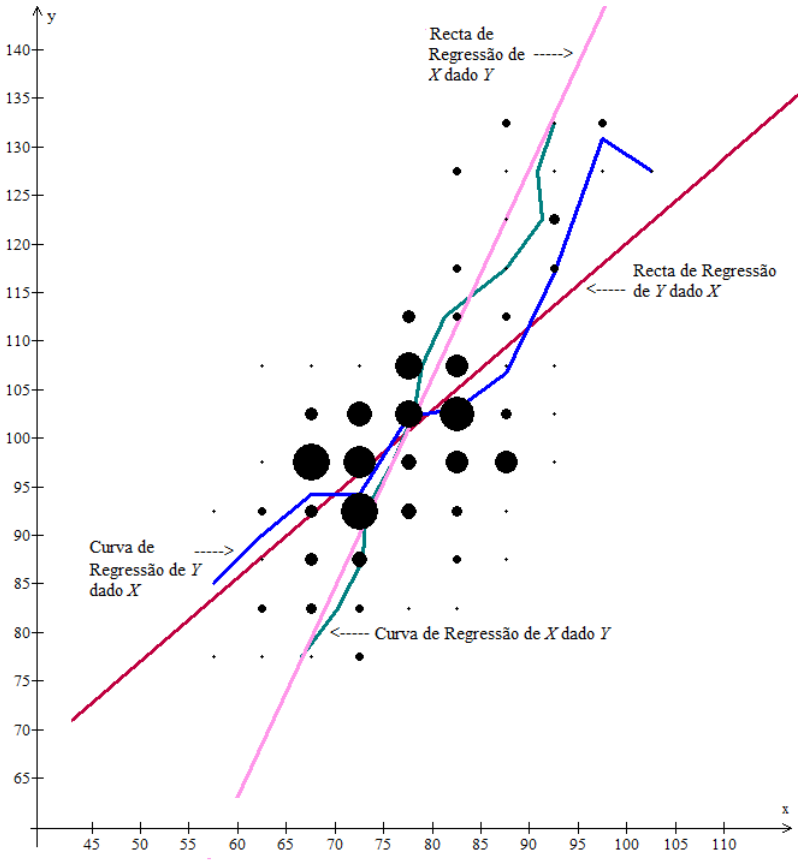


Figura 4 – Distribuição da resistência à ruptura de fibras de linho e respectivo comprimento. Curvas de regressão condicionadas e recta de regressão.

1.3. Organização da Dissertação

No capítulo 2, é feita uma revisão dos conceitos teóricos usados na tese e são abordadas várias metodologias numa tentativa de dar resposta a algumas das questões levantadas na secção anterior e outras que possam surgir.

No capítulo 3, estas metodologias são aplicadas aos conjuntos de dados já apresentados e a outros conjuntos de dados individuais e feita a interpretação dos resultados.

No capítulo 4, recorrendo a simulações obtém-se a distribuição amostral do coeficiente de correlação global e discute-se de que forma pode ser usada para testar se a existência de não linearidade num conjunto de dados é ou não significativa.

No capítulo 5 apresenta-se alguns comentários sobre o trabalho desenvolvido e as questões que nos foram surgindo no decorrer do nosso estudo.

Capítulo 2 : Elementos da Teoria

Neste capítulo optamos por esta ordem das secções, mas as implicações duns aspectos nos outros poderiam sugerir outra ordem.

2.1. Correlação e Regressão

Um estudo de Correlação consiste em analisar se duas variáveis estão ligadas em algum sentido, ou seja, verificar se têm comportamento ligado. A correlação global pode ser perfeita num sentido e ser nula no outro, ou seja, pode a variável X dar-nos informação completa sobre a variável Y, mas conhecida a variável Y não obtermos qualquer informação sobre a variável X. Naturalmente, todas as situações intermédias são possíveis.

A Regressão é um método estatístico muito utilizado que permite modelar a "ligação" entre variáveis e neste trabalho será abordado sob o ponto de vista bidimensional. Este método consiste, no caso mais simples, em conhecida uma variável aleatória X determinar a variável aleatória Y, ou seja, efectua-se apenas num sentido. A esta variável costuma-se designar por variável dependente e representa-se por $Y|_{X=x}$.

Supondo assim que se pretende determinar a curva de regressão de Y dado X (de forma análoga se obtém a curva de regressão de X dado Y), a curva de regressão teórica é dada por $f(x) = E(Y|_{X=x})$, onde $Y|_{X=x}$ costuma ser designada por variável condicional e a estimativa da curva de regressão de Y condicionada por $X = x$ é dada por:

$$\tilde{y}_i = f(x_i + h) = \frac{\sum y_i|_{x \in "classe"}}{n_i} =$$ "média dos valores de y naquela "classe".

A técnica para se determinar cada uma das curvas num conjunto de dados consiste em, dividir a nuvem de pontos em classes segundo o eixo dos *xx* e/ou o eixo dos *yy*, ou seja, determinar respectivamente $Y|_{X \in]X_0, X_0+h[}$ e/ou $X|_{Y \in]Y_0, Y_0+h[}$. Em cada classe calcula-se a média condicional e substitui-se cada variável condicional pelo seu valor esperado – média condicional. Desta forma está-se a minimizar o erro quadrático médio, Grais (1964). Assim, a curva de regressão é a curva do plano que melhor aproxima os dados pelo critério dos mínimos quadrados.

Uma representação gráfica do ajuste de uma nuvem de pontos (x_i, y_i), $i = 1...n$ através da curva de regressão global pode ser a apresentada na Figura 5. E a curva estimada, constituída pelos pontos $(x_i, \overline{Y}|_{X=x_i})$, pode ter a representação da Figura 6.

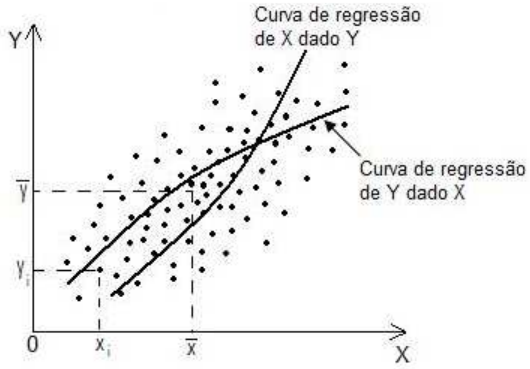


Figura 5 – Curvas de regressão condicionadas

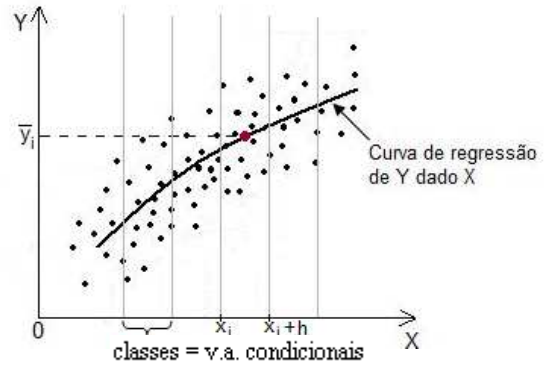


Figura 6 – Curva de Regressão relativa a uma distribuição de duas variáveis

Uma forma de sumariar os pares de pontos observados, (x_i, y_i), é recorrer a uma tabela de contingência. Por vezes, e vai ser do nosso interesse, descrever a variável através de classes. Assim, os valores do par (x_i, y_i) corresponderão às marcas de cada classe. Como usualmente designamos por marca de uma classe o valor atribuído para efeito do cálculo das características descritivas algébricas. Quase sempre é o ponto médio das classes. Para classes abertas escolhe-se um valor por algum critério que pareça razoável.

A notação que adopta neste trabalho será a mesma que a de Grais (1974) e é a seguinte:

Tabela 11 – Representação de uma distribuição estatística com duas variáveis

Variável Y / Variável X	y_1	y_2	…	y_j	…	y_l	Total marginal de X
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1l}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2l}	$n_{2\bullet}$
⋮	⋮	⋮		⋮		⋮	⋮
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{il}	$n_{i\bullet}$
⋮	⋮	⋮		⋮		⋮	⋮
x_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{kl}	$n_{k\bullet}$
Total marginal de Y	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet j}$	…	$n_{\bullet l}$	n

Com os dados representados na forma da Tabela 11 podemos dizer que a curva de regressão estimada de y dado x, definida pela fixação dos valores da variável x terá a seguinte expressão:

$$\tilde{y}_i = \sum_j y_j \frac{n_{ij}}{n_{i\cdot}}.$$

As características e definições associadas à Tabela 11 podem então ser escritas da seguinte forma:

☞ Características amostrais marginais da variável X

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet} x_i \;\rightarrow \text{ média aritmética marginal}$$

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}\left(x_i - \bar{x}\right)^2 \;\rightarrow \text{ variância marginal}$$

$$\gamma_{1x} = \frac{\mu_3}{s_x^3} \text{ onde } \mu_3 = \frac{1}{n}\sum_{i=1}^{k} n_{i\cdot}(x_i - \bar{x})^3 \;\rightarrow \text{ assimetria}$$

$$\gamma_{2x} = \frac{\mu_4}{s_x^4} - 3 \text{ onde } \mu_4 = \frac{1}{n}\sum_{i=1}^{k} n_{i\cdot}(x_i - \bar{x})^4 \;\rightarrow \text{ “achatamento”}$$

☞ Características amostrais marginais da variável Y

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{l} n_{\bullet j} y_j \;\rightarrow \text{ média aritmética marginal}$$

$$s_y^2 = \frac{1}{n}\sum_{j=1}^{l} n_{\bullet j}\left(y_j - \bar{y}\right)^2 \;\rightarrow \text{ variância marginal}$$

$$\gamma_{1y} = \frac{\mu_3}{s_y^3} \text{ onde } \mu_3 = \frac{1}{n}\sum_{j=1}^{l} n_{\cdot j}\left(y_j - \bar{y}\right)^3 \;\rightarrow \text{ assimetria}$$

$$\gamma_{2y} = \frac{\mu_4}{s_y^4} - 3 \text{ onde } \mu_4 = \frac{1}{n}\sum_{j=1}^{l} n_{\cdot j}\left(y_j - \bar{y}\right)^4 \;\rightarrow \text{ “achatamento”}$$

☞ Características amostrais condicionais de $X \mid y_j$, $\forall_j$

$$\bar{x}\,|_j = \frac{1}{n_{\bullet j}}\sum_{i=1}^{k} n_{ij} x_i \text{ e } s_{x|j}^2 = \frac{1}{n_{\bullet j}}\sum_{i=1}^{k} n_{ij}\left(x_i - \bar{x}\,|_j\right)^2 \;\rightarrow \text{ média e variância}$$

$$\gamma_{1x|_j} = \frac{\mu_3}{s_{x|_j}^3} \text{ onde } \mu_3 = \frac{1}{n_{\cdot j}}\sum_{i=1}^{k} n_{ij}\left(x_i - \bar{x}|_j\right)^3 \;\rightarrow \text{ assimetria}$$

$$\gamma_{2x|_j} = \frac{\mu_4}{s_{x|_j}^4} - 3 \text{ onde } \mu_4 = \frac{1}{n_{\cdot j}}\sum_{i=1}^{k} n_{ij}\left(x_i - \bar{x}|_j\right)^4 \;\rightarrow \text{ “achatamento”}$$

☞ Características amostrais condicionais de $Y \mid x_i$, $\forall_i$

$$\bar{y}\,|_i = \frac{1}{n_{i\bullet}}\sum_{j=1}^{l} n_{ij} y_j \text{ e } s_{y|i}^2 = \frac{1}{n_{i\bullet}}\sum_{j=1}^{l} n_{ij}\left(y_j - \bar{y}\,|_i\right)^2 \;\rightarrow \text{ média e variância}$$

$$\gamma_{1y|_i} = \frac{\mu_3}{s_{y|_i}^3} \text{ onde } \mu_3 = \frac{1}{n_{i\cdot}}\sum_{j=1}^{l} n_{ij}\left(y_j - \bar{y}|_i\right)^3 \;\rightarrow \text{ assimetria}$$

$\gamma_{2y|_i} = \frac{\mu_4}{s^4_{y|_i}} - 3$ onde $\mu_4 = \frac{1}{n_{i.}} \sum_{j=1}^{l} n_{ij} \left(y_j - \bar{y}|_i \right)^4$ $\rightarrow$ "achatamento"

☞ Covariância amostral de duas variáveis *X* e *Y*

$$\text{cov}(x, y) = \frac{1}{n} \sum_i \sum_j n_{ij} \left(x_i - \bar{x} \right) \left(y_j - \bar{y} \right) = \frac{1}{n} \sum_i \sum_j n_{ij} x_i y_j - \bar{x}\bar{y}$$

☞ Relações entre as características marginais e condicionais

$\bar{x} = \frac{1}{n} \sum_i \sum_j n_{ij} x_i = \frac{1}{n} \sum_{j=1}^{l} n_{\bullet j} \bar{x}|_j$ e $\bar{y} = \frac{1}{n} \sum_i \sum_j n_{ij} y_j = \frac{1}{n} \sum_{i=1}^{k} n_{i\bullet} \bar{y}|_i$ $\rightarrow$ a média marginal é igual à média ponderada das médias condicionais

$$s_x^2 = \frac{1}{n} \sum_{j=1}^{l} n_{\bullet j} \left(\bar{x}|_j - \bar{x} \right)^2 + \frac{1}{n} \sum_{j=1}^{l} n_{\bullet j} s^2_{x|j} \quad \text{e} \quad s_y^2 = \frac{1}{n} \sum_{i=1}^{k} n_{i\bullet} \left(\bar{y}|_i - \bar{y} \right)^2 + \frac{1}{n} \sum_{i=1}^{k} n_{i\bullet} s^2_{y|i}$$

A variância marginal (variância total de cada variável) é igual à variância das médias condicionais – $s^2_{\bar{x}|_j}$ e $s^2_{\bar{y}|_i}$, mais a média das variâncias condicionais, devidamente ponderadas.

Desta forma, pode-se interpretar que a variância total (variância marginal) é a soma da variância explicada pela regressão e a variância residual (variância que não é explicada pela regressão).

Com estas definições está-se em condições de definir o coeficiente de correlação global e o coeficiente de correlação linear.

O coeficiente de correlação global, η^2_{**}, representa a proporção da variância explicada pela curva, isto é,

$$\eta^2_{**} = \frac{\text{variância explicada}}{\text{variânciatotal}} = 1 - \frac{\text{variânciaresidual}}{\text{variânciatotal}}, \; 0 \le \eta^2_{**} \le 1.$$

A correlação de Y dado X é portanto dada por

$$\eta^2_{Y|X} = \frac{s^2_{\bar{y}|_i}}{s_y^2} = \frac{\sum_{i=1}^{k} n_{i\bullet} \left(\bar{y}|_i - \bar{y} \right)^2}{\sum_{j=1}^{l} n_{\bullet j} \left(y_j - \bar{y} \right)^2} \text{ ou } \eta^2_{Y|X} = 1 - \frac{\frac{1}{n} \sum_{i=1}^{k} n_{i\bullet} s^2_{y|i}}{s_y^2},$$

e de forma análoga a correlação de *X* dado Y

$$\eta^2_{X|Y} = \frac{s^2_{\bar{x}|_j}}{s_x^2} = \frac{\sum_{j=1}^{l} n_{\bullet j} \left(\bar{x}|_j - \bar{x} \right)^2}{\sum_{i=1}^{k} n_{i\bullet} \left(x_i - \bar{x} \right)^2} \text{ ou } \eta^2_{X|Y} = 1 - \frac{\frac{1}{n_{\bullet\bullet}} \sum_{j=1}^{l} n_{\bullet j} s^2_{x|j}}{s_x^2},$$

O coeficiente de correlação linear amostral, r_{xy} é dado por:

$$\text{r}_{\text{xy}} = \frac{\text{variânciaexplicada}}{\text{variânciatotal}} = \frac{\text{cov}(x, y)}{s_x s_y}, \; -1 \le r_{xy} \le 1.$$

Este coeficiente é visto como uma medida do grau de ligação ou intensidade da relação linear entre a variável *X* e a variável *Y*. Indica-nos também a direcção dessa relação.

Notas:

1. Como $r_{xy} = r_{yx}$, no que se segue, designa-se simplesmente por r. Além disso, é de ter em conta que se queremos tratar os dados considerando classes usa-se as fórmulas atrás mencionadas para a covariância e para os desvios padrões, caso contrário usa-se a definição usual, ou seja,

$$r = \frac{\frac{1}{n}\sum_{w=1}^{n}(x_w - \bar{x})(y_w - \bar{y})}{\sqrt{\frac{1}{n}\sum_{w=1}^{n}(y_w - \bar{y})^2}\sqrt{\frac{1}{n}\sum_{w=1}^{n}(x_w - \bar{x})^2}}.$$

2. Com o η^2_{**} obtém-se toda a correlação, linear ou não.
3. Sabendo que η^2_{**} nos dá o valor da correlação global e r apenas o valor da correlação linear facilmente se constata que $0 \le r^2 \le \eta^2_{**} \le 1$.
4. Do ponto de vista prático, o tamanho das classes tem fortes consequências. Se escolhermos classes "finas" numa variável, temos grande precisão nessa variável, mas o número de pontos pode ser pequeno. Assim, a variância da estimativa da média de uma variável condicionada por essa classe "fina" pode ser muito grande. Fica muito dependente de *outliers* ou de pontos influentes.
5. As estimativas da curva de regressão têm as propriedades da média:
   - ✓ A classes "finas" correspondem poucos valores e portanto sujeitas a flutuações de amostragem levando a uma pouca precisão em $\tilde{y}_i$, ou seja, tem-se uma boa precisão na variável condicionante e uma má precisão na variável condicionada;
   - ✓ A classes "grossas" correspondem muitos valores e portanto pouca precisão no $\tilde{x}_i = x_i + h$, ou seja, tem-se uma má precisão na variável condicionante e uma boa precisão na variável condicionada.

2.2. A Recta de Regressão pelo Método dos Mínimos Quadrados

O termo "regressão" foi introduzido por Sir Francis Galton em 1885 e Gauss "descobriu" no início do séc XIX o método dos mínimos quadrados. Desde então, a análise de regressão linear foi muito desenvolvida, e neste trabalho vai-se cingir apenas a descrever os resultados mais importantes e que são necessários para a percepção do que é apresentado. Para mais pormenores, a teoria desenvolvida pode ser encontrada, por exemplo, em Tomassone *et al* (1983) e/ou Kutner *et al* (2004).

Consideremos então o modelo de regressão base, que é o modelo linear, no qual temos apenas uma variável independente, isto é,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1 \ldots n \qquad (1)$$

onde

Y_i é o valor da variável dependente na i-ésima observação;

β_0 e β_1 são constantes;

X_i é o valor da variável independente na i-ésima observação, que se admite conhecida.

Para este modelo tem-se os seguintes pressupostos:

ε_i é o erro aleatório com média $E(\varepsilon_i) = 0$ e variância $\sigma^2(\varepsilon_i) = \sigma^2$;

A covariância entre pares de erros é $\sigma(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$, pois os termos erros são não correlacionados;

$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Kutner *et al* (2004) prova que no modelo de regressão linear simples, a variável Y_i provém de uma distribuição em que as médias são $E(Y_i) = \beta_0 + \beta_1 X_i$, as variâncias são σ^2 e quaisquer dois valores Y_i e Y_j, $i \neq j$, são não correlacionados.

Este modelo tem parâmetros desconhecidos e que portanto é necessário estimar, são eles β_0, β_1 e σ^2. Para se encontrar então os estimadores destes parâmetros pode utilizar-se o método de estimação pelos mínimos quadrados (MMQ).

O MMQ considera que deve ser mínima a soma das distâncias na vertical entre os valores observados para Y e a recta ajustada. Assim, considera os desvios $Y_i - E(Y_i)$, eleva ao quadrado estes desvios e aplica o somatório obtendo-se:

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \quad (2)$$

As estimações de β_0 e β_1, b_0 e b_1, são obtidas aquando da minimização de Q para a amostra $(X_1, Y_1), \cdots, (X_n, Y_n)$. Ou seja, tem-se que diferenciar (2) relativamente a β_0 e β_1, e obtemos as seguintes condições de estacionaridade

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i)$$

De seguida, iguala-se a zero e obtém-se o minimizante que é único.

Da resolução das equações normais tem-se:

$$\begin{cases} \sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i Y_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2 \end{cases}$$

sendo a solução dada por:

$$\begin{cases} \hat{\beta}_0 = \dfrac{\sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i}{n} \\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} X_i Y_i - \bar{Y} \sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i^2 - \bar{X} \sum_{i=1}^{n} X_i} \end{cases}$$

As estimativas b_0 e $b_1$são dadas por:

$$\begin{cases} b_0 = \bar{y} - b_1\bar{x} \\ b_1 = \dfrac{\sum_{i=1}^{n} x_i\, y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \end{cases}$$

As variâncias e covariância dos estimadores, $\hat{\beta}_0$ e $\hat{\beta}_1$, que são não enviesados para os casos em que a função de distribuição da variável dependente é gaussiana, são dadas pelas seguintes equações, Tomassone *et al* (1983):

$$V(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$$

$$V(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{X}}{\sum(X_i - \bar{X})^2} \times \sigma^2$$

onde, uma estimativa para σ^2 pode ser $s^2 = \frac{\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i^0)^2}{n-2}$.

Os intervalos de confiança a $100(1-\alpha)\%$, para os valores de β_0 e de β_1, podem ser determinados por:

- $b_0 \pm t_{n-2;\alpha/2} \times \sqrt{v(\hat{\beta}_0)}$;
- $b_1 \pm t_{n-2;\alpha/2} \times \sqrt{v(\hat{\beta}_1)}$;

onde v designa a estimativa da variância do estimador.

A recta de regressão permite-nos, conhecendo um valor de *X*, predizer o valor de *Y*, no entanto, esta predição pode ser de dois tipos. Podemos querer predizer um valor médio ou um valor individual. Este último caso é o caso anterior (1). No caso em que queremos predizer um valor médio o modelo a usar é

$$\mu_Y(x) = \beta_0 + \beta_1 x$$

com parâmetros desconhecidos e portanto uma sua estimativa é dada por

$$\hat{\mu}_Y(x) = b_0 + b_1 x.$$

Dado um valor fixo x_i predizer um valor médio ou predizer um valor individual conduz ao mesmo resultado, mas os erros associados a cada predição não são os mesmos.

Para a predição de um valor médio, o erro padrão associado à estimativa no ponto x_i resulta de,

$$\hat{\mu}_Y(x_i) - \mu_Y(x_i) = b_0 + b_1 x_i - (\beta_0 + \beta_1 x_i)$$

e a variância do estimador correspondente é

$$V(\hat{\mu}_Y(x_i) - \mu_Y(x_i)) = V(\hat{\mu}_Y(x)) = \sigma^2\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right] \qquad (3)$$

Para a predição de um valor individual, o erro padrão associado à estimativa no ponto x_i resulta de,

$$E(Y_i) - Y_i = b_0 + b_1 x_i - (\beta_0 + \beta_1 x_i + \varepsilon_i)$$

e a variância do estimador correspondente é

$$V[E(Y_i) - Y_i] = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right].$$

Como se pode constatar a variância para um valor individual é simplesmente deduzida pela adição de σ^2 a (3) pois o novo erro não é correlacionado com os anteriores.

Note-se que, se para algum *i*,

- ✓ $x_i = \bar{x}$, o erro é mínimo e tem-se

$$V(\hat{\mu}_Y(x_i) - \mu_Y(x_i)) = \sigma^2 \left[\frac{1}{n} + \frac{0}{\sum (x_i - \bar{x})^2}\right] = \frac{\sigma^2}{n};$$

$$V[E(Y_i) - Y_i] = \sigma^2 \left[1 + \frac{1}{n} + \frac{0}{\sum (x_i - \bar{x})^2}\right] = \sigma^2 + \frac{\sigma^2}{n}.$$

  Se *n* for muito grande $\sigma^2 \approx \sigma^2 + \frac{\sigma^2}{n}$ e se *n* for muito pequeno $\frac{\sigma^2}{n} \ll \sigma^2 + \frac{\sigma^2}{n}$.

- ✓ $x_i \gg \bar{x}$ ou $x_i \ll \bar{x}$ o termo $(x_i - \bar{x})^2$ toma valores maiores e portanto as variâncias também.

Daqui pode-se concluir que quanto mais distante estiver x_i do centro da nuvem de pontos maior será o erro, menos precisa será a estimativa quer para um valor médio, quer para um único individuo.

Este efeito pode ser observado numa representação gráfica do intervalo de confiança para cada valor de *x*. No entanto, pode-se desde já concluir que o I.C. para predizer um valor individual vai ser maior do que o I.C. para a média.

O I.C. para um valor médio é determinado da seguinte forma,

$$\hat{\mu}_Y(x) \pm \sqrt{v(\hat{\mu}_Y(x_i) - \mu_Y(x_i))} \times t(n-2, 1-\alpha/2)$$

onde v é uma estimativa da variância e $t(n-2, 1-\alpha/2)$ é o quantil $1-\alpha/2$ da distribuição t-student.

Para se obter a região de confiança para toda a recta de regressão a distribuição t-student deve ser substituída pela distribuição de Fisher-Snedecor, da seguinte forma,

$$\mu_Y(x) \pm \sqrt{v(\hat{\mu}_Y(x_i) - \mu_Y(x_i))} \times [2F(n-2, 1-\alpha/2)]^{1/2}$$

em que v é uma estimativa da variância e $F(n-2, 1-\alpha/2)$ é o quantil $1-\alpha/2$ da distribuição Fisher-Snedecor.

Da mesma forma se pode deduzir a um nível de significância α, um I.C. para um valor individual, ou seja,

$$E(Y_i) \pm \sqrt{v[E(Y_i) - Y_i]} \times t(n-2, 1-\alpha/2);$$

$$Y_i \pm \sqrt{v[E(Y_i) - Y_i]} \times [2F(n-2, 1-\alpha/2)]^{1/2}.$$

2.3. Distribuição Binormal

Um dos objectivos neste trabalho, será aplicar a teoria em amostras de variáveis com distribuições normais. Para este estudo serão geradas variáveis aleatórias binormais através do método de Box Muller (ver anexo A).

Considerando que estamos a estudar o caso em que temos a distribuição de duas variáveis agrupadas em classes, e que a relação entre estas é linear, então somos levados a concluir que a curva de regressão será uma recta. Grais (1974) faz o ajustamento da recta de regressão através das médias móveis e chega à conclusão que:

- A recta de ajustamento de Y dado X, $Y-\mu_Y=\beta_1(X-\mu_X)$, passa pelo ponto médio (μ_X,μ_Y) e o valor de b_1 é dado por $\beta_1=\dfrac{\operatorname{cov}(X,Y)}{\sigma_X^2}=\rho\dfrac{\sigma_Y}{\sigma_X}$;

- A recta de ajustamento de X dado Y, $X-\mu_X=\dfrac{1}{\beta_1}(Y-\mu_Y)$, passa pelo ponto médio (μ_X,μ_Y) e o valor de $\frac{1}{\beta_1}$ é dado por. $\dfrac{1}{\beta_1}=\dfrac{\operatorname{cov}(X,Y)}{\sigma_Y^2}=\rho\dfrac{\sigma_X}{\sigma_Y}$ Pelo que facilmente se verifica que $Y-\mu_Y=\frac{1}{\rho}\frac{\sigma_Y}{\sigma_X}(X-\mu_X)$.

Assim, considerando a expressão da recta de ajustamento $Y=\mu_Y+\rho\dfrac{\sigma_X}{\sigma_Y}(X-\mu_x)$ que nos dá a curva de regressão linear e tendo em conta que $X\approx N(\mu_X,\sigma_X)$ e $Y\approx N(\mu_Y,\sigma_Y)$ tem-se que. $Y|_{X=x}\approx N\left(\mu_Y+\rho\dfrac{\sigma_Y}{\sigma_X}(x-\mu_X),\sigma_Y\sqrt{1-\rho^2}\right)$

Neste trabalho e numa primeira abordagem, será gerado um par binormal de variáveis independentes, ou seja, $\rho=0$ e portanto $\mathrm{Y}-\mu_{\mathrm{Y}}=0$. Numa segunda abordagem, será gerada uma binormal, tal que $\rho\neq 0$ e será considerado que $\mu_{\mathrm{Y}}=\mu_{\mathrm{X}}=0$ e $\sigma_X=\sigma_Y=1$ (se não for o caso reduz-se a este caso por *Standardização).* Desta forma vem que $E(Y|_{X=x})=\rho x$, $V(Y|_{X=x})=1-\rho^2$ e as variáveis serão tais que $X\approx N(0,1)$, $Y\approx N(0,1)$ e $Y\approx N\left(\rho x,\sqrt{1-\rho^2}\right)$.

2.4. ANOVA

A Análise de Variância surgiu com a necessidade de se comparar as médias de dois ou mais grupos de amostras aleatórias e independentes, visto que com a aplicação do teste t-Student, a probabilidade de se cometer o erro Tipo I aumentava consideravelmente. Esta situação deve-se ao facto de quando se efectua cada comparação de pares de médias a probabilidade de se cometer o erro Tipo I não é alterado, mas no conjunto das comparações é aumentado.

Para realizar o teste global são necessários efectuar *m* testes independentes, se considerarmos que temos *m* grupos, com um nível de significância α (probabilidade de erradamente se considerar significativa cada comparação entre duas médias).

Considerando α^* o nível de significância quando efectuamos simultaneamente o conjunto dos m testes independentes temos que:

α^* = probabilidade de rejeitar pelo menos uma das hipóteses nulas verdadeira

= 1 – probabilidade de não rejeitar nenhuma hipótese nula verdadeira

$= 1-(1-\alpha)^m$

Uma primeira metodologia foi proposta por Sir Ronald Fisher – 1935, que abreviadamente corresponde à ANOVA, "**An**alysis **o**f **Va**riance", usual. Esta análise pressupõe que a distribuição da variável em estudo seja normal em cada grupo e que os grupos sejam homogéneos. Para testar se as amostras possuem as médias dos grupos significativamente diferentes, verifica-se se a variância residual é significativamente inferior à variância entre grupos, estudando um coeficiente associado ao quociente das variâncias.

Na regressão, este coeficiente é uma medida do mesmo tipo se as classes forem caracterizadas de forma equivalente.

Se as variáveis independentes em estudo (factores) forem fixadas pelo investigador então tem-se uma ANOVA de efeitos fixos, se existem variáveis que ele não controla então está-se perante uma ANOVA de efeitos aleatórios, ou seja, ou estamos perante o caso em que sabemos que grupos queremos estudar ou de um conjunto de grupos extraímos aleatoriamente um número inferior de grupos. Uma combinação destas duas resulta numa ANOVA de efeitos mistos.

Quando na ANOVA se rejeita a igualdade das médias, é porque existe pelo menos uma delas que é significativamente diferente das outras. Para verificar de que forma elas se diferenciam, duas a duas, usa-se a "Comparação múltipla de médias". Os testes usados neste procedimento são vários, mas os mais usados são Tukey, Bonferroni, Scheffé e LSD. O mesmo pode ser feito na regressão, pois existe um paralelismo entre os parâmetros, como podemos constatar comparando os dois modelos a seguir.

O modelo de efeitos fixos pode ser descrito da seguinte forma:

$$X_{ij} = \mu + \alpha_j + \varepsilon_{ij}, \quad i = 1,...,n_j \text{ e } j = 1,...,m$$

X_{ij} – representa a observação para a i-ésima observação no j-ésimo grupo;

μ – valor esperado da variável X_{ij};

α_j – representa o efeito principal do j-ésimo grupo;

ε_{ij} – são os resíduos ou erros e são assumidos como sendo independentes e tendo distribuição normal com valor esperado zero e variância constante, σ^2.

O modelo a ser considerado na regressão pode ser formulado da seguinte forma:

$$Y_j = f(x_j) + \varepsilon_j$$

em que no cálculo $f(x_j)$ pode ser usada uma média, os resíduos são geralmente normais e numa situação ideal (correlação binormal, por exemplo), para além de média nula, todos os resíduos têm a mesma variância. Muitas vezes, se existir heterocedasticidade,

as diferenças de variâncias podem ser compensadas por pesos w_j (correspondendo aos efeitos fixos) tal que se minimize $\sum w_j \left(y_j^{observado} - y_j^{calculado}\right)^2$,

Com estes modelos e tendo em conta que se pretende comparar as médias de dois ou mais grupos de amostras aleatórias e independentes, a hipótese a ser testada pode ser formulada do seguinte modo:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_m = \mu \qquad \text{(ou seja, o factor não produz efeito)}$$

Na regressão, espera-se que $\mu_j = f\left(x_j\right)$ para uma dada função f em classes fixadas à priori.

Para se testar a hipótese nula tem-se de proceder à estimação da variação dentro dos grupos (VDG) e da variação entre os grupos (VEG).

Cada uma daquelas variações são variáveis assim definidas:

$$VDG = \sum_j \sum_i \left(X_{ij} - \bar{X}_j\right)^2 \quad \text{e} \quad VEG = \sum_j n_j \left(\bar{X}_j - \bar{X}\right)^2,$$

para as quais se verifica $\frac{VDG}{\sigma^2} = \frac{\sum_j \sum_i \left(X_{ij} - \bar{X}_j\right)^2}{\sigma^2} \sim \chi^2_{\sum_j n_j - m}$ quer H_0 seja verdadeira ou não mas, $\frac{VEG}{\sigma^2} = \frac{\sum_j n_j \left(\bar{X}_j - \bar{X}\right)^2}{\sigma^2} \sim \chi^2_{m-1}$, apenas se H_0 é verdadeira.

A partir destas variáveis, pode obter-se quadrados médios dentro dos grupos (QMDG) e os quadrados médios entre grupos (QMEG). Assim, obtemos:

$$QMDG = \frac{VDG}{\sum_j n_j - m} \quad \text{e} \quad QMEG = \frac{VEG}{m-1},$$

que são variáveis independentes.

Então a comparação entre aqueles quadrados médios permite realizar o teste para validar a hipótese H_0 . Sob a validade de H_0 tem-se que,

$$F = \frac{QMEG}{QMDG} = \frac{\frac{VEG/\sigma^2}{m-1}}{\frac{VDG/\sigma^2}{\sum_j n_j - m}} \sim \frac{\chi^2_{m-1}/(m-1)}{\chi^2_{\sum_j n_j - m}/\sum_j n_j - m}$$

e como o quociente de Qui-Quadrados independentes divididos pelo respectivo número de graus de liberdade tem distribuição F de Snedecor, vem que, $F \sim \mathcal{F}_{(m-1, \sum_j n_j - m)}$. A prova destes resultados pode ser vista, por exemplo, em Pestana e Velosa (2006).

Para a correlação basta-nos estabelecer a relação que existe entre a estatística F e o valor do coeficiente η^2_{**} para efectuar um paralelismo entre as estatísticas.

Seja m o número de grupos em x. O coeficiente de correlação global pode descrito como sendo, $\eta^2_{**} = \frac{V_{exp}}{V_{exp}+V_{res}} = \frac{1}{\frac{V_{exp}+V_{res}}{V_{exp}}} = \frac{1}{1+\frac{V_{res}}{V_{exp}}}$,

$\frac{1}{\eta^2_{**}} = 1 + \frac{V_{res}}{V_{exp}} \Leftrightarrow \frac{1}{\eta^2_{**}} - 1 = \frac{V_{res}}{V_{exp}} \Leftrightarrow \frac{V_{exp}}{V_{res}} = \frac{1}{\frac{1}{\eta^2_{**}}-1} \Leftrightarrow \frac{V_{exp}}{V_{res}} = \frac{\eta^2_{**}}{1-\eta^2_{**}} > 0$.

O coeficiente de Snedecor pode vir escrito em função das variâncias explicadas pois:

$VEG \rightarrow V_{exp}$ e $VDG \rightarrow V_{res}$, ou seja, $F = \left(\frac{\sum n_j - m}{m-1}\right) \times \frac{VEG}{VDG} = \left(\frac{\sum n_j - m}{m-1}\right) \times \frac{V_{exp}}{V_{res}}$,

$$F = \left(\frac{\sum n_j - m}{m-1}\right) \times \frac{\eta_{**}^2}{1-\eta_{**}^2} \Leftrightarrow \frac{\eta_{**}^2}{1-\eta_{**}^2} = \frac{m-1}{\sum n_j - m} F \Leftrightarrow \eta_{**}^2 = \frac{m-1}{\sum n_j - m} F \times (1 - \eta_{**}^2) \Leftrightarrow$$

$$\Leftrightarrow \eta_{**}^2 + \frac{m-1}{\sum n_j - m} F \times \eta_{**}^2 = \frac{m-1}{\sum n_j - m} F \Leftrightarrow \eta_{**}^2 = \frac{\frac{m-1}{\sum n_j - m} F}{\frac{m-1}{\sum n_j - m} F + 1}$$

Quando se agrupa num caso concreto de uma regressão, temos de ter em atenção que podem ficar classes sem elementos e podem ficar classes muito desequilibradas. Logo vamos ter *F*'s com graus de liberdade diferentes, o que sugere que numa regressão vamos ter uma mistura de F's diferentes.

A estatística F detecta se as médias são ou não são todas iguais! Se forem todas iguais o $\eta_{**}^2 = 0$, caso contrário $\eta^2 \neq 0$, tem-se que fazer as previsões $\hat{y}(x)$ usando médias condicionais.

Recorde-se que conhecido um x determinar $\hat{y}(x)$ pode ser feito de duas maneiras:

1. Dado por uma recta ou por uma outra curva analítica (muitas alternativas);
2. Média condicionada.

Qual a melhor? A que tiver menor variância e de preferência com distribuição conhecida.

2.5. Momentos cruzados de terceira ordem (Dodge)

Como se pode constatar, a ANOVA e a Correlação-Regressão têm uma metodologia muito parecida. Nestas metodologias é crucial que as distribuições nas classes sejam gaussianas, mas é robusto a pequenas assimetrias.

Para a correlação ser diferente da ANOVA teria de haver produtos cruzados x por y pois só assim os valores das covariâncias seriam diferentes. De facto, isto normalmente acontece, levando a pensar que talvez seja por isso que o Dodge and Rousson (1999) calculem momentos cruzados para avaliar a linearidade.

Num dos artigos de Dodge and Rousson (1999), este começa por considerar um modelo de regressão linear e recorrendo à covariância entre duas variáveis e à covariancia entre três variáveis, *covtri* e chega à seguinte expressão:

$$\rho_{XY}^3 = \frac{\gamma_{1Y}}{\gamma_{1X}} \tag{4}$$

Como iremos ver, ela só é válida se Y for a variável resposta, X a variável independente e a correlação entre as duas variáveis for linear.

Considerando a recta de regressão na forma já introduzida anteriormente

$$Y - \bar{Y} = \rho_{XY} \frac{\sigma_Y}{\sigma_X} (X - \bar{X})$$

onde $b_1 = \rho_{XY} \frac{\sigma_Y}{\sigma_X} \Longrightarrow \rho_{XY} = b_1 \frac{\sigma_X}{\sigma_Y}$.

Como já é sabido, $\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \Longrightarrow cov(X,Y) = \rho_{XY} \sigma_X \sigma_Y$, logo $cov(X,Y) = b_1 \sigma_X^2$.

Dodge and Rousson (1999) define a covariância entre três variáveis como sendo

$$cov3(X,Y,Z) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])(Z - \mathrm{E}[Z])]$$

e a correlação entre três variáveis como sendo

$$\rho_{XYZ} = \frac{cov3(X,Y,Z)}{\sigma_X \sigma_Y \sigma_Z}.$$

Com estas definições está-se em condições de chegar à expressão (4) e vai-se proceder à sua prova.

Através da definição de covariância entre três variáveis e considerando a equação da recta de regressão (1), pode-se concluir que:

$$\begin{aligned}
cov3(X,X,Y) &= \mathrm{E}[(X-\bar{X})(X-\bar{X})(Y-\bar{Y})]= \\
&= \mathrm{E}[(X-\bar{X})(X-\bar{X})(b_0+b_1X-b_0-b_1\bar{X})]= \\
&= b_1\mathrm{E}[(X-\bar{X})(X-\bar{X})(X-\bar{X})]= \\
&= b_1 cov3(X,X,X) \\
cov3(X,Y,Y) &= \mathrm{E}[(X-\bar{X})(Y-\bar{Y})(Y-\bar{Y})]= \\
&= \mathrm{E}[(X-\bar{X})b_1(X-\bar{X})b_1(X-\bar{X})] \\
&= b_1^2\mathrm{E}[(X-\bar{X})(X-\bar{X})(X-\bar{X})] \\
&= b_1^2 cov3(X,X,X) \\
cov3(Y,Y,Y) &= b_1^3\mathrm{E}[(X-\bar{X})(X-\bar{X})(X-\bar{X})] \\
&= b_1^3 cov3(X,X,X)
\end{aligned}$$

Quanto aos coeficientes de correlação tem-se que:

$$\rho_{XXX} = \frac{cov3(X,X,X)}{\sigma_X\sigma_X\sigma_X} = \Upsilon_{1X}$$

pois por definição $\Upsilon_{1X} = \frac{\mathrm{E}[(X-\bar{X})^3]}{\sigma_X^3}$ logo $\Upsilon_{1X} = \frac{cov3(X,X,X)}{\sigma_X^3}$.

Assim sendo, $\rho_{XXY} = \frac{b_1 cov3(X,X,X)}{\sigma_X\sigma_X\sigma_Y} = \frac{b_1\Upsilon_{1X}\sigma_X^3}{\sigma_X^2\sigma_Y} = b_1\Upsilon_{1X}\frac{\sigma_X}{\sigma_Y}$

$$\rho_{XYY} = \frac{b_1^2 cov3(X,X,X)}{\sigma_X\sigma_Y\sigma_Y} = \frac{b_1^2\Upsilon_{1X}\sigma_X^3}{\sigma_X\sigma_Y^2} = b_1^2\Upsilon_{1X}\frac{\sigma_X^2}{\sigma_Y^2}$$

$$\rho_{YYY} = \frac{b_1^3 cov3(X,X,X)}{\sigma_Y\sigma_Y\sigma_Y} = \frac{b_1^3\Upsilon_{1X}\sigma_X^3}{\sigma_Y^3} = b_1^3\Upsilon_{1X}\frac{\sigma_X^3}{\sigma_Y^2}.$$

Conjugando a covariancia e correlação de três variáveis vem que,

$cov3(Y,Y,Y) = b_1^3\, cov3(X,X,X) \Leftrightarrow$

$\Leftrightarrow \Upsilon_{1Y} = \frac{b_1^3\sigma_X^3\Upsilon_{1X}}{\sigma_Y^3} \Leftrightarrow \frac{\gamma_{1Y}}{\Upsilon_{1X}} = b_1^3\frac{\sigma_X^3}{\sigma_Y^3} \Leftrightarrow \frac{\gamma_{1Y}}{\Upsilon_{1X}} = \left(b_1\ \frac{\sigma_X}{\sigma_Y}\right)^3 \Leftrightarrow \frac{\gamma_{1Y}}{\Upsilon_{1X}} = \rho_{XY}^3$ (c.q.d.).

Como $|\rho_{XY}^3| \leq 1$ a escolha para a variável dependente, Y, será aquela em que $\left|\frac{\gamma_{1Y}}{\gamma_{1X}}\right| \leq 1$.

A expressão (4) pode também ser usada para verificar se o modelo linear se ajusta aos dados, ou seja, se a parte não linear não é significativa. Como a expressão só é válida se a correlação entre as duas variáveis é linear, basta analisar $\left|\frac{\gamma_Y}{\gamma_X}\right|$ e verificar se este se afasta muito do valor 1.

2.6. *Outliers* e Pontos Influentes

Na análise de regressão podem existir observações que se distinguem pelo comportamento da maioria dos outros valores e a que chamamos *outliers* ou observações influentes.

Em particular, numa regressão gaussiana, há mais pontos *outliers* do que pontos influentes onde ambos exigem ser pontos extremos locais, mas para se ser influente tem que ser extremo local simultaneamente nas variáveis x e y e para ser *outlier* tem de ser extremo local numa e não o ser na outra.

A maneira mais simples, e numa primeira abordagem ao problema, a detecção destas observações pode ser feita através da análise do diagrama de dispersão das próprias variáveis e em seguida dos resíduos de *Y* versus X_i.

Numa segunda abordagem, será melhor considerar diferentes técnicas estatísticas analíticas para detectar estas observações.

Para a identificação de *outliers* pode recorrer-se à matriz "chapéu" (hat matrix), $H = X'(X'X)^{-1}X$, de dimensão $n \times p$, onde p é o número de variáveis explicativas e n a dimensão de cada uma das variáveis. Através dela pode-se estudar o tamanho dos "resíduos apagados studentardizados". Analiticamente estes resíduos podem ser calculados através da fórmula $t_i = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}} = e_i \left[\frac{n-p-1}{SSE(1-h_{ii})-{e_i}^2}\right]^{1/2}$, onde e_i são os resíduos, $MSE_{(i)}$[1] é o erro quadrático médio quando o $i-ésimo$ caso é omitido na obtenção da regressão, SSE é a variabilidade não explicada pelo modelo mas sim pelos erros e h_{ii} são os elementos da diagonal da matriz "chapéu". Segundo Dodge and Rousson (1999), a i-observação é um *outlier* se $|t_i| \geq 2$.

Aos valores h_{ii} dá-se o nome de *leverage* e traduz-se numa medida de distância entre os valores do $i-ésimo$ caso e a média dos n casos, ou seja, média das variáveis em X. Se o $i-ésimo$ caso é um *outlier* então possui um valor de *leverage* grande. Considera-se um valor grande se $h_{ii} > 2 \times \frac{p}{n}$, onde p é o número de parâmetros da regressão, incluindo o associado à ordenada na origem.

A matriz "chapéu" pode ainda ser usada para ver se uma nova observação vai influenciar ou não na inferência. Como este estudo não é um dos objectivos do trabalho não será aqui abordado, mas pode ser visto mais pormenorizadamente em Kutner (2004).

Após a identificação dos valores extremos interessa saber se estas observações influenciam ou não na determinação da curva de regressão. Para esta identificação existem três medidas de influência que são usadas com frequência, são elas influência de apenas um valor ajustado – DFFITS, influência de todos os valores ajustados – Distância Cook e influência dos coeficientes da regressão – DFBETAS. Para o cálculo de cada medida omite-se a observação que se pensa ser o valor extremo.

No cálculo da medida DFFITS pretende-se verificar se o caso *i* que pensamos ser influente tem interferência na sua predição. Assim, determina-se o valor da medida que é dado por:

[1] Optou-se por usar a notação usual neste contexto.

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}} = t_i \left(\frac{h_{ii}}{1 - h_{ii}}\right)^{1/2}$$

onde $\hat{Y}_i$ é o valor predito da observação *i* quando estamos a considerar a regressão com as *n* observações e $\hat{Y}_{i(i)}$ é o valor predito da observação *i* quando excluímos o suposto caso influente *i*. Considera-se que a observação i é influente se o valor absoluto de DFFITS for superior a 1 para pequenas e médias amostras e se for superior a $2\sqrt{p/n}$ para amostras grandes.

No caso da medida distância de *Cook,* esta determina a influência que uma observação tem na predição de todos os outros valores sem tomar em atenção o sinal desta influência. Assim, esta medida é dada por:

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{pMSE} = \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1 - h_{ii})^2}\right]$$

onde $\hat{Y}_{j(i)}$ é o valor predito da observação *j* quando excluímos a observação *i*. Verificámos que existe interferência da observação *i* na regressão se o valor do percentil estiver perto de 50% ou for superior.

A medida de influência DFBETAS vê a interferência que a observação *i* tem sobre o cálculo dos valores dos coeficientes de regressão. Esta medida é muito similar à medida DFFITS, mas em vez de usar os valores preditos com e sem a observação *i*, utiliza os valores dos coeficientes de regressão obtidos com e sem a observação *i*, que se designam por b_k e $b_{k(i)}$, respectivamente. Assim sendo, o seu cálculo é feito através da seguinte expressão:

$$(DFBETAS)_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)} c_{kk}}}$$

onde c_{kk} é o $k - ésimo$ elemento da matriz $(X'X)^{-1}$. Um elevado valor desta medida leva-nos a concluir que a observação *i* tem uma forte influência no cálculo do $k - ésimo$ coeficiente de correlação. Considera-se que esta influência é significativa se o módulo da medida exceder 1 no caso de amostras pequenas e médias ou exceder $2/\sqrt{n}$ para amostras grandes.

Estas metodologias não são adaptadas a regressões não gaussianas que são o caso de muitas regressões não lineares.

2.7. Métodos Alternativos

O método dos mínimos quadrados para a estimação dos coeficientes de uma regressão linear, apesar de ser o método mais utilizado pela sua fácil implementação e por constar na maioria dos softwares estatísticos, é extremamente sensível a valores extremos (no *Y* e pontos influentes no *X*) da variável dependente e heterocedasticidade da amostra, pelo que outros estimadores mais robustos foram elaborados para modelos lineares. Assim, no sentido de corrigir estes problemas, pode-se recorrer a estimadores não paramétricos, como por exemplo, a Regressão dos Mínimos Desvios Absolutos, MDA, que por sua vez é um caso particular da Regressão Quantílica.

Este recurso a estimadores mais robustos aquando da observação de uma variância dos erros bastante elevada, deve-se ao facto de se pressupor que a linearidade na variável *Y* e o não enviesamento são características do MMQ em vez de se pensar que são pressupostos que devem de ser validados.

Laplace, em 1818, provou que para o modelo de regressão bivariado que passa na origem, o estimador dos coeficientes de regressão MDA tem uma variância inferior à obtida pelo MMQ.

2.7.1. Regressão MDA

A Regressão dos Mínimos Desvios Absolutos, MDA, muitas vezes conhecida como regressão mediana, é um método que consiste na estimação dos coeficientes da regressão através da minimização da soma dos valores absolutos dos erros residuais, isto é, minimizar

$$\sum |Y_i - (b_0 + b_1 X_i)|$$

O conceito do MDA não é assim mais difícil do que o de MMQ e tem a vantagem do valor dos resíduos traduzir melhor o desvio do ponto (x_i, y_i) da curva de regressão $E(Y_i) = b_0 + b_1 X_i$. Este método foi muitas vezes preterido em detrimento do MMQ pela sua pesada computação. Hoje em dia isto deixou de ser um problema e por isso o MDA começou a ser mais usado. No entanto, o cálculo dos estimadores não é feito através de fórmulas existentes, mas sim através de um algoritmo.

Na regressão MDA a recta passa por dois pontos da amostra e por isso o procedimento começa com um dado ponto da amostra (x_0, y_0) e procura qual o ponto, dos restantes da amostra, por onde passa a recta com menores desvios absolutos. Depois de encontrado, este passa a ser o ponto inicial e repete-se o procedimento. O algoritmo termina quando se voltar a ter um mesmo ponto inicial, ou seja, a melhor recta voltar a ser a anterior.

De seguida, apresenta-se como o algoritmo funciona.

- ✓ Dado (x_0, y_0) e para cada (x_i, y_i), calcular o declive $\frac{y_i - y_0}{x_i - x_0}$ que passa pelos dois pontos;
- ✓ Se $x_i = x_0$ para algum *i*, o declive não é calculado e o ponto ignorado;
- ✓ Reindexar os pontos de forma a que $\frac{y_1 - y_0}{x_1 - x_0} \leq \frac{y_2 - y_0}{x_2 - x_0} \leq \cdots \leq \frac{y_n - y_0}{x_n - x_0}$;
- ✓ Seja $T = \sum |x_i - x_0|$. Encontrar o índice *k* que satisfaz

$$|x_1 - x_0| + \cdots + |x_{k-1} - x_0| < \frac{1}{2} T$$

$$|x_1 - x_0| + \cdots + |x_{k-1} - x_0| + |x_k - x_0| < \frac{1}{2} T$$

- ✓ A melhor recta que passa por (x_0, y_0) é $Y = b_0^* + b_1^* X$, onde

$$b_1^* = \frac{y_k - y_0}{x_k - x_0} \text{ e } b_0^* = y_0 - b_1^* x_0.$$

Neste algoritmo, ocasionalmente, ocorrem dois factos que provocam problemas técnicos. São eles a não unicidade e a degeneração. A não unicidade implica que a melhor recta de regressão, que passa por um ponto, não é única e a degeneração implica

que a melhor recta que passa por esse ponto também passa por dois ou mais pontos da amostra. Quando isto acontece, a má escolha do ponto que irá continuar o algoritmo, pode levar a que este entre em ciclo ou resultar numa recta que não é a melhor curva de regressão pelo MDA. Estes casos podem ser vistos no calculo do índice *k* quando temos uma igualdade em vez de desigualdade ou quando temos dois declives iguais. Para contornar estes problemas Dodge and Rousson (1999) indicam uma solução, que não será aqui apresentada.

As provas do porquê deste algoritmo e de que a recta passa realmente por dois pontos da amostra pode ser vistas também em Dodge and Rousson (1999).

Além destes dois problemas, Koenker and Bassett (1978) mencionam que este método tem uma outra desvantagem, que é o facto de puder ser afectado por uma única observação, ou seja, possui um ponto de ruptura (break point) de $1/n$. É através deste ponto de ruptura que se quantifica a sensibilidade do MMQ, Giloni *et al* (2005), e numa tentativa de melhorar este valor foi desenvolvida uma nova técnica chamada e Regressão MDA Pesada, que através de pesos adequados leva a um acréscimo do valor do ponto de ruptura e por conseguinte a uma maior robustez dos estimadores MDA. Esta técnica não será aqui abordada mas pode ser vista em Giloni *et al* (2005).

2.7.2. Regressão Quantílica

A Regressão Quantílica permite analisar a associação entre a variável dependente e variáveis independentes nos diversos quantis da distribuição condicional, tendo-se assim um estudo que não se cinge ao caso da previsão de um valor médio.

Algumas das vantagens apontadas por Koenker and Bassett (1978) são as seguintes:

- ✓ A regressão quantílica é usada quando a distribuição não é gaussiana, particularmente quando não é simétrica. No entanto se a distribuição for gaussiana a regressão quantílica consegue ser relativamente eficiente, enquanto que o estimador dos mínimos quadrados é muito pobre em muitas distribuições não gaussianas principalmente quando a distribuição possui caudas longas;
- ✓ Os erros não terem distribuição normal e/ou não serem homocedásticos proporciona a obtenção de estimadores de regressão quantílica mais eficientes do que os obtidos por regressão dos mínimos quadrados;
- ✓ A regressão quantílica é mais robusta a *outliers*;
- ✓ É uma regressão mais informativa pois permite obter uma regressão em cada quantil de interesse.

Neste trabalho, será exposta a teoria aplicada ao caso bidimensional. Para mais pormenores consultar Koenker and Bassett (1978).

A regressão quantílica, para o caso de um modelo de regressão linear (1), é obtida minimizando uma soma, pesada assimetricamente, dos valores absolutos dos erros, ou seja, encontrando a solução de:

$$\min\left(\sum_{y_i \geq b_0+b_1x_i} \theta|y_i - b_0 - b_1x_i| + \sum_{y_i \leq b_0+b_1x_i} (1-\theta)|y_i - b_0 - b_1x_i|\right) =$$

$$= \min \sum_{i=1}^{n} \Psi(\varepsilon)(y_i - b_0 - b_1x_i)n^{-1}$$

onde

θ é o *j-ésimo* quantil da amostra e $0 < \theta < 1$;

$\Psi(\varepsilon) = \begin{cases} \theta\varepsilon, & \varepsilon \geq 0 \\ (\theta-1)\varepsilon, & \varepsilon < 0 \end{cases}$, ou seja, Ψ multiplica os resíduos por θ se estes forem positivos ou por $\theta - 1$ caso contrário, fazendo com que estes sejam tratados assimetricamente.

No caso em que se tem $\theta = 1/2$ estamos no caso da mediana, ou seja, na regressão dos mínimos desvios absolutos descrita em na secção 2.7.1.
Os estimadores para um número fixo de quantis são combinações lineares de estatísticas de ordem.

Capítulo 3: Aplicações/Resultados

O estudo inicia-se com o cálculo da correlação, isto é, dos valores de $\eta^2_{X|Y}$, $\eta^2_{Y|X}$ e r^2 para o mesmo agrupamento de dados. No entanto, tendo em mente que r^2 é sempre menor do que $\eta^2_{X|Y}$ e $\eta^2_{Y|X}$, enumera-se já alguns resultados evidentes e que irão ser verificados com os exemplos. Assim, temos que:

- $\Rightarrow$ Se $r^2 \approx 1$, está tudo resolvido. Se as variáveis têm distribuição bivariada normal e temos o valor de r^2 também, não vamos precisar procurar os valores de η^2_{**}, pois o coeficiente de correlação de Pearson traduz uma correlação linear perfeita quando estamos perante variáveis binormais ("Toda a correlação é linear").
- $\Rightarrow$ Se $r^2 \ll 1$ calcula-se os η^2_{**}. Uma dúvida que defrontámos é as variáveis não serem gaussianas. Se não forem, os resíduos ainda podem ser gaussianos ou podemos estar perante uma situação em que nada é gaussiano. Os resíduos podem mesmo ser gaussianos, mas não terem a mesma variância (é o caso do conjunto de dados "Virgin", que iremos ver mais à frente, onde os resíduos não apresentam variância constante).
  - ↳ Se os $\eta^2_{**} \approx r^2$, a distribuição (X, Y) pode ser binormal, isto é, tem que ser avaliada a possibilidade das rectas de regressão terem toda a informação.
  - ↳ Se houver um $\eta^2_{**} \gg r^2$ (ou se pelo menos for possível distinguir estatisticamente η^2_{**} de r^2) então a distribuição (X, Y) não é binormal e por isso tem-se que analisar as duas curvas condicionadas, ou seja, avaliar a não linearidade e o que isso significa. Além disso, os resíduos podem ou não ser normais;

Após a análise da correlação, num estudo da regressão vamos querer à custa de uma variável obter informação sobre outra. Neste caso podemos ter quatro situações:

- ✓ Conhecida uma das variáveis, podemos querer estimar uma média (erro da ordem de $\frac{\sigma^2}{n}$) ou estimar um valor individual (erro da ordem de σ^2);
- ✓ Podemos querer um valor no meio da amostra (uma "interpolação") ou um valor para além do limite da amostra (uma "extrapolação");
- ✓ Podemos querer estimar um valor que foi efectivamente observado, ou seja, um valor da amostra ou um valor que não foi observado mas pertence à população;
- ✓ Pode-se fazer estimativas com uma recta ou uma curva de regressão usando o método de mínimos quadrados ou com uma curva de regressão global usando as médias condicionais.

Assim, temos 2^4 situações diferentes.

Nota: No estudo da curva de regressão global, para o agrupamento de classes considera-se o seguinte:

1. Se tiver sentido, deve-se considerar todas as classes com a mesma amplitude ou com o mesmo efectivo.
2. Quando não for possível e se houver uma regra, deve-se tanto quanto possível segui-la. Por exemplo, classes com extremos 10, 100, 1000, é natural que se escolha a classe seguinte com extremo 10000 (significa que temos classes da

mesma amplitude em logaritmo). Para as classes extremas o problema é mais complicado pela possibilidade de criar pontos influentes para uma regressão, possível existência de *outliers* e também a possibilidade de se estar a supor valores aberrantes para uma variável (por exemplo, obter chuva negativa).

3. Se não há regra, tomamos a variável X e isolamos duas classes consecutivas, $[x_a, x_b[$ e $[x_b, x_c[$. Sejam a média e a variância de cada uma destas classes $\bar{x}_A$, $\bar{x}_B$, s_A^2 e s_B^2, respectivamente. Supondo que temos n_A elementos na primeira classe e n_B elementos na segunda classe, uma estimativa da média feita com a primeira classe será de $\bar{x}_A \pm \frac{s_A}{\sqrt{n_A}}$ e na segunda classe será de $\bar{x}_B \pm \frac{s}{\sqrt{n_B}}$. Se juntarmos as duas classes a média será dada por, $\bar{x}_{AB} = \frac{n_A\bar{x}_A + n_B\bar{x}_B}{n_A+n_B}$ (média das médias devidamente ponderadas) e a variância da classe agrupada é dada por, $s_{AB}^2 = \frac{n_A s_A^2 + n_B s_B^2}{n_A+n_B} + \frac{n_A(\bar{x}_A - \bar{x}_{AB})^2 + n_B(\bar{x}_B - \bar{x}_{AB})^2}{n_A+n_B}$ (média das variâncias devidamente ponderadas e a variância das médias), supondo as classes A e B independentes, ou seja, com covariância nula.

   Deve-se agrupar as classes se o segundo valor for mais pequeno que algum dos outros dois primeiros valores. Pode ser útil agrupar para a classe da ponta e não ser para a classe anterior, visto que é nos extremos que a situação é mais sensível (*outliers* ou pontos influentes). Ao fazer este novo agrupamento não nos podemos esquecer qual o efeito da perda de um grau de liberdade ($n - 1$).

3.1. Linearidade e não linearidade em diferentes exemplos

Para se obter uma estimativa da distribuição dos valores do coeficiente de correlação não linear, que é um dos objectivos deste trabalho, é essencial que não haja uma dependência das classes dos valores do *x* e do *y* (tanto no tamanho das classes como na sua posição).

Neste sentido, numa primeira abordagem, através de dois dos exemplos já referenciados, Grais (1974) e Calot (1969), e um novo conjunto de dados individuais referentes à "Conservação do contexto de pares de codões", Moura *et al* (2005), vai-se expor algumas das diferenças que se verificam entre os resultados obtidos usando as classes estipuladas pelos autores e quando existe mudança das classes, para os dois primeiros exemplos e no terceiro analisa-se os resultados com diferentes agrupamentos.

Mas antes de passar a esta análise vai-se descrever mais pormenorizadamente o terceiro exemplo.

O conjunto dos dados individuais caracteriza a relação que existe entre a conservação do contexto de pares de codões, variável y e a conservação do codão, variável x (Figura 7). Este estudo é feito com base em seqüências de codificação de 22 espécies de fungos e é analisado através do software *Anaconda*, de onde são obtidas as variáveis aqui apresentadas.

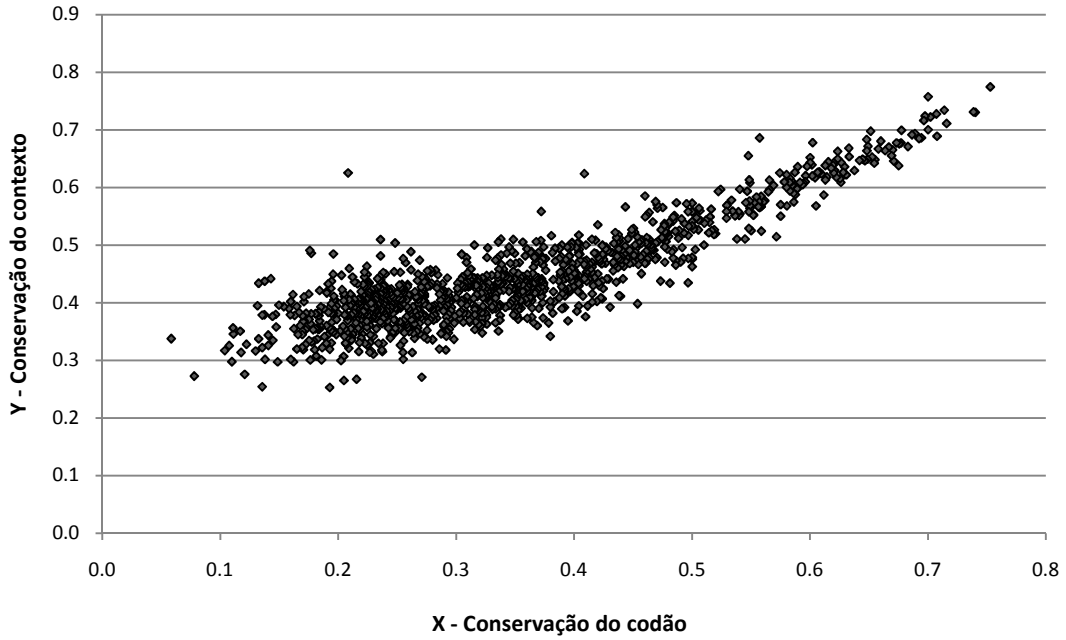


Figura 7 – Distribuição da conservação de pares de codões em 22 espécies de fungos

Para este conjunto de dados e dada a grandeza dos valores das variáveis, começou-se por considerar a divisão de cada variável em 24 classes e que são apresentadas na Tabela 12.

Tabela 12 – Dados agrupados da conservação de pares de codões em 22 espécies de fungos

Conservação do contexto / Conservação do codão		< 0,2640	[0,264;0,286[	[0,286;0,308[	[0,308;0,33[	[0,33;0,352[	[0,352;0,374[	[0,374;0,396[	[0,396;0,418[	[0,418;0,4561[	[0,4561;0,4781[	[0,4781;0,5001[	[0,5001;0,5221[	[0,5221;0,5441[	[0,5441;0,5661[	[0,5661;0,5881[	[0,5881;0,6101[	[0,6101;0,6321[	[0,6321;0,6541[	[0,6541;0,6761[	[0,6761;0,6981[	[0,6981;0,7201[	[0,7201;0,7421[	> 0,7421	Total
		0,2530	0,2750	0,2970	0,3190	0,3410	0,3630	0,3850	0,4070	0,4371	0,4671	0,4891	0,5111	0,5331	0,5551	0,5771	0,5991	0,6211	0,6431	0,6651	0,6871	0,7091	0,7311	0,7531	
< 0,0734	0,0584	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
[0,0734;0,1034[	0,0884	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
[0,1034;0,1334[	0,1184	0	1	1	5	3	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13
[0,1334;0,1634[	0,1484	1	0	4	2	2	2	9	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24
[0,1634;0,1934[	0,1784	1	0	3	8	9	16	13	4	3	0	2	0	0	0	0	0	0	0	0	0	0	0	0	59
[0,1934;0,2234[	0,2084	0	2	2	5	16	15	25	21	13	1	1	0	0	0	0	0	1	0	0	0	0	0	0	102
[0,2234;0,2534[	0,2384	0	0	0	9	8	23	36	36	34	1	1	2	0	0	0	0	0	0	0	0	0	0	0	150
[0,2534;0,2834[	0,2684	0	1	1	2	11	18	26	22	22	6	1	0	0	0	0	0	0	0	0	0	0	0	0	110
[0,2834;0,3134[	0,2984	0	0	0	2	7	11	16	18	30	3	2	0	0	0	0	0	0	0	0	0	0	0	0	89
[0,3134;0,3639[	0,3387	0	0	0	0	2	12	29	29	70	8	12	4	0	0	0	0	0	0	0	0	0	0	0	166
[0,3639;0,3939[	0,3789	0	0	0	0	1	4	8	10	37	14	11	3	0	1	0	0	0	0	0	0	0	0	0	89
[0,3939;0,4239[	0,4089	0	0	0	0	0	1	4	6	34	19	14	7	1	0	0	0	1	0	0	0	0	0	0	87
[0,4239;0,4539[	0,4389	0	0	0	0	0	0	1	3	6	16	21	11	2	1	0	0	0	0	0	0	0	0	0	61
[0,4539;0,4839[	0,4689	0	0	0	0	0	0	0	0	3	8	11	17	9	6	3	0	0	0	0	0	0	0	0	57
[0,4839;0,5139[	0,4989	0	0	0	0	0	0	0	0	1	7	5	3	12	9	3	0	0	0	0	0	0	0	0	40
[0,5139;0,5439[	0,5289	0	0	0	0	0	0	0	0	0	0	0	5	3	10	3	3	0	0	0	0	0	0	0	24
[0,5439;0,5739[	0,5589	0	0	0	0	0	0	0	0	0	0	0	2	3	5	12	5	2	0	1	1	0	0	0	31
[0,5739;0,6039[	0,5889	0	0	0	0	0	0	0	0	0	0	0	0	0	1	4	11	7	5	0	1	0	0	0	29
[0,6039;0,6339[	0,6189	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	12	9	2	0	0	0	0	26
[0,6339;0,6639[	0,6489	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	8	4	3	0	0	0	16
[0,6639;0,6939[	0,6789	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	5	5	1	0	0	13
[0,6939;0,7239[	0,7089	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3	4	1	10
> 0,7239	0,7389	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	3
Total		2	5	11	33	60	103	168	151	256	83	81	54	30	33	27	20	24	24	12	12	4	6	2	1201

Passando agora ao nosso estudo, as mudanças que se efectuaram consistiram em fazer diferentes agrupamentos de classes. No exemplo dos Salários, começou-se por juntar as classes duas a duas, mas como a variável "idade" possui um número impar de classes, considerou-se duas situações. Numa, a primeira classe ficou isolada e na outra foi a última classe a ser considerada sozinha. Numa última abordagem juntou-se as duas primeiras classes, as duas seguintes e as últimas três, na variável "idade", mantendo-se a variável "salário mensal" com os agrupamentos anteriores. No exemplo das Fibras, começou-se por considerar em cada variável a junção de classes duas a duas. De seguida, mantendo-se esta junção na variável "resistência à ruptura", considerou-se na variável "comprimento" a união das duas primeiras classes, das três seguintes, de mais três e por último das duas restantes. Outra situação, foi considerar a junção das três classes das pontas. Por fim, manteve-se esta junção na variável "comprimento" e uniu-se as classes da variável "resistência à ruptura", três a três. No exemplo dos Codões, para cada uma das variáveis, optou-se por juntar as três primeiras classes e depois as restantes duas a duas e em outra situação, quatro a quatro.

Foram então determinadas em cada junção diferente de classes, as medidas já analisadas no capítulo 1 (Tabela 13). Além disso, acrescentou-se o valor da expressão (4) da secção 2.6.. Entre parênteses na primeira linha encontram-se os valores obtidos considerando os dados individuais apresentados na secção 3.2..

Analisando assim a Tabela 13, observa-se que a junção de duas classes afecta quer a correlação linear, quer a correlação global. Esta diminui, traduzindo-se numa menor não linearidade. No entanto, nos exemplos das Fibras e dos Codões, na junção de apenas duas classes, os valores são menos afectados do que no exemplo dos Salários. Esta afectação parece no entanto afectar a não linearidade quando fixamos as classes na variável X, no sentido em que esta parece tornar-se insignificante, valores inferiores a uma centésima nos exemplos dos Salários e das Fibras. Para as restantes medidas, podemos ainda concluir que a junção de classes, para a variável y, diminui o "achatamento", trazendo assim a variável para mais perto da normal. Já no exemplo linear isto não acontece, o " achatamento" aumenta com a junção de classes.

Tabela 13 – Valores dos coeficientes de linearidade e não linearidade consoante as classes escolhidas

	r	$\eta^2_{Y\|X}$	$\eta^2_{X\|Y}$	r^2	$\eta^2_{Y\|X} - r^2$	$\eta^2_{X\|Y} - r^2$	γ_{1x}	γ_{1y}	γ_{2x}	γ_{2y}	r^3	$\left\|\frac{\gamma_{1x}}{\gamma_{1y}}\right\|$	$\left\|\frac{\gamma_{1y}}{\gamma_{1x}}\right\|$
Salários													
Inicial: 7 × 8	0,5012	0,3609	0,2706	0,2512	0,1097	0,0195	0,0504 (0,0069)	1,8535 (1,8494)	-0,5740 (-0,378)	7,1434 (7,049)	0,1259	0,0272 (0,0037)	36,7758 (268,0290)
2 a 2, prim.: 4 × 4	0,4370	0,2175	0,1951	0,1910	0,0265	0,0042	0,0582	1,2463	-0,6524	2,8978	0,0835	0,0467	21,4141
2 a 2, últ.: 4 × 4	0,4548	0,2669	0,2212	0,2069	0,0600	0,0143	0,0582	1,2980	-0,6524	2,3464	0,0941	0,0448	22,3024
2+2+3: 3 × 4	0,4436	0,2660	0,2116	0,1968	0,0692	0,0148	0,0582	1,3096	-0,6524	1,4457	0,0873	0,0444	22,5017
Fibras													
Inicial: 10 × 12	0,6321	0,4533	0,4162	0,4162	0,0537	0,0167	0,2385	0,7077	-0,3312	0,7032	0,2526	0,3370	2,9673
2 a 2: 5 × 6	0,6219	0,4077	0,3950	0,3868	0,0209	0,0082	0,2807	0,6377	-0,2770	0,4446	0,2405	0,4402	2,2718
2+3+3+2: 4 × 6	0,5456	0,3292	0,3218	0,2977	0,0315	0,0241	0,2262	0,6377	-0,6437	0,4446	0,1624	0,3547	2,8192
3+2+2+3: 4 × 6	0,5603	0,3574	0,3443	0,3140	0,0434	0,0303	0,1175	0,6377	-0,5121	0,4446	0,1759	0,1843	5,4272
idem, 3 a 3: 4 × 4	0,5311	0,3226	0,3006	0,2821	0,0405	0,0184	0,1175	0,6334	-0,5121	0,2243	0,1498	0,1855	5,3906
Codões													
Inicial: 23× 23	0,8771	0,8130	0,7799	0,7692	0,0438	0,0107	0,6775	0,9781	-0,1228	0,7272	0,6747	0,6927	1.4436
3+2 a 2: 11 × 11	0,8665	0,8068	0,7637	0,7509	0,0560	0,0128	0,7444	1,0542	0,0497	0,9702	0,6506	0,7061	1,4162
3+4 a 4: 6 × 6	0,8140	0,7239	0,6712	0,6626	0,0613	0,0086	0,7482	1,1256	-0,0580	1,0127	0,5393	0,6647	1,5044

Para analisar a existência de não linearidade na regressão, em vez de recorrer à diferença entre η_{**}^2 e r^2, pode-se usar o seu quociente, ou seja:

- Se $\frac{\eta_{**}^2}{r^2} = 1$, não existe parte não linear;
- Se $\frac{\eta_{**}^2}{r^2} > 1$, existe parte não linear.

Assim, de seguida, fez-se uma comparação dos coeficientes de linearidade dos quatro exemplos já enunciados e juntou-se este quociente de coeficientes de correlações.

Tabela 14 – Medidas de linearidade

	Betão	**Betão (log)**	**Salários**	**Fibras**	**Codões**
r^2	0,6071	0,9697	0,2512	0,3996	0,7692
$\eta_{Y\|X}^2$	0,9691	0,9630	0,3609	0,4533	0,8130
$\eta_{X\|Y}^2$	0,8423	0,9759	0,2706	0,4162	0,7799
$\eta_{Y\|X}^2 - r^2$	0,3620	0,0133	0.1097	0.0537	0,0438
$\eta_{X\|Y}^2 - r^2$	0,2352	0,0062	0,0194	0.0166	0,0107
$\frac{\eta_{Y\|X}^2}{r^2}$	1,5963	1,0064	1,4367	1,1345	1,0569
$\frac{\eta_{X\|Y}^2}{r^2}$	1,3874	1,0137	1,0772	1,0417	1,0139
r^3	0,4730	-0,9548	0,1259	0,2526	0,6747
$\left\|\frac{\gamma_{1x}}{\gamma_{1y}}\right\|$	2,4827	1,0631	0,0272	0,3370	0,6927
$\left\|\frac{\gamma_{1y}}{\gamma_{1x}}\right\|$	0,4028	0,9406	36,7758	2,9673	1.4436

No exemplo do Betão, a parte da correlação que não é linear baixou de 36%, para 1% . Esta conclusão tem, no entanto, que ser lida com um certo cuidado visto estar a trabalhar com "logaritmos". Apesar de neste exemplo a transformação ter resultado muito bem, no sentido em que linearizou os dados, tal como já foi referido em Bessa (2007) a utilização de transformações pode melhorar a performance do método estatístico, como também pode não se adequar à realidade, em particular, disfarçando situações importantes, como por exemplo misturas. Nos restantes exemplos, a linearidade tem um decréscimo menos acentuado.

Analisando o quociente entre os coeficientes de correlação, conclui-se que o exemplo do Betão é o que possui maior parte não linear, como era de se esperar. Tal como era de se esperar, quando consideramos o Betão (log) o quociente é aproximadamente 1, ou seja, não existe parte não linear. Analisando os coeficientes de Dodge, para estes

conjuntos de dados, verifica-se que estão de acordo com o valor de $\frac{\eta_{**}^2}{r^2}$, no caso linear e quando há indicação da existência de não linearidade. Este coeficiente pode ainda ser um indicador de qual das variáveis deve ser a variável dependente, o que não tem nada a ver com causadidade. Será portanto aquela em que o coeficiente se aproxima de 1 sem nunca o ultrapassar. Observando os valores, verifica-se que apenas no exemplo do Betão (log) a curva de regressão deve de ser feita fixando as classes de y. A par destes resultados temos ainda a medida $\eta_{X|Y}^2 - r^2$, que chega a ser inferior a $0{,}01$ indicando que não existe não linearidade.

Apesar de todas estas novas medidas serem uma mais-valia, no sentido que possuímos assim alguns coeficientes que nos podem traduzir a existência ou não da linearidade num conjunto de dados, temos de ter muito cuidado com a sua análise. Por exemplo, no conjunto de dados Salários, o valor de $\eta_{**}^2 - r^2$ é muito pequeno, que sugere não haver não linearidade, mas o valor de $\left|\frac{\gamma_{1x}}{\gamma_{1y}}\right|$ sugere que também não temos linearidade, ou seja, não se pode concluir que haja uma linearidade expressiva. Apesar de r^3 se aproximar de $\left|\frac{\gamma_{1x}}{\gamma_{1y}}\right|$, este valor não é 1. Estas conclusões parecem contraditórias, no entanto, o que se está a concluir é que a correlação linear é praticamente a mesma que a correlação global e que este exemplo não é tipicamente linear, no sentido que uma recta de regressão linear pode não ser a melhor opção para se aproximar os dados.

3.2. Impacto do Agrupamento em Classes

Para se obter uma representação mais realista dos valores e para um melhor tratamento dos dados, começou-se por fazer uma transformação dos valores apresentados na Tabela 6.

Essa transformação consiste em primeiro considerar os pontos médios das classes, segundo o que indica Grais (1974) no tratamento que dá a estes dados. De seguida, construiu-se dois vectores, um para a idade e outro para o salário mensal de acordo com o número de indivíduos apresentados na Tabela 6. Para obter uma melhor distribuição dos valores e de acordo com as amplitudes das classes, construiu-se dois novos vectores que consistiam em somar a cada valor médio da classe, um número aleatório de forma a preencher a classe com uma lei uniforme (não sabemos que distribuição cada classe possui, aqui optou-se pela uniforme), isto é, entre menos metade e mais metade da amplitude da classe. Um exemplo de uma representação gráfica obtida foi a apresentada na Figura 8.

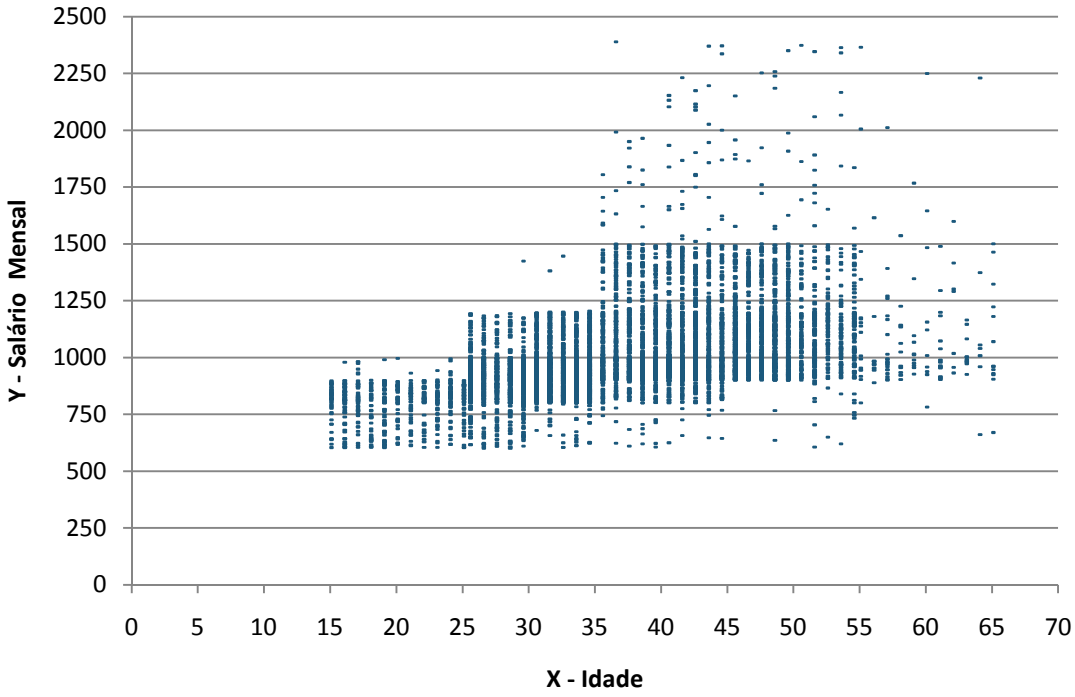


Figura 8 – Distribuição do salário mensal de trabalhadores segundo a idade

Ora, se se estudar o resultado das diferentes variáveis obtidas com diferentes simulações ficamos com estimativas do erro de agrupamento das diferentes medidas e podemos desta forma avalia-lo. Assim, escolheu-se efectuar 100 simulações e em cada uma delas obter os valores de r, $\eta^2_{X|Y}$, $\eta^2_{Y|X}$, b_0 e b_1 para a recta de regressão da variável y em função de x e para a recta de regressão da variável x em função de y. Os valores obtidos através destas simulações e em comparação com os valores sem a transformação, encontram-se resumidos na seguinte tabela.

Tabela 15 – Valores descritivos das simulações feitas do conjunto de dados dos Salários

	Média	Desvio-Padrão	Variância	"Achatamento"	Assimetria	Valores iniciais
$\boldsymbol{r}$ com agrupamento	0,5008	0,0017	2,7810E-06	0,0557	-0,1173	0,5012
$\boldsymbol{\eta^2_{Y\|X}}$	0,2706	0,0019	3,4915E-06	0,1960	-0,0044	0.2706
$\boldsymbol{\eta^2_{X\|Y}}$	0,3610	0,0010	1,0033E-06	-0,1350	-0,0465	0.3609
$\boldsymbol{r}$ sem agrupamento	0,4740	0,0026	6,8621E-06	-0,5395	-0,0521	0.4982
$\boldsymbol{b_0}$ de X|Y	15,0270	0,2034	0,04138818	-0,1496	0,0597	13,462
$\boldsymbol{b_1}$ de X|Y	0,0223	0,0002	4,1698E-08	-0,1538	-0,0308	0,0238
$\boldsymbol{b_0}$ de Y|X	630,3676	2,6131	6,8281	-0,1784	-0,0555	618,15
$\boldsymbol{b_1}$ de Y|X	10,08580	0,0726	0,0052	-0,2774	0,0414	10,411

Analisando os valores obtidos, para os dados agrupados, conclui-se que o erro é inferior a uma milésima para o $\eta^2_{Y|X}$ e inferior a duas centésimas para o coeficiente de correlação

linear e para $\eta^2_{X|Y}$. Para os valores considerados individualmente existe uma menor normalidade do coeficiente de correlação como nos indica o respectivo valor de "achatamento". Os valores das constantes de regressão linear também apresentam alguma variabilidade levando a concluir que se a distribuição não se afastar muito da uniforme, dentro das classes, a substituição dos valores individuais por agrupamentos não parece modificar muito os resultados, exceptuando o valor de $\boldsymbol{b_0}$ na regressão X|Y, com erro superior a 10%.

As correcções de agrupamento para momentos ímpares são nulas. No caso gaussiano (contacto de alta ordem nas caudas) são $\frac{h^2}{12}$ para o desvio padrão, Kendall e Stuart (1969) se as classes tiverem todas a amplitude h.

As correcções tipo Sheppard, não são directamente aplicáveis porque, apesar da distribuição no agrupamento ser considerada uniforme, as amplitudes das classes não são todas iguais e a variável "salário mensal" não tem distribuição simétrica, duas condições necessárias para se aplicar as correcções tipo Sheppard.

3.3. Estimativas de valores duma variável, conhecida outra

Feita esta análise, e utilizando os vectores transformados ilustrados na Figura 8, propõe-se de seguida responder a algumas questões, para este exemplo.

1. Qual a estimativa do rendimento de um funcionário entre 30 e 35 anos, usando a recta de regressão e usando a curva de regressão? Qual a melhor estimativa?
2. Qual a estimativa do rendimento médio dos funcionários entre 30 e 35 anos, usando a recta de regressão e usando a curva de regressão? Qual a melhor estimativa?
3. Qual a estimativa do rendimento de um funcionário com 65 anos, usando a recta de regressão e usando a curva de regressão? Qual a melhor estimativa?

Para responder a cada uma destas questões precisamos de analisar o desvio padrão quando estamos a falar de estimação usando a recta de regressão ou da estimação usando a curva de regressão. Além disso, não nos podemos esquecer que o erro associado à estimativa de um valor individual é diferente do associado a um valor médio.

De seguida, apresentam-se os gráficos dos erros padrão obtidos para estimações com uma recta de regressão (Figuras 9 e 10). Com estes gráficos podemos verificar desde já, que os erros são maiores nas classes extremas superiores, como era de se esperar. Além disso, a curva é mais suave quando se obtém uma estimativa para valores médios.

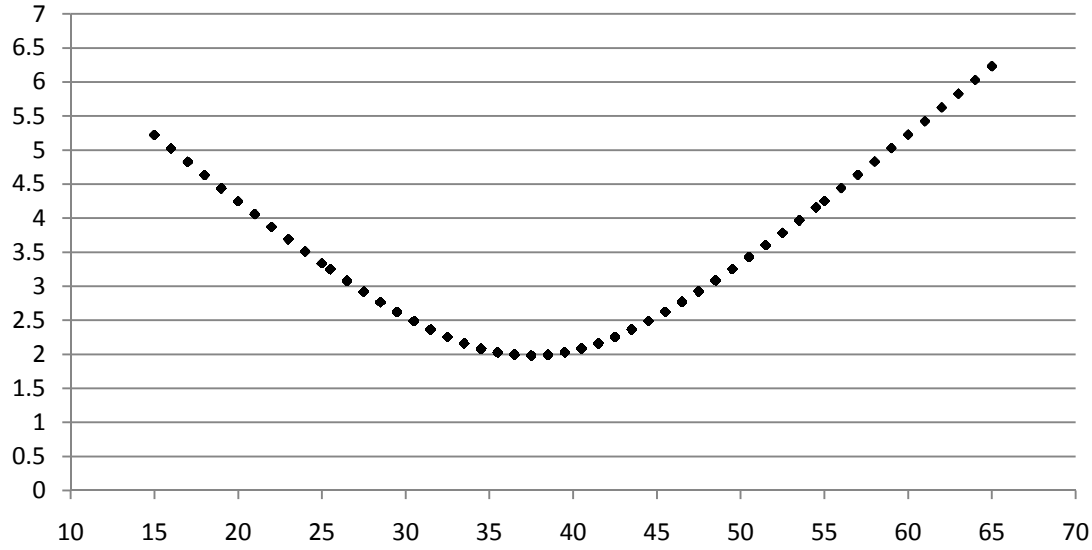


Figura 9 – Distribuição dos erros padrão segundo a idade dos trabalhadores

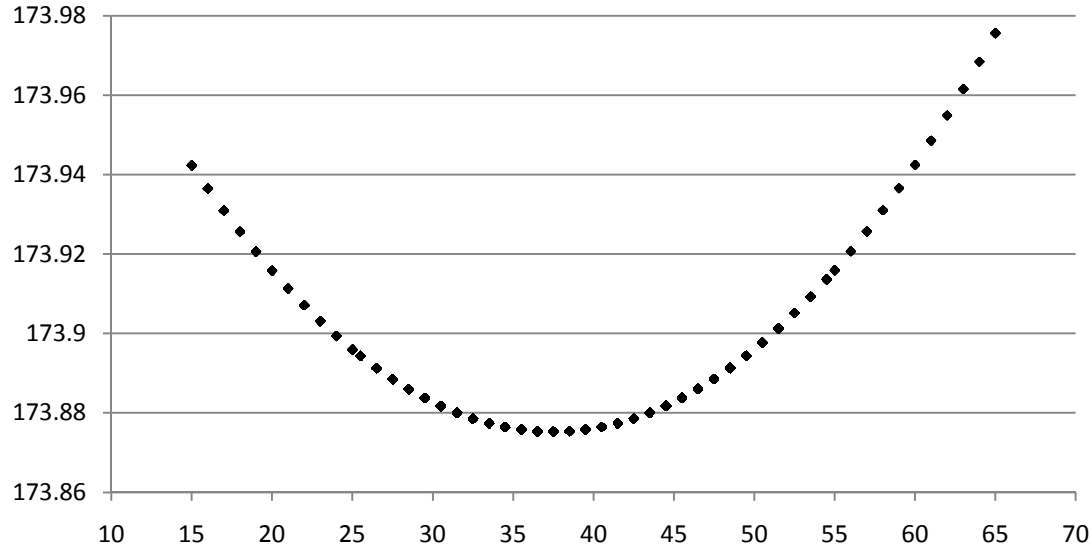


Figura 10 – Distribuição dos erros padrão segundo a idade média dos trabalhadores

Os erros associados aos rendimentos estimados com a curva de regressão, $(s_y)|_{x\in[x_0,x_{0+h}[}$, encontram-se descritos na Tabela 9 para valores individuais. Para valores médios basta retirar nessa tabela o erro associado ao valor individual e dividi-lo pela raiz do número de observações na classe, ou seja, neste caso o erro será $\left(\frac{s_y}{\sqrt{n}}\right)|_{x\in[x_0,x_{0+h}[}$.

Para a questão 1, temos que a estimativa para o rendimento de um funcionário entre os 30 e os 35 anos, usando a curva de regressão será $944{,}5 \pm 99{,}4$ e usando a recta de regressão será: $938{,}3 \pm 173{,}9$. Claramente a melhor estimativa será a primeira, pois é a que tem um menor desvio padrão.

Para a questão 2, temos que uma estimativa para o rendimento médio de um funcionário entre os 30 e os 35 anos, usando a curva de regressão será $944{,}5 \pm 2{,}50$ e usando a

recta de regressão será: $938{,}3 \pm 2{,}25$. Neste caso os desvios padrões são muito idênticos, ou seja, não se perde muito em ter o conjunto de dados agrupados em classes, sendo viável este estudo.
Na questão 3, a estimativa do rendimento de um trabalhador com 65 anos usando a curva de regressão será de $1119{,}5 \pm 293{,}4$ enquanto que com a recta de regressão será de $1081{,}5 \pm 174{,}0$, ou seja, em casos de extrapolação a segunda estimativa parece ser a melhor.

3.4. Métodos alternativos – MDA

Como já foi referenciado na secção 2.7., existem métodos alternativos ao MMQ. Um dos métodos aí referenciados é o MDA e nesse sentido será feita aqui uma pequena análise gráfica, do comportamento das diferentes curvas de regressão, que estabelecem a "ligação" entre as variáveis do conjunto de dados representado na Figura 8. Assim, na seguinte figura encontram-se estas curvas de regressão.

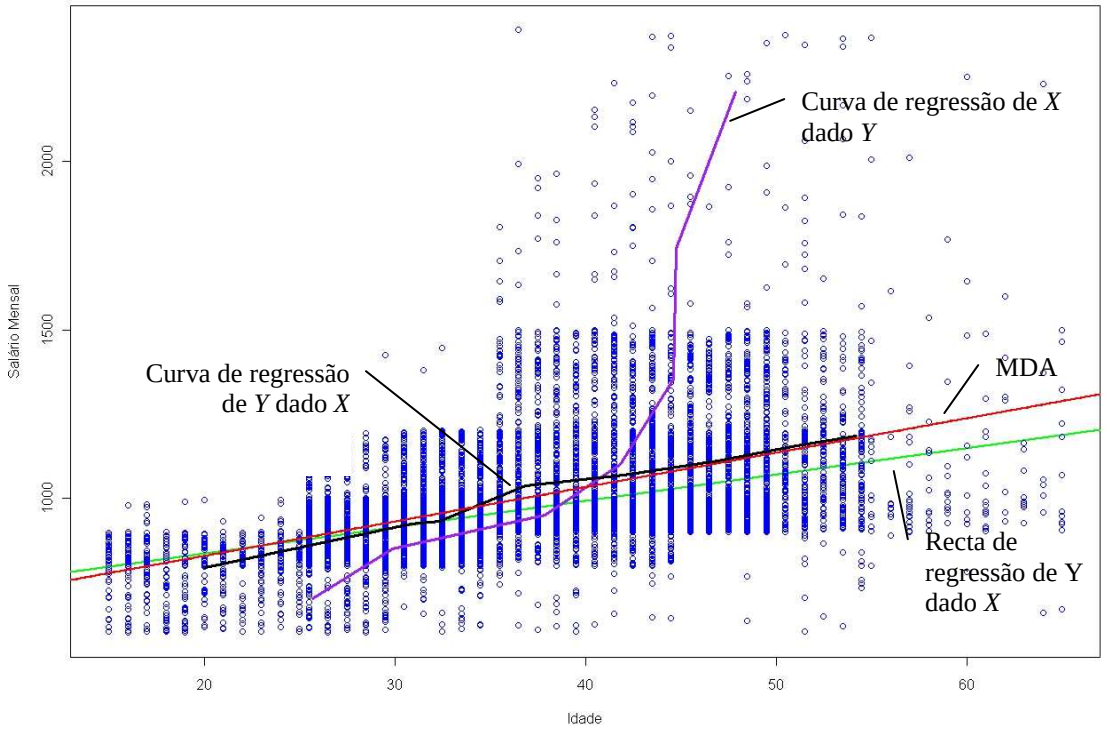


Figura 11 – Distribuição do salário mensal de trabalhadores segundo a idade, Curvas de regressão condicionadas, recta de regressão dos mínimos quadrados e dos desvios absolutos.

Analisando a Figura 11, pode-se concluir que a regressão MDA aproxima-se da curva de regressão da variável *Y* condicionada á variável *X*. Tendo em conta esta observação e os resultados obtidos na secção 3.3., em que a curva de regressão só não obtinha bons resultados para extrapolações, parece-nos que a regressão MDA será a curva que melhor representa este conjunto de dados. Para extrapolações ela tem o mesmo comportamento que a recta de regressão.

3.5. Viadutos para comboios de alta velocidade

Consideremos agora um novo exemplo intitulado de "Virgin"[2]. Este conjunto de dados é referente a um estudo feito num viaduto onde se mediu a aceleração vertical (m/s^2) e deslocamento vertical (m), aquando da passagem de um comboio, em diferentes velocidades (km/h). É de salientar que a variável "velocidade" não é observada, é fixada à priori e portanto não será de interesse estudá-la em todos os sentidos, mas sim ver o que acontece com as outras duas variáveis dada esta.
Nas Tabelas 16 e 17 encontra-se a descrição destes dados, agrupados em classes.

Tabela 16 – Deslocamento vertical (m) em viadutos para comboios de alta velocidade e a sua velocidade

Desloc. / Velocid.	< 0,00208	[0,00208;0,00256[	[0,00256;0,00305[	[0,00305;0,00354[	[0,00354;0,00402[	[0,00402;0,00451[	[0,00451;0,00500[	[0,00500;0,00549[	[0,00549;0,00597[	[0,00597;0,00646[	> 0,00646	Total
	0,00159	0,00232	0,00281	0,00329	0,00378	0,00427	0,00475	0,00524	0,00573	0,00622	0,00670	
140	3	18	10	0	0	0	0	0	0	0	0	31
150	2	23	6	0	0	0	0	0	0	0	0	31
160	5	21	5	0	0	0	0	0	0	0	0	31
170	5	19	7	0	0	0	0	0	0	0	0	31
180	3	19	9	0	0	0	0	0	0	0	0	31
190	2	19	10	0	0	0	0	0	0	0	0	31
200	2	18	11	0	0	0	0	0	0	0	0	31
210	1	20	10	0	0	0	0	0	0	0	0	31
220	1	19	8	3	0	0	0	0	0	0	0	31
230	1	15	10	4	1	0	0	0	0	0	0	31
240	1	11	10	5	3	1	0	0	0	0	0	31
250	0	8	9	6	4	1	2	1	0	0	0	31
260	1	4	7	6	4	3	1	1	2	1	1	31
270	0	3	2	7	4	3	2	2	4	2	2	31
280	0	1	4	0	7	2	4	3	8	2	0	31
290	0	1	0	4	2	5	7	8	3	1	0	31
300	0	0	1	6	7	3	5	6	3	0	0	31
310	0	0	6	8	4	3	7	2	1	0	0	31
320	0	0	12	8	6	1	2	2	0	0	0	31
330	0	2	21	3	2	1	2	0	0	0	0	31
340	0	10	16	4	0	0	1	0	0	0	0	31
350	0	18	12	0	0	1	0	0	0	0	0	31
360	0	25	5	1	0	0	0	0	0	0	0	31
370	1	27	3	0	0	0	0	0	0	0	0	31
380	2	27	2	0	0	0	0	0	0	0	0	31
390	4	26	1	0	0	0	0	0	0	0	0	31
400	5	25	1	0	0	0	0	0	0	0	0	31
410	6	24	1	0	0	0	0	0	0	0	0	31
420	8	21	2	0	0	0	0	0	0	0	0	31
Total	53	424	201	65	44	24	33	25	21	6	3	899

[2] Dados fornecidos pelo o professor Rui Calçada do DEC-FEUP.

Tabela 17 – Aceleração vertical (m/s^2) em viadutos para comboios de alta velocidade e a sua velocidade

Aceler. / Velocid.	< 0,196	[0,196;0,392[	[0,392;0,588 [	[0,588;0,784[	[0,784;0,981[	[0,981;1,177[	[1,177;1,373[	[1,373;1,569[	[1,569;1,765[	[1,765;1,961[	> 1,961	Total
	0,098	0,294	0,490	0,686	0,882	1,079	1,275	1,471	1,667	1,863	2,059	
140	30	1	0	0	0	0	0	0	0	0	0	31
150	27	4	0	0	0	0	0	0	0	0	0	31
160	28	3	0	0	0	0	0	0	0	0	0	31
170	30	1	0	0	0	0	0	0	0	0	0	31
180	31	0	0	0	0	0	0	0	0	0	0	31
190	31	0	0	0	0	0	0	0	0	0	0	31
200	31	0	0	0	0	0	0	0	0	0	0	31
210	30	1	0	0	0	0	0	0	0	0	0	31
220	24	7	0	0	0	0	0	0	0	0	0	31
230	16	14	1	0	0	0	0	0	0	0	0	31
240	5	18	6	2	0	0	0	0	0	0	0	31
250	2	14	6	6	2	1	0	0	0	0	0	31
260	1	4	11	6	4	1	2	1	1	0	0	31
270	0	3	2	11	3	1	2	6	1	2	0	31
280	0	1	2	2	10	0	5	6	4	1	0	31
290	0	0	2	3	6	3	3	8	5	1	0	31
300	0	0	6	4	3	4	2	3	3	6	0	31
310	0	2	9	4	3	1	3	4	3	2	0	31
320	0	8	7	5	5	1	2	0	1	2	0	31
330	0	12	8	5	2	0	1	2	0	1	0	31
340	0	17	9	2	2	0	0	0	0	0	1	31
350	1	24	3	2	0	0	0	0	1	0	0	31
360	4	23	3	0	0	1	0	0	0	0	0	31
370	9	20	1	1	0	0	0	0	0	0	0	31
380	12	18	0	1	0	0	0	0	0	0	0	31
390	17	13	1	0	0	0	0	0	0	0	0	31
400	21	10	0	0	0	0	0	0	0	0	0	31
410	22	9	0	0	0	0	0	0	0	0	0	31
420	23	8	0	0	0	0	0	0	0	0	0	31
Total	395	235	77	54	40	13	20	30	19	15	1	899

As Figuras 12 e 13 descrevem as distribuições dos valores individuais, primeiramente através de uma nuvem de pontos e de seguida através de um gráfico de linhas. Cada uma das curvas representa um comboio e a curva a vermelho representa a curva dos valores médios, ou seja, a curva de regressão fixada a velocidade a que percorre cada comboio.

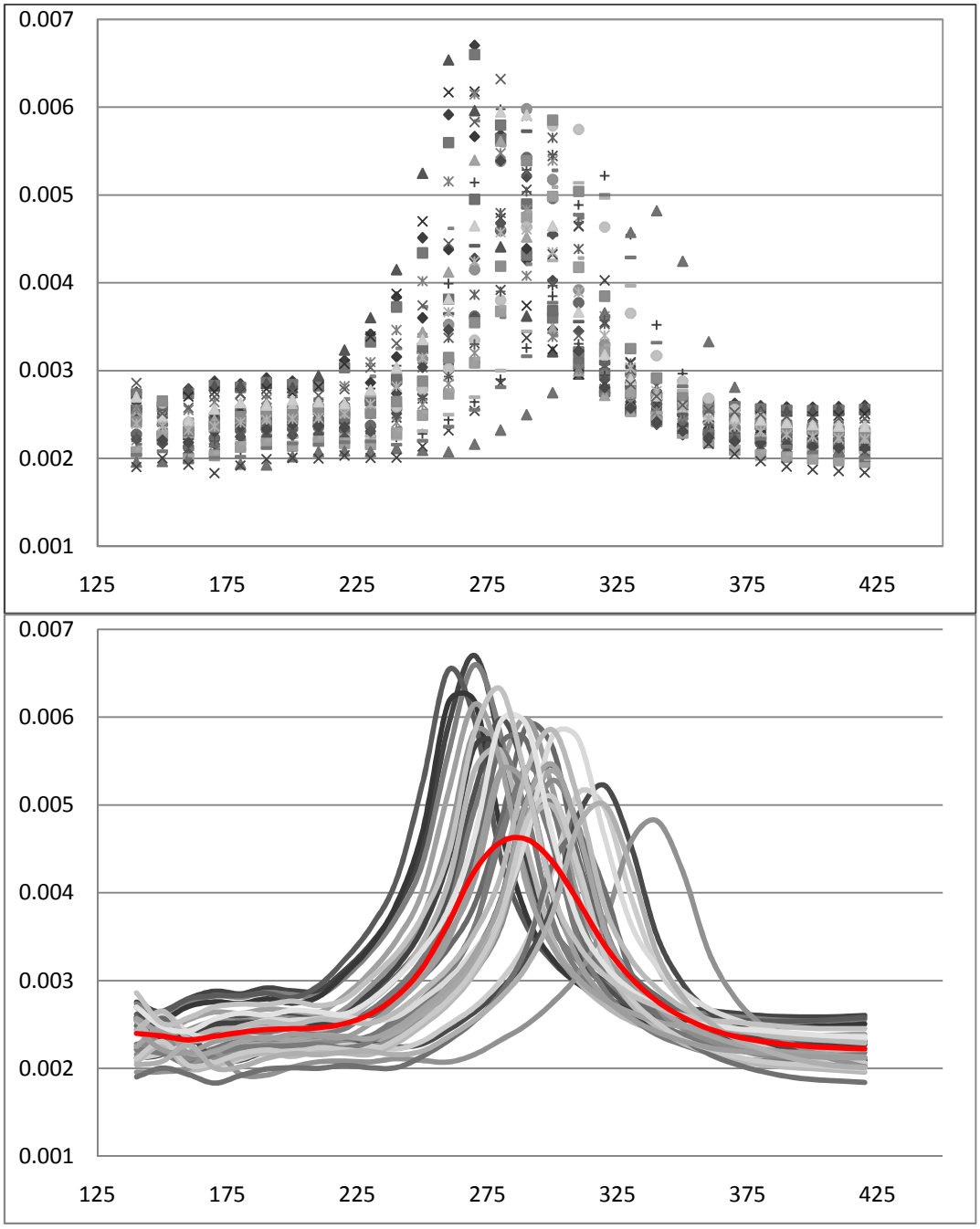


Figura 12 – Distribuição do deslocamento vertical em viadutos para comboios de alta velocidade segundo a sua velocidade

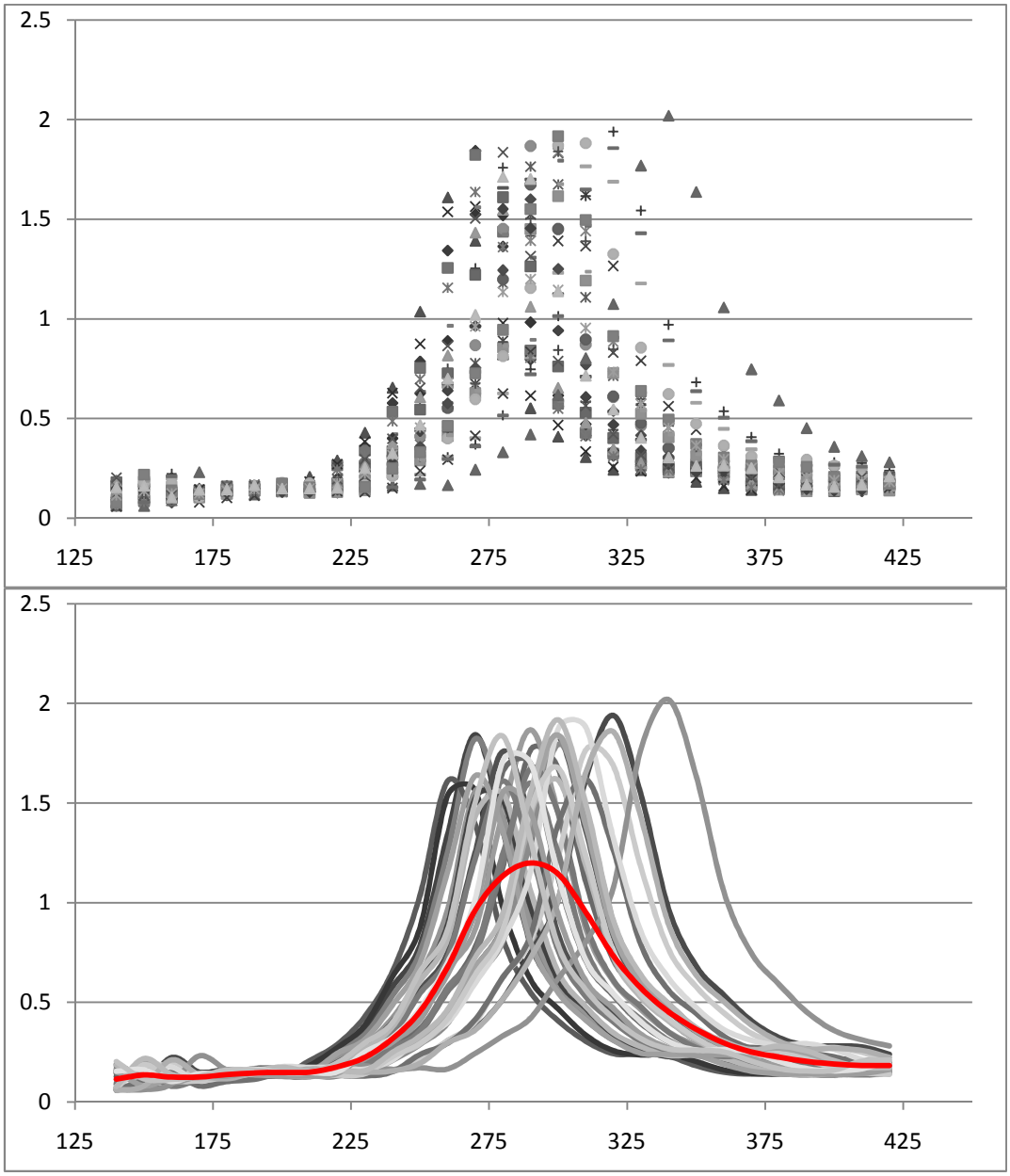


Figura 13 – Distribuição da aceleração vertical em viadutos para comboios de alta velocidade segundo a sua velocidade

Na Tabela 18, apresentam-se os valores das estatísticas descritivas deste conjunto de dados. Nela, podemos observar que o $\eta^2_{X|Y}$ vai dar aproximadamente zero, o que significa que analisar neste sentido é inútil. A correlação é, por isso, aproximadamente nula. Como $\eta^2_{Y|X} \cong 0{,}61$, num caso e $\eta^2_{Y|X} \cong 0{,}66$ no outro há correlação e praticamente toda não é linear.

Tabela 18 – Valores descritivos do deslocamento/aceleração vertical

	Deslocamento vertical	Aceleração vertical
$\bar{x}$	280	280
$\bar{y}$	0,0029	0,4040
s_x	83,666	83,666
s_y	9.7364E-004	0,4400
$\eta^2_{Y\|X}$	0,6128	0,6622
$\eta^2_{X\|Y}$	0,0153	0,1692
r	0,0088	0,1614
r^2	7.8257E-005	0,0261
γ_{1y}	1,6393	1,7895
γ_{2y}	-1,2029	2,4256

Nas Tabelas 19 e 20 faz-se um estudo dos valores condicionais. Nestas tabelas verifica-se que nos valores extremos a média condicionada é praticamente constante e baixa e que nas classes centrais há um razoável aumento. Nas classes extremas a assimetria e o achatamento são pequenos, sugerindo normalidade. Ao contrário, nas classes intermédias os valores são grandes não havendo, por isso, normalidade.

Tabela 19 – Valores descritivos condicionais do deslocamento vertical (m) dada a velocidade

Condicionais – $Y\mid_{X\in]X_0,X_0+h[}$		**Valores das medidas**	**Condicionais –** $Y\mid_{X\in]X_0,X_0+h[}$		**Valores das medidas**
$\boldsymbol{y\mid_{x=140}}$	$\bar{y}$	0,00241	$\boldsymbol{y\mid_{x=290}}$	$\bar{y}$	0,00461
	s_y	0,000347		s_y	0,000877
	γ_1	-0,773		γ_1	-0,591
	γ_2	0,532		γ_2	-0,084
$\boldsymbol{y\mid_{x=150}}$	$\bar{y}$	0,00237	$\boldsymbol{y\mid_{x=300}}$	$\bar{y}$	0,00433
	s_y	0,000280		s_y	0,000855
	γ_1	-0,639		γ_1	0,089
	γ_2	2,055		γ_2	-1,263
$y\mid_{x=160[}$	$\bar{y}$	0,00228	$\boldsymbol{y\mid_{x=310}}$	$\bar{y}$	0,00389
	s_y	0,000350		s_y	0,000855
	γ_1	-0,691		γ_1	0,335
	γ_2	0,263		γ_2	-1,103
$y\mid_{x=170}$	$\bar{y}$	0,00231	$\boldsymbol{y\mid_{x=320}}$	$\bar{y}$	0,00345
	s_y	0,000374		s_y	0,000726
	γ_1	-0,643		γ_1	1,136
	γ_2	-0,047		γ_2	0,372
$\boldsymbol{y\mid_{x=180}}$	$\bar{y}$	0,00239	$\boldsymbol{y\mid_{x=330}}$	$\bar{y}$	0,00306
	s_y	0,000340		s_y	0,000592
	γ_1	-0,740		γ_1	1,744
	γ_2	0,644		γ_2	2,254
$\boldsymbol{y\mid_{x=190}}$	$\bar{y}$	0,00243	$\boldsymbol{y\mid_{x=340}}$	$\bar{y}$	0,00277
	s_y	0,000314		s_y	0,000478
	γ_1	-0,702		γ_1	2,175
	γ_2	0,972		γ_2	6,927
$\boldsymbol{y\mid_{x=200}}$	$\bar{y}$	0,00244	$\boldsymbol{y\mid_{x=350}}$	$\bar{y}$	0,00257
	s_y	0,000321		s_y	0,000389
	γ_1	-0,756		γ_1	2,613
	γ_2	0,867		γ_2	8,871
$\boldsymbol{y\mid_{x=210}}$	$\bar{y}$	0,00245	$\boldsymbol{y\mid_{x=360}}$	$\bar{y}$	0,00243
	s_y	0,000276		s_y	0,000238
	γ_1	-0,387		γ_1	2,099
	γ_2	1,032		γ_2	3,628
$\boldsymbol{y\mid_{x=220}}$	$\bar{y}$	0,00252	$\boldsymbol{y\mid_{x=370}}$	$\bar{y}$	0,00234
	s_y	0,000363		s_y	0,000199
	γ_1	0,450		γ_1	-0,534
	γ_2	0,573		γ_2	6,493
$\boldsymbol{y\mid_{x=230}}$	$\bar{y}$	0,00263	$\boldsymbol{y\mid_{x=380}}$	$\bar{y}$	0,00230
	s_y	0,000437		s_y	0,000222
	γ_1	0,480		γ_1	-1,396
	γ_2	0,418		γ_2	5,561
$\boldsymbol{y\mid_{x=240}}$	$\bar{y}$	0,00281	$\boldsymbol{y\mid_{x=390}}$	$\bar{y}$	0,00224
	s_y	0,000575		s_y	0,000265
	γ_1	0,534		γ_1	-1,582
	γ_2	0,021		γ_2	2,385
$\boldsymbol{y\mid_{x=250}}$	$\bar{y}$	0,00315	$\boldsymbol{y\mid_{x=400}}$	$\bar{y}$	0,00222
	s_y	0,000784		s_y	0,000289
	γ_1	0,987		γ_1	-1,355
	γ_2	0,274		γ_2	1,198
$\boldsymbol{y\mid_{x=260}}$	$\bar{y}$	0,00363	$\boldsymbol{y\mid_{x=410}}$	$\bar{y}$	0,00219
	s_y	0,001226		s_y	0,000309
	γ_1	0,865		γ_1	-1,154
	γ_2	0,054		γ_2	0,385
$\boldsymbol{y\mid_{x=270}}$	$\bar{y}$	0,00427	$\boldsymbol{y\mid_{x=420}}$	$\bar{y}$	0,00216
	s_y	0,001315		s_y	0,000358
	γ_1	0,324		γ_1	-0,626
	γ_2	-1,078		γ_2	-0,607
$\boldsymbol{y\mid_{x=280}}$	$\bar{y}$	0,00457			
	s_y	0,001132			
	γ_1	-0,313			
	γ_2	-1,100			

Tabela 20 – Valores descritivos condicionais da aceleração vertical (m/s^2) dada a velocidade

Condicionais – $Y\mid_{X\in]X_0,X_0+h[}$		**Valores das medidas**	**Condicionais –** $Y\mid_{X\in]X_0,X_0+h[}$		**Valores das medidas**
$\boldsymbol{y\mid_{x=140}}$	$\bar{y}$	0,104	$\boldsymbol{y\mid_{x=290}}$	$\bar{y}$	1,205
	s_y	0,035		s_y	0,384
	γ_1	5,295		γ_1	-0,243
	γ_2	26,033		γ_2	-1,138
$\boldsymbol{y\mid_{x=150}}$	$\bar{y}$	0,123	$\boldsymbol{y\mid_{x=300}}$	$\bar{y}$	1,154
	s_y	0,066		s_y	0,505
	γ_1	2,213		γ_1	0,102
	γ_2	2,898		γ_2	-1,446
$y\mid_{x=160[}$	$\bar{y}$	0,117	$\boldsymbol{y\mid_{x=310}}$	$\bar{y}$	0,965
	s_y	0,058		s_y	0,498
	γ_1	2,728		γ_1	0,379
	γ_2	5,440		γ_2	-1,312
$y\mid_{x=170}$	$\bar{y}$	0,104	$\boldsymbol{y\mid_{x=320}}$	$\bar{y}$	0,731
	s_y	0,035		s_y	0,447
	γ_1	5,295		γ_1	1,203
	γ_2	26,033		γ_2	0,671
$\boldsymbol{y\mid_{x=180}}$	$\bar{y}$	0,098	$\boldsymbol{y\mid_{x=330}}$	$\bar{y}$	0,604
	s_y	0,000		s_y	0,401
	γ_1			γ_1	1,640
	γ_2			γ_2	1,954
$\boldsymbol{y\mid_{x=190}}$	$\bar{y}$	0,098	$\boldsymbol{y\mid_{x=340}}$	$\bar{y}$	0,471
	s_y	0,000		s_y	0,335
	γ_1			γ_1	3,476
	γ_2			γ_2	13,391
$\boldsymbol{y\mid_{x=200}}$	$\bar{y}$	0,098	$\boldsymbol{y\mid_{x=350}}$	$\bar{y}$	0,376
	s_y	0,000		s_y	0,263
	γ_1			γ_1	3,884
	γ_2			γ_2	16,018
$\boldsymbol{y\mid_{x=210}}$	$\bar{y}$	0,104	$\boldsymbol{y\mid_{x=360}}$	$\bar{y}$	0,313
	s_y	0,035		s_y	0,168
	γ_1	5,295		γ_1	2,901
	γ_2	26,033		γ_2	11,420
$\boldsymbol{y\mid_{x=220}}$	$\bar{y}$	0,142	$\boldsymbol{y\mid_{x=370}}$	$\bar{y}$	0,256
	s_y	0,082		s_y	0,126
	γ_1	1,312		γ_1	0,929
	γ_2	-0,280		γ_2	2,457
$\boldsymbol{y\mid_{x=230}}$	$\bar{y}$	0,199	$\boldsymbol{y\mid_{x=380}}$	$\bar{y}$	0,231
	s_y	0,110		s_y	0,126
	γ_1	0,486		γ_1	1,147
	γ_2	-0,795		γ_2	3,045
$\boldsymbol{y\mid_{x=240}}$	$\bar{y}$	0,326	$\boldsymbol{y\mid_{x=390}}$	$\bar{y}$	0,193
	s_y	0,150		s_y	0,110
	γ_1	0,579		γ_1	0,613
	γ_2	0,272		γ_2	-0,668
$\boldsymbol{y\mid_{x=250}}$	$\bar{y}$	0,459	$\boldsymbol{y\mid_{x=400}}$	$\bar{y}$	0,161
	s_y	0,234		s_y	0,092
	γ_1	0,768		γ_1	0,759
	γ_2	-0,076		γ_2	-1,424
$\boldsymbol{y\mid_{x=260}}$	$\bar{y}$	0,680	$\boldsymbol{y\mid_{x=410}}$	$\bar{y}$	0,155
	s_y	0,357		s_y	0,089
	γ_1	1,039		γ_1	0,924
	γ_2	0,636		γ_2	-1,146
$\boldsymbol{y\mid_{x=270}}$	$\bar{y}$	0,965	$\boldsymbol{y\mid_{x=420}}$	$\bar{y}$	0,149
	s_y	0,459		s_y	0,086
	γ_1	0,427		γ_1	1,106
	γ_2	-1,008		γ_2	-0,777
$\boldsymbol{y\mid_{x=280}}$	$\bar{y}$	1,136			
	s_y	0,402			
	γ_1	-0,125			
	γ_2	-0,970			

Com base nestas tabelas e sabendo que neste exemplo, a variação do desvio padrão é importante para o projecto dos viadutos, sendo assim, analisou-se como estes variam fixada a velocidade.

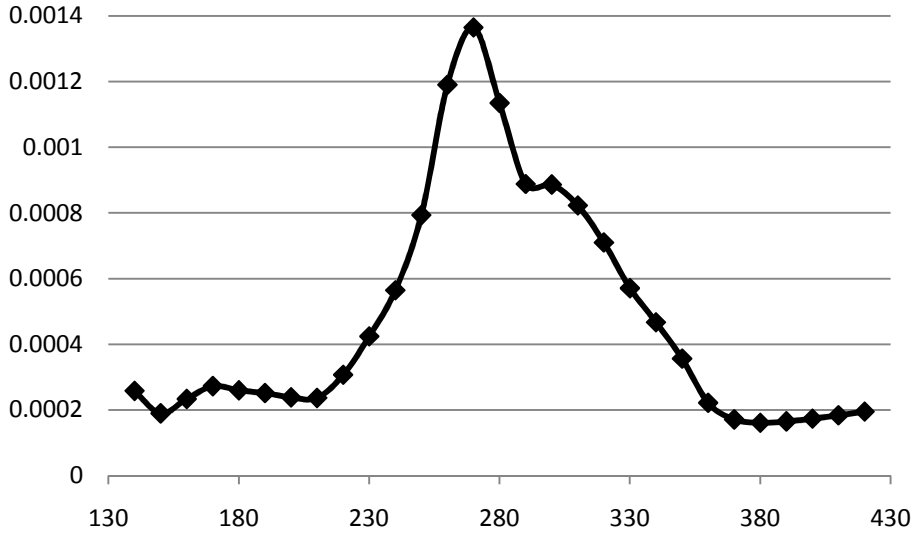


Figura 14 – Distribuição do desvio padrão segundo a velocidade para os deslocamentos verticais

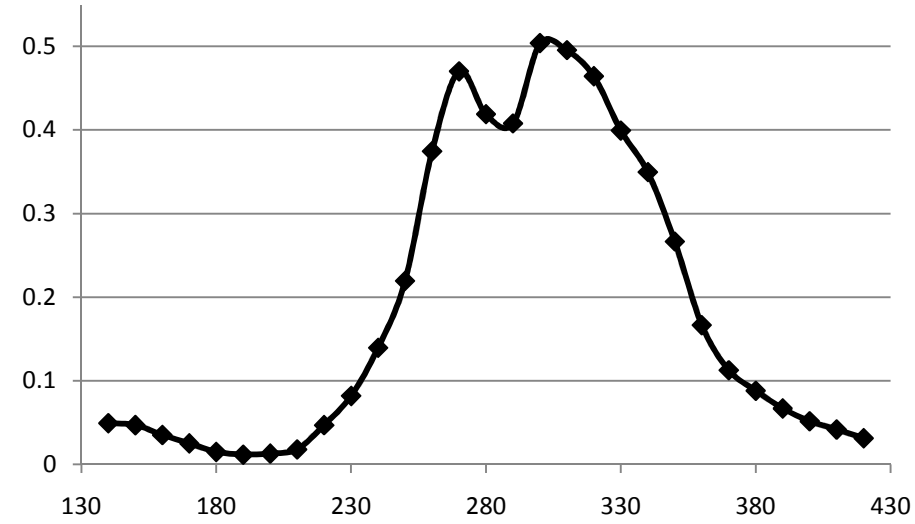


Figura 15 – Distribuição do desvio padrão segundo a velocidade para as acelerações verticais

Como se pode observar, a partir dos $200km/h$ e até perto de $350km/h$ os deslocamentos, as acelerações e também a incerteza sobre eles aumenta.

Capítulo 4: A distribuição amostral do coeficiente de correlação global

Como já foi referido, o principal objectivo deste trabalho é obter um teste de não linearidade, para isso vai-se estudar a quantidade $\eta_{**}^2 - \rho^2$ – "coeficiente de não linearidade". Este teste será deduzido utilizando a metodologia clássica dos testes de hipóteses.

Num teste estatístico clássico é conhecida (pelo menos aproximadamente) a distribuição de uma variável aleatória função da estatística de interesse e o objectivo é verificar se o valor obtido numa amostra é muito ou pouco provável. Se o valor for muito pouco provável recusa-se a hipótese nula. As hipóteses nulas que podem ser testadas seriam:

H_0: Toda a correlação é linear, ou

$H_0: \eta_{**}^2 - \rho^2$ não é significativamente diferente de zero, ou seja, existe parte não linear no modelo, ou

$H_0: \eta_{**}^2 = \rho^2$, "o par (X, Y), como variável aleatória bidimensional, é gaussiano"

Testar uma hipótese é verificar/medir se a amostra que eu tenho é fracamente ou fortemente compatível com a hipótese nula, medição que é feita com probabilidades. Face a uma amostra determina-se o valor de uma "quantidade (variável)" chamada Estatística de Teste. Usando a distribuição amostral daquela variável, determina-se a probabilidade do valor obtido para a estatística, supondo que se verifica a hipótese nula.

Se é muito provável que o valor observado ocorra no caso H_0, não rejeito H_0, se o valor observado for pouco provável sob a hipótese H_0, rejeito H_0.

Tendo então em mente, que para este efeito será necessário obter as distribuições de η_{**}^2 e ρ^2, supondo a normalidade com todas as suas propriedades, e determinar a estatística teste, surge a seguinte questão: Qual a melhor maneira de analisar a distribuição amostral do η_{**}^2, visto que a de ρ^2 já é conhecida? É recorrer à simulação.

Desta forma, obteve-se então de diversas maneiras aproximações da distribuição amostral do η_{**}^2 e o valor do ρ^2 foi usado como um dos controlos para a escolha da "melhor distribuição" do η_{**}^2.

O cálculo de η_{**}^2 para além das dificuldades usuais dos cálculos de amostragem, tem uma dificuldade suplementar que são as classes. Assim, tem-se que ter em consideração a amplitude das classes, a posição das classes e também o tamanho da amostra.

Para se testar se a metodologia nos vai permitir obter valores razoáveis para o η_{**}^2, vamos fixar alguns tamanhos da amostra (n) e fazer variar: 1º teste – as amplitudes das classes – Δh_x e Δh_y, 2º teste – a posição das classes.

Relativamente às amplitudes das classes, procura-se ver o efeito do Δh de forma a estimar η_{**}^2 correspondente à malha nula (pois trata-se de uma variável contínua). Por outro lado, queremos para um $\Delta h \neq 0$ pois é desta forma que numa primeira abordagem se conseguirá estudar a amostra de forma a tirar conclusões. Neste estudo será, também necessário, analisar o seu efeito no ρ^2.

Vamos considerar que a distribuição amostral da estatística η_{**}^2 está calculada correctamente, se se verificar cumulativamente as seguintes condições:

1. os valores estimados de ρ^2 estarem correctos;
2. a dependência da amplitude das classes está controlada;
3. a dependência da posição das classes está controlada (sabemos o efeito delas).

4.1. Resultados das Simulações

Neste trabalho considerou-se $\rho = 0$ e $\rho = 0{,}5$ para determinar η_{**}^2 através da simulação. A escolha da "melhor" distribuição será aquela em que os três pontos referidos anteriormente sejam satisfeitos e em que haja a garantia de que o desvio padrão em cada "amostra" é inferior a uma centésima.

Através da geração de uma binormal, com $\rho = 0$ (variáveis independentes) e de seguida com $\rho = 0{,}5$ (existe correlação entre as variáveis), pretendo estudar a variável aleatória $\eta_{**}^2 - \rho^2$ que corresponde ao "coeficiente de não linearidade". Assim, para esse efeito tive-se em consideração:

1. Tamanho das "classes", $\Delta h = 0{,}5$ e $\Delta h = 0{,}1$;
2. Os 3 tipos de "classes", $\left]0-\frac{\Delta h}{2}, 0+\frac{\Delta h}{2}\right[$, $]0, \Delta h[$ *vs* $]-\Delta h, 0[$ ou $\left]\bar{x}-\frac{\Delta h}{2}, \bar{x}+\frac{\Delta h}{2}\right[$ e $\left]\bar{y}-\frac{\Delta h}{2}, \bar{y}+\frac{\Delta h}{2}\right[$;
3. Considerando fixo a tamanho da amostra, $N = 30, 100, 300$ avaliar o efeito das classes nos η_{**}^2's e escolher o "óptimo". Esta análise pode ser feita através da construção de tabelas de cada uma das classes e respectivas amplitudes.

Nota: Podia-se ter em conta os desvios padrões amostrais em vez das amplitudes, ou seja, em vez de $\Delta h = 0{,}5$ tinha-se meio desvio padrão. Mas o que se verifica é que se na teoria se tem $\sigma = 1$, na prática a "grosso modo", também o temos.

Através do software *Matlab 7.7.0*, foram geradas m variáveis aleatórias bidimensionais normais. A partir destas foram estudadas as distribuições amostrais de: $\eta_{Y|X}^2$, $\eta_{X|Y}^2$, ρ^2 (calculado de duas maneiras diferentes – pela definição com dados agrupados em classes e usando a função"corr" do Matlab) e $\eta_{**}^2 - \rho^2$. Os resultados dos valores médios destas medidas encontram-se expostos nas seguintes tabelas[3]:

[3] No anexo C, encontram-se graficamente algumas das distribuições dos coeficientes de correlação global e linear (para N=100 e N=300) É de ter em atenção que os valores expostos aqui podem não corresponder aos descritos em anexo pois de cada vez que as variáveis são geradas os valores mudam.

$\Rightarrow \rho = 0 \ (\rho^2 = 0)$

Tabela 21 – Resultados dos valores estimados para r^2

Tamanho da amostra (N)		30		100		300	
Nº de avaliações (m)		100	1000	100	1000	100	1000
"Uma classe com zero como extremo"	$\Delta h = 0{,}5$	0,03463	0,03425	0,00896	0,01087	0,00355	0,00349
	$\Delta h = 0{,}1$	0,03592	0,03447	0,00912	0,01061	0,00314	0,00346
"Uma classe centrada em zero"	$\Delta h = 0{,}5$	0,03494	0,03464	0,00844	0,01033	0,00309	0,00338
	$\Delta h = 0{,}1$	0,03531	0,03451	0,00903	0,01054	0,00313	0,00344
"Uma classe centrada na média de x e de y"	$\Delta h = 0{,}5$	0,03809	0,03477	0,00978	0,01060	0,00295	0,00339
	$\Delta h = 0{,}1$	0,03528	0,03437	0,00909	0,01059	0,00318	0,00347
usando a função "corr" do matlab		0,03563	0,03451	0,00913	0,01060	0,00314	0,00345

Tabela 22 – Resultados dos valores estimados para $\eta^2_{Y|X}$

Tamanho da amostra (N)		30		100		300	
Nº de avaliações (m)		100	1000	100	1000	100	1000
"Uma classe com zero como extremo"	$\Delta h = 0{,}5$	0,24227	0,25571	0,10342	0,09907	0,03961	0,03834
	$\Delta h = 0{,}1$	0,65861	0,68078	0,37845	0,37029	0,16552	0,15933
"Uma classe centrada em zero"	$\Delta h = 0{,}5$	0,26684	0,25837	0,10038	0,09642	0,04030	0,03876
	$\Delta h = 0{,}1$	0,66922	0,67638	0,37200	0,37048	0,16236	0,16015
"Uma classe centrada na média de x e de y"	$\Delta h = 0{,}5$	0,25807	0,25489	0,10454	0,09825	0,03962	0,03895
	$\Delta h = 0{,}1$	0,63402	0,63553	0,36374	0,36270	0,16051	0,15887

Tabela 23 – Resultados dos valores estimados para $\eta^2_{X|Y}$

Tamanho da amostra (N)		30		100		300	
Nº de avaliações (m)		100	1000	100	1000	100	1000
"Uma classe com zero como extremo"	$\Delta h = 0{,}5$	0,25986	0,26347	0,09418	0,09709	0,03743	0,03787
	$\Delta h = 0{,}1$	0,67047	0,67549	0,37527	0,36470	0,15860	0,15932
"Uma classe centrada em zero"	$\Delta h = 0{,}5$	0,25714	0,25876	0,09741	0,09577	0,03978	0,03848
	$\Delta h = 0{,}1$	0,66183	0,67445	0,36503	0,36637	0,16384	0,15984
"Uma classe centrada na média de x e de y"	$\Delta h = 0{,}5$	0,26394	0,25573	0,09865	0,09648	0,03912	0,03841
	$\Delta h = 0{,}1$	0,63104	0,62927	0,35845	0,35941	0,16106	0,15886

$\Rightarrow \ \rho = 0{,}5 \ (\rho^2 = 0{,}25)$

Tabela 24 – Resultados dos valores estimados para r^2

Tamanho da amostra (N)		30		100		300	
Nº de avaliações (m)		100	1000	100	1000	100	1000
"Uma classe com zero como extremo"	$\Delta h = 0{,}5$	0,2516	0,2423	0,2557	0,2445	0,2400	0,2464
	$\Delta h = 0{,}1$	0,2545	0,2517	0,2647	0,2547	0,2505	0,2552
"Uma classe centrada em zero"	$\Delta h = 0{,}5$	0,2393	0,2432	0,2548	0,2459	0,2412	0,2437
	$\Delta h = 0{,}1$	0,2542	0,2525	0,2654	0,2549	0,2506	0,2539
"Uma classe centrada na média de x e de y"	$\Delta h = 0{,}5$	0,2485	0,2426	0,2548	0,2454	0,2415	0,2441
	$\Delta h = 0{,}1$	0,2558	0,2511	0,2646	0,2545	0,2507	0,2543
usando a função "corr" do matlab		0,2550	0,2525	0,2655	0,2553	0,2510	0,2548

Tabela 25 – Resultados dos valores estimados para $\eta^2_{Y|X}$

Tamanho da amostra (N)		30		100		300	
Nº de avaliações (m)		100	1000	100	1000	100	1000
"Uma classe com zero como extremo"	$\Delta h = 0{,}5$	0,4320	0,4205	0,3214	0,3128	0,2662	0,2732
	$\Delta h = 0{,}1$	0,7689	0,7547	0,5309	0,5232	0,3674	0,3691
"Uma classe centrada em zero"	$\Delta h = 0{,}5$	0,4273	0,4213	0,3233	0,3131	0,2672	0,2704
	$\Delta h = 0{,}1$	0,7533	0,7523	0,5325	0,5235	0,3667	0,3682
"Uma classe centrada na média de x e de y"	$\Delta h = 0{,}5$	0,4277	0,4177	0,3238	0,3108	0,2676	0,2713
	$\Delta h = 0{,}1$	0,7206	0,7137	0,5208	0,5164	0,3668	0,3682

Tabela 26 – Resultados dos valores estimados para $\eta^2_{X|Y}$

Tamanho da amostra (N)		30		100		300	
Nº de avaliações (m)		100	1000	100	1000	100	1000
"Uma classe com zero como extremo"	$\Delta h = 0{,}5$	0,4336	0,4185	0,3148	0,3107	0,2658	0,2729
	$\Delta h = 0{,}1$	0,7554	0,7532	0,5230	0,5214	0,3666	0,3710
"Uma classe centrada em zero"	$\Delta h = 0{,}5$	0,4170	0,4165	0,3123	0,3113	0,2670	0,2701
	$\Delta h = 0{,}1$	0,7569	0,7489	0,5288	0,5239	0,3671	0,3692
"Uma classe centrada na média de x e de y"	$\Delta h = 0{,}5$	0,4270	0,4144	0,3124	0,3123	0,2671	0,2703
	$\Delta h = 0{,}1$	0,7255	0,7186	0,5206	0,5165	0,3672	0,3688

Segundo Aïvazian[4], para um intervalo de confiança a 95% os valores de r deveriam estar de acordo com os da Tabela 27.

Observando a Tabela 21 e a Tabela 24, tomando em atenção que os valores do coeficiente de correlação linear estão ao quadrado e comparando com a Tabela 27, verifica-se que os valores médios amostrais encontram-se todos dentro dos intervalos apresentados na tabela.

[4] De acordo com a figura 18 pág 110.

Tabela 27 – Valores amostrais admitidos para r

	N=30	N=100	N=300
$\rho = 0$	$]-0{,}36;0{,}36[$	$]-0{,}2;0{,}2[$	$]-0{,}125;0{,}125[$
$\rho = 0{,}5$	$]0{,}14;0{,}71[$	$]0{,}33;0{,}63[$	$]0{,}4;0{,}58[$

Analisando agora todas as tabelas, facilmente se verifica que, para os casos estudados, o tamanho da amostra mais adequado é $N = 300$ e que a amplitude das classes é "óptima" para $\Delta h = 0{,}1$. Com o aumento do tamanho da amostra verifica-se claramente uma crescente precisão, sendo que para $\rho = 0{,}5$ esta se verifica na casa das centésimas qualquer que seja a classe que consideramos. Relativamente à dimensão da amostra de simulação, não existem diferenças notórias quer se considere $m = 100$ ou $m = 1000$. O mesmo se passa para os 3 tipos de classes. Através da análise gráfica (anexo C) das distribuições, parece sugerir que para $m = 1000$ se obtém uma maior estabilidade pois temos uma linha mais "suave", mas isto deve-se ao facto de termos mais valores condensados. Enquanto que, mais uma vez, se verifica que para diferentes tipos de classes não existem diferenças significativas que sejam visíveis.

Quanto à parte não linear, se fixarmos o tamanho da amostra em $N = 300$, ela é aproximadamente a mesma pois já foi obtida uma estabilidade razoável. Uma outra análise que se pode fazer é que a parte não linear decresce se considerarmos uma amostra cada vez maior, como era de se esperar.

Analisando e comparando os gráficos no anexo C, 1, 2 e 3 com 4, 5 e 6, do ponto 1, para $\rho = 0$, verifica-se que com classes "finas" o erro de localização do x não pesa muito, pois para o quantil de ordem 90%, η^2_{**} é sempre $\approx 0{,}45$ o que não se verifica nas mais "grossas".

Outra análise que se pode ter em atenção e tendo em mente Dodge and Rousson (1999), para classes "grossas" tem que se ter muito cuidado com a simetria e o primeiro elemento que se coloca, pois a assimetria é fortemente afectada pelas classes extremas e estas ficam fixadas pela primeira.

Para uma melhor compreensão, na Figura 17 encontra-se os valores de $\eta^2_{**} - r^2$.

A um nível de significância de 5%, o valor crítico é de cerca de $0{,}25$. Assim, o teste não rejeita a hipótese nula se $\eta^2_{**} - r^2 < 0{,}25$ e rejeita caso contrário.

Pode ainda verificar-se que a correlação não linear cresce imenso com as flutuações da amostragem, de facto, o valor mínimo encontrado é de aproximadamente $0{,}2$. A correlação global chega a atingir o valor de $0{,}45$ em cerca de 10% dos casos. Em 22% dos casos ultrapassa o valor de $0{,}40$, quando na verdade este valor deveria de ser zero porque foi gerado como tal.

Note-se ainda que os dois valores de η^2_{**} são sempre idênticos, isto é, a componente não linear é aproximadamente igual nos dois sentidos.

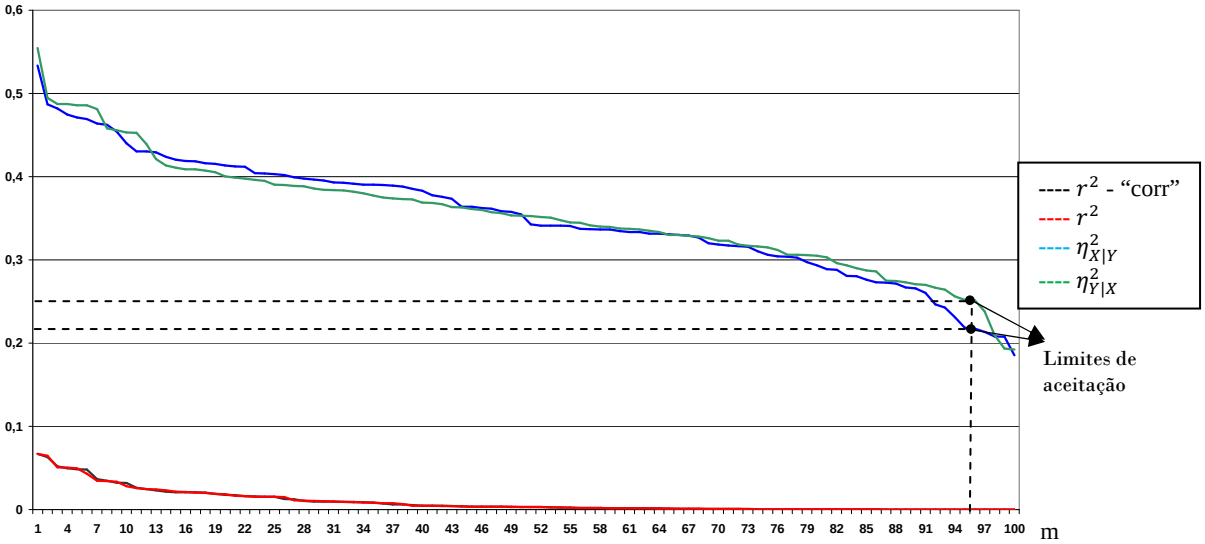


Figura 16 – Distribuição amostral simulada de η^2_{**} e r^2 para variáveis independentes, N=100, m=100, $\Delta h = 0.1$ e uma classe é centrada na média de *x* e de *y*.

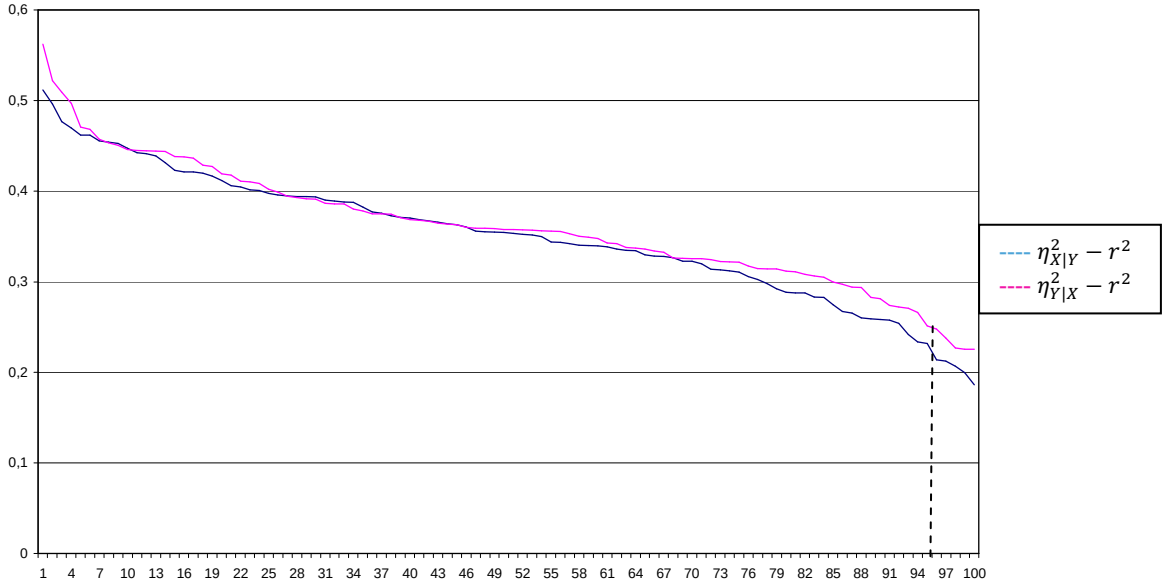


Figura 17 – Distribuição amostral simulada de $\eta^2_{**} - r^2$

Outras análises que podemos fazer é que, para $\rho = 0$, conforme vai aumentando a correlação linear a correlação não linear aumenta ainda mais e para $\rho = 0{,}5$, a componente não linear é aproximadamente constante (anexo C, ponto 2). O aumento da componente linear é acompanhado pelo aumento da componente não linear (linhas "paralelas").

De seguida, fez-se um estudo ao quociente $\frac{\eta^2_{**}}{r^2}$ (optou-se por, $\frac{\eta^2_{X|Y}}{r^2}$, pois como já se referiu os η^2_{**} são idênticos) e outro ao logaritmo do quociente. Sob a hipótese $\rho = 0$, considerou-se o logaritmo do quociente, para $\rho = 0{,}5$, apenas o quociente. Em cada um dos casos optou-se pelas classes em que uma delas possui o zero como extremo.

Analisando os gráficos do anexo C, ponto 3, podemos constatar que as diferenças surgem quando comparamos os coeficientes de correlação linear.

Quando $\rho = 0{,}5$, o crescimento não é tão acentuado. Este resultado está de acordo com os valores já obtidos, pois neste caso existe um certo paralelismo entre as duas componentes e os seus valores obtidos na simulação estão mais próximos.

Para $\rho = 0$ quer se considere $N = 100$ ou $N = 300$, não parece influenciar nos resultados. O mesmo não acontece para $\rho = 0{,}5$, existe uns valores simulados em que a componente linear é relativamente inferior à componente não linear ($N = 100$).

Com estes gráficos consegue-se ainda realçar a interferência do tamanho das classes na não linearidade, ou seja, para $\Delta h = 0{,}5$ a componente não linear é menor do que quando consideramos $\Delta h = 0{,}1$. Isto deve-se a quê? Quanto maior for o tamanho das classes menor será o valor da variabilidade global.

Nota:

Pelo que sabemos dos exemplos, fora os casos “Virgin” e Betão, em todos os outros este teste não detectaria a não linearidade.

Capítulo 5: Comentários Finais

Contrariamente ao problema do estudo da existência de uma relação linear entre duas variáveis, a análise da não linearidade raramente é abordada. Motivada por exemplos clássicos começou por fazer-se um estudo descritivo do coeficiente de correlação global em simultâneo com o coeficiente de correlação linear.

Recorrendo a simulações apresenta-se um estudo da distribuição amostral de $\eta_{**}^2 - r^2$, que fornece um teste para a não linearidade. O refinamento do teste constitui trabalho futuro. De momento, mostrou-se pouco capaz para detectar pequenas linearidades.

A possibilidade de uma regressão não linear fica condicionada pela componente não linear da correlação. Esta componente por vezes é de tal forma pequena que não chega a ser significativamente diferente de zero e portanto torna-se abusivo fazer um tratamento da não linearidade porque não nos irá trazer nada de novo.

Foram estudados diferentes exemplos de dados publicados, nos quais se analisou as alternativas. Temos de concluir que não há um processo sistemático de análise de dados bidimensionais, pelo que um estudo cuidadoso obriga à análise das diversas alternativas.

Referências Bibliográficas

Aïvazian, S. (1970). *Étudie Statistique Des Dépendences*, Éditions De Moscou.

Bessa, C. (2007). *Explorações do Coeficiente de Correlação Global*, relatório para a cadeira de Simulação, do Mestrado Em Estatística Aplicada e Modelação, FEUP.

Birkes, D. and Dodge, Y. (1993). *Alternative Methods of Regression*, John Wiley & Sons, New York.

Blakeman, J. (1905). *On tests for the linearity of regression in frequency distributions*, Biometrika, 4, 332-350.

Calheiros, F. e Faria, S. (2002). *Sobre a Assimetria e "Achatamento" de Misturas de Distribuições*, Actas do VIII Congresso da SPE, pág.171-178.

Calheiros, F. (2002). *Mean and standard deviation in mixtures*, Joclad, pag. 24-27, 2002.

Calot, G. (1969). *Cours de statistique descriptive*, Ed. Dunod.

Dagnelie, P. (1973). *Estatística-Teoria e Métodos* 1º volume, Publicações Europa-América, Biblioteca Universitária, nº 331[5].

Dodge, Y. and Rousson, V. (1999). *Testing Linearity in Regression Model*, ASMDA, 57-61.

Durand, J. F. and Sabatier, R. (1997). *Additive Splines for Partial Least Squares Regression*, Journal of the American Statistical Association, Vol. 92, No. 440, pp. 1546-1554.

Fisher, R. A. (1925). *Statistical Methods for research Workers*, Ed. Oliver and Boyd.

Friedman, J. H. (1991). *Multivariate Adaptive Regression Splines (with discussion)*, Annals of Statistics, 19, 1-141.

Giloni, A.; Simonoff, J. S. and Sengupta, B. (2005). *Robust Weighted LAD Regression*, Computational Statistics & Data Analysis, Volume 50, Issue 11, 20 July 2006, pages 3124-3140.

Goodman, J. (2005). *Simple Sampling of Gaussians*, Courant Institute of Mathematical Sciences, NYU.

Grais, B. (1974). *Méthodes Statistiques*, Ed Dunod.

Iman, R. L.; Davenport, J. M. and Zeigler, D. K. (1980). *Latin hypercube sampling (program user's guide)*.

Kelley, T. L. (1935). *An unbiased correlation ratio measure*, proceedings of the National Academy of Sciences of the United States of America, Vol. 21, No. 9, 554-559.

Kendall, M. G. and Stuart, A. (1969). *The Advanced Theory of Statistics*, Vol 1, London, Charles Griffin.

[5] Tradução do Professor A. St. Aubyn, sem data edição e, com o nome do autor errado na capa!!!

Koenker, R. and Bassett, G. (1978). *Regression Quantiles*, Econometrica, Vol. 46, Nº1, pp. 33-50.

Kutner, M. H.; Nachtsheim, C. and Neter, J. (2004). *Applied Linear Regression Models*, McGraw-Hill.

Magalhães, R. (2006). *Um Sistema de Informação e Modelo de Diagnóstico Diferencial no Âmbito de um Rastreio Populacional*, Tese de Mestrado Em Estatística Aplicada e Modelação, FEUP.

McKay, M. D.; Beckman, R.J. and Conover, W.J. (May 1979). *A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code*, Technometrics, Vol. 21, Nº 2, pp. 239-245.

Moura, G.; Pinheiro, M.; Silva, R.; Miranda, I.; Afreixo, V.; Dias, G.; Freitas, A.; Oliveira, J. L. e Santos, M. (2005). *Comparative context analysis of codon pairs on na ORFeome scale*, Genome Biology, 6, R28.

Pestana, D. e Velosa, S. (2006). *Introdução à Probabilidade e à Estatísta*, Murteira, 2ª Edição, Lisboa, Fundação Calouste Gulbenkian.

Pinto, J. R.; Calçada, R. e Calheiros, F. (2008). *Modelação probabilística da acção de tráfego em pontes ferroviárias em linhas de alta velocidade*, Encontro Nacional Betão Estrutural 2008, Guimarães

Santos, J. A. Amaral and Neves, M. Manuela (2007). *A Local Maximum Likelihood Estimator for Poisson Regression*, Metrika, Vol 68, 257-270

Sereno, F. (2000). *The Aplication of Radial Basis Functions and Support Vector Machines to the Foetal Weight Prediction*, Intell Eng Syst Through Artif Neural Networks, 10: 801-806.

Student (1913). *The Correction to be made to the Correlation Ratio for Grouping*, Biometrika, Vol. 9, No. 1/2 pp. 316-320.

Tomassone, R. ; Lesquoy, E. et Miller C. (1983). *La Régression*, Ed. Masson.

Weisstein, Eric W., *Bivariate Normal Distribution,* from MathWorld--A Wolfram. Web Resource: http://mathworld.wolfram.com/BivariateNormalDistribution.html .

Weisstein, Eric W., *Sheppard's Correction,* from MathWorld--A Wolfram Web Resource: http://mathworld.wolfram.com/SheppardsCorrection.html .

Anexo A: Distribuição de Variáveis Binormais

O método de Box Muller surgiu com o intuito de gerar duas distribuições normais e independentes a partir de duas distribuições uniformes e independentes, Goodman (2005). Começando por definir a função distribuição de variáveis binormais, temos que esta pode ser pode ser definida pela seguinte função de densidade de probabilidade, tal como nos refere Weisstein:

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}} exp\left[-\frac{z}{2(1-\rho_{XY}^2)}\right],$$

onde

$$z = \frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho_{XY}(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}.$$

Assim, o par (X,Y), de variáveis aleatórias independentes normalmente distribuídas, com médias $\mu = 0$ e variâncias $\sigma^2 = 1$, possui a seguinte função densidade de probabilidade: $f(x,y) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \times \frac{1}{\sqrt{2\pi}} e^{-y^2/2} = \frac{1}{2\pi} e^{-(x^2+y^2)/2}$. A partir daqui o método de Box Muller consiste em considerar o par de variáveis aleatórias em coordenadas polares, (R,Θ), em que $0 \le \Theta < 2\pi$, $X = R\cos(\Theta)$ e $Y = R\sin(\Theta)$ e verificar que X e Y podem ser obtidas a partir de variáveis uniformes, U_1 e U_2.

Começa-se por verificar que Θ é uniformemente distribuída no intervalo $[0,2\pi]$ é obtida considerando como sendo $\Theta = 2\pi U_1$. A função de distribuição de R é dada por

$$G(R) = P(R \le r) = \int_{r'=0}^{r} \int_{\Theta=0}^{2\pi} \frac{1}{2\pi} e^{-r'^2/2}\, r\, dr d\theta = \int_{r'=0}^{r} e^{-r'^2/2}\, r\, dr\,,$$

onde R é facilmente obtido tendo em consideração um certo "truque" do cálculo de integração. Ou seja, considerando $r^2/2 = s$, $r\,dr = ds$, tal que $r' = r$ e $s = r^2/2$, tem-se que $G(r) = \int_{s=0}^{r^2/2} e^{-s} ds = 1 - e^{-r^2/2} \Rightarrow G(R) = 1 - e^{-R^2/2} = 1 - U_2$ e portanto $R = \sqrt{-2\ln(U_2)}$.

Anexo B: Funções do Matlab

- No Matlab foi elaborada uma função que constrói uma Binormal de média zero e desvio padrão um (facilmente se produz uma análoga, para o caso em que a média e o desvio são diferentes dos valores referidos, alterando-se os valores, convenientemente de $n1$ e $n2$ de acordo com o que foi visto na secção 2.3). A função construída foi a seguinte, em que $V1$ e $V2$ é o par binormal gerado usando Box Muller:

```
function[V1,V2]=BiNormal(n)
A1=rand(n,1);
A2=rand(n,1);
V1=zeros(n,1);
V2=zeros(n,1);
for i=1:n
    n1=sqrt(-2*log(A1(i)))*sin(2*pi*A2(i));
    n2=sqrt(-2*log(A1(i)))*cos(2*pi*A2(i));
V1(i)=n1;
V2(i)=n2;
end
% binormal com variáveis não independentes, ρ=0.5:
% V3=V2*sqrt(0.5)+0.5*V1
```

No Matlab se a correr, a título de exemplo para n=10 obtêm-se:

```
>> n=10
n =
   10
>> [V1,V2]=BiNormal(n)
V1 =
  -0.2122
  -1.6525
  -0.4715
  -1.1980
   0.4292
   0.4116
  -0.4940
  -1.4088
   0.3352
  -0.7888
V2 =
  -0.2394
   0.4458
   0.8813
  -0.0889
   0.2144
  -0.6115
   1.1508
   2.4485
  -0.5302
   0.9993
```

- De seguida, foi feita outra função com o intuito de representar a tabela de contingência – A, resultante da aplicação de diferentes tamanhos de "classes" no x e no y – (deltah), ou seja, as variáveis condicionadas por intervalos realizam uma partição dos conjuntos de valores das variáveis que se encontram apresentados na matriz A. Os valores de xm e ym correspondem aos valores médios das classes. Apresenta-se de seguida três funções, em que a primeira possui o valor zero como centro de uma das classes, $]-deltah/2, deltah/2[$, a segunda o valor zero é um dos extremos das classes, $]0, deltah[$ ou $]-deltah, 0[$ e na terceira a média de x e a de y são os centros, $]\bar{x} - deltah/2, \bar{x} + deltah/2[$ e $]\bar{y} - deltah/2, \bar{y} + deltah/2[$:

```
function[xm,ym,A]=Tabela2(V1,V2,n,deltah)
% o valor zero é o centro de uma das "classes"
m1=max(V1)+deltah;
m2=min(V1)-deltah;
x1=deltah/2:deltah:m1;
x2=deltah/2:deltah:-m2;
x3=-x2(length(x2):-1:1);
x=[x3 x1]';
for i=1:length(x)-1
    xm(i)=(x(i)+x(i+1))/2;
end
n1=max(V2)+deltah;
n2=min(V2)-deltah;
y1=deltah/2:deltah:n1;
y2=deltah/2:deltah:-n2;
y3=-y2(length(y2):-1:1);
y=[y3 y1]';
for j=1:length(y)-1
    ym(j)=(y(j)+y(j+1))/2;
end
A=zeros(length(x),length(y));
for i=1:length(x)-1
    for j=1:length(y)-1
        for z=1:n
            if (V1(z)>=x(i) & V1(z)<x(i+1)) & (V2(z)>=y(j) &
V2(z)<y(j+1))
                som=1;
             A(i,j)=A(i,j)+som;
            end
        end
    end
end
z1=zeros(length(x)-1,1);
z2=zeros(1,length(y)-1);
for i=1:length(x)-1
    z1(i)=sum(A(i,:))+z1(i);
    A(i,length(y))=A(i,length(y))+z1(i);
end
for j=1:length(y)-1
    z2(j)=sum(A(:,j))+z2(j);
    A(length(x),j)=A(length(x),j)+z2(j);
end
A(length(x),length(y))=A(length(x),length(y))+n;
```

A sublinhado encontram-se as mudanças feitas à função anterior:

```
function[xm,ym,A]=Tabela(V1,V2,n,deltah)
% o valor zero é um dos extremos das classes
m1=max(V1)+deltah;
m2=min(V1)-deltah;
x1=0:deltah:m1;
x2=deltah:deltah:-m2;
x3=-x2(length(x2):-1:1);
x=[x3 x1]';
for i=1:length(x)-1
    xm(i)=(x(i)+x(i+1))/2;
end
n1=max(V2)+deltah;
n2=min(V2)-deltah;
y1=0:deltah:n1;
y2=deltah:deltah:-n2;
y3=-y2(length(y2):-1:1);
y=[y3 y1]';
for j=1:length(y)-1
    ym(j)=(y(j)+y(j+1))/2;
end
A=zeros(length(x),length(y));
for i=1:length(x)-1
    for j=1:length(y)-1
        for z=1:n
            if (V1(z)>=x(i) & V1(z)<x(i+1)) & (V2(z)>=y(j) &
V2(z)<y(j+1))
                som=1;
             A(i,j)=A(i,j)+som;
            end
        end
    end
end
z1=zeros(length(x)-1,1);
z2=zeros(1,length(y)-1);
for i=1:length(x)-1
    z1(i)=sum(A(i,:))+z1(i);
    A(i,length(y))=A(i,length(y))+z1(i);
end
for j=1:length(y)-1
    z2(j)=sum(A(:,j))+z2(j);
    A(length(x),j)=A(length(x),j)+z2(j);
end
A(length(x),length(y))=A(length(x),length(y))+n;
```

```
function[xm,ym,A]=Tabela3(V1,V2,n,deltah)
% a média de V1 e de V2 são centros das "classes"
m1=max(V1)+deltah;
m2=min(V1)-deltah;
x1=abs(mean(V1)+deltah/2):deltah:m1;
x2=abs(mean(V1)-deltah/2):deltah:-m2;
x3=-x2(length(x2):-1:1);
x=[x3 x1]';
for i=1:length(x)-1
    xm(i)=(x(i)+x(i+1))/2;
end
n1=max(V2)+deltah;
n2=min(V2)-deltah;
y1=abs(mean(V2)+deltah/2):deltah:n1;
y2=abs(mean(V2)-deltah/2):deltah:-n2;
y3=-y2(length(y2):-1:1);
y=[y3 y1]';
for j=1:length(y)-1
    ym(j)=(y(j)+y(j+1))/2;
end
A=zeros(length(x),length(y));
for i=1:length(x)-1
    for j=1:length(y)-1
        for z=1:n
            if (V1(z)>=x(i) & V1(z)<x(i+1)) & (V2(z)>=y(j) &
V2(z)<y(j+1))
                som=1;
             A(i,j)=A(i,j)+som;
            end
        end
    end
end
z1=zeros(length(x)-1,1);
z2=zeros(1,length(y)-1);
for i=1:length(x)-1

    z1(i)=sum(A(i,:))+z1(i);
    A(i,length(y))=A(i,length(y))+z1(i);
end
for j=1:length(y)-1
    z2(j)=sum(A(:,j))+z2(j);
    A(length(x),j)=A(length(x),j)+z2(j);
end
A(length(x),length(y))=A(length(x),length(y))+n;
```

Correndo no matlab, apenas uma das funções, obtém-se o seguinte (os valores de x e de y não deveriam aparecer porque não são elementos de saída da função, mas apresenta-se aqui para uma melhor compreensão da função):

```
>> [xm,ym,A]=Tabela2(V1,V2,n,deltah)
x =
   -1.7500
   -1.2500
   -0.7500
   -0.2500
    0.2500
    0.7500
y =
   -0.7500
   -0.2500
    0.2500
    0.7500
    1.2500
    1.7500
    2.2500
    2.7500
xm =
   -1.5000   -1.0000   -0.5000         0    0.5000
ym =
   -0.5000         0    0.5000    1.0000    1.5000    2.0000    2.5000
A =
     0     0     1     0     0     0     1     2
     0     1     0     1     0     0     0     2
     0     0     0     2     0     0     0     2
     0     1     0     0     0     0     0     1
     2     1     0     0     0     0     0     3
     2     3     1     3     0     0     1    10
```

- Para determinar o valor e eta – coeficiente de correlação global e de rho – coeficiente de linearidade, elaborou-se a seguinte função:

```
function[xj,yi,medy,medx,rho,etaYX,etaXY]=etarho(ym,xm,n,A)
lx=length(xm);
ly=length(ym);
ni=A(1:lx,ly+1);
nj=A(lx+1,1:ly);
for j=1:ly
    if A(lx+1,j)~=0
        xj(j)=1/A(lx+1,j)*(xm*A(1:lx,j));
    else
        xj(j)=0;
    end
end
medx=1/n*xm*ni;
mx=ones(1,lx)*medx;
desvx=sqrt(1/n*(xm-mx).^2*ni);

for i=1:lx
    if A(i,ly+1)~=0
        yi(i)=1/A(i,ly+1)*ym*A(i,1:ly)';
    else
        yi(i)=0;
    end
end
medy=1/n*ym*nj';
my=ones(1,ly)*medy;
desvy=sqrt(1/n*nj*(ym-my).^2');
cov1=0;
for i=1:lx
    for j=1:ly
        cov2=A(i,j)*xm(i)*ym(j);
        cov1=cov1+cov2;
    end
end
covXY=1/n*cov1-medx*medy;
rho=covXY/(desvx*desvy);
ny=(yi-ones(1,length(yi))*medy).^2*ni;
dy=nj*(ym-my).^2';
etaYX=ny/dy;
nx=(xj-ones(1,length(xj))*medx).^2*nj';
dx=(xm-mx).^2*ni;
etaXY=nx/dx;
```

Para apresentar um exemplo em matlab usou-se os resultados que se obteve através da função *Tabela2*.

```
>> [xj,yi,medy,medx,rho,etaYX,etaXY]=etarho(ym,xm,n,A)
xj =
    0.5000  -0.1667  -1.5000  -0.6667   0   0  -1.5000
yi =
    1.5000    0.5000    1.0000    0   -0.3333
```

(xi e yj são as médias de cada uma das classes)

```
medy =
    0.5000
medx =
   -0.4500
```

(são as medias de x e de y respectivamente)

```
rho =
   -0.7249
etaYX =
    0.6444
etaXY =
    0.7671
```

- A próxima função calcula m valores dos coeficientes e depois retorna a sua média (para todo o processo basta utilizar esta função pois ela faz referência às outras três funções descritas anteriormente):

```
function
[vrho,detaYX,detaXY,drho,deta2YX,deta2XY,d2rho,deta3YX,deta3XY,
d3rho]=medias(n,m,deltah)
for z=1:m

   [V1,V2]=Binormalrho(n);
   vrho(z)=corr(V1,V2)^2;
     %o valor zero é um dos extremos das classes
        [xm,ym,A]=Tabela(V1,V2,n,deltah);
        [xj,yi,medy,medx,rho,etaYX,etaXY]=etarho(ym,xm,n,A);
        detaYX(z)=etaYX;
        detaXY(z)=etaXY;
        drho(z)=rho^2;
     %o valor zero é o centro de uma das "classes"
       [xm,ym,A]=Tabela2(V1,V2,n,deltah);
       [xj,yi,medy,medx,rho,etaYX,etaXY]=etarho(ym,xm,n,A);
       deta2YX(z)=etaYX;
       deta2XY(z)=etaXY;
       d2rho(z)=rho^2;
     % a média de V1 e de V2 são centros das "classes"
       [xm,ym,A]=Tabela3(V1,V2,n,deltah);
       [xj,yi,medy,medx,rho,etaYX,etaXY]=etarho(ym,xm,n,A);
       deta3YX(z)=etaYX;
       deta3XY(z)=etaXY;
       d3rho(z)=rho^2;
end
vrho= vrho';
detaYX=detaYX'; detaXY=detaXY'; drho=drho';
deta2YX=deta2YX';deta2XY=deta2XY'; d2rho=d2rho';
deta3YX=deta3YX'; deta3XY=deta3XY'; d3rho=d3rho';
%mdetaYX=mean(detaYX);
%mdetaXY=mean(detaXY);
%mdrho=mean(drho);
%mvrho=mean(dvrho);
```

No matlab, obtendo apenas algumas das saídas da função *medias*, para n=100, m=500, deltah=0.5 (usando a função *Tabela*):

```
>> n=100
n =
  100
>> m=500
m =
  500
>> deltah=0.5
deltah =
   0.5000
>> [mdetaYX,mdetaXY,mdrho]=medias(n,m,deltah)
mdetaYX =
   0.1010
mdetaXY =
   0.0968
mdrho =
   0.0099
```

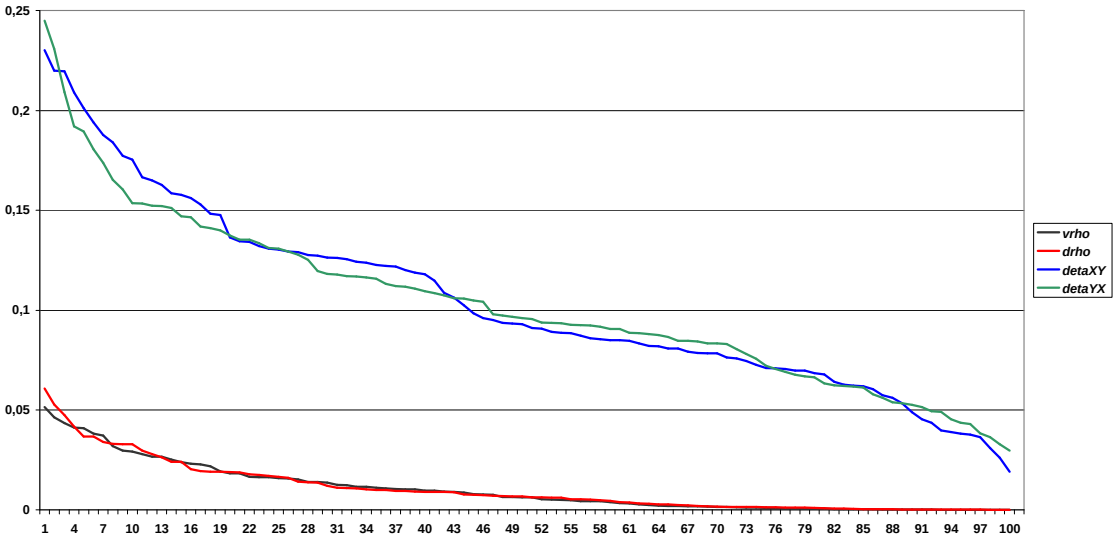
- A função *kurtskew* determina os valores da assimetria e do "achatamento" das variáveis *X* e *Y* associadas a uma tabela de contingência.

```
function[gama1x,gama1y,gama2x,gama2y]=kurtskew(A,n,xm,ym)
x1=length(xm);
y1=length(ym);
a=size(A);
x=zeros(n,1);
y=zeros(n,1);
csx=cumsum(A(1:x1,a(2)));
csy=cumsum(A(a(1),1:y1));
a1=[0 csx'];
a2=[0 csy];
for i=1:a(1)-1
    for j=a1(i)+1:a1(i+1)
        x(j)=xm(i);
    end
end
for i=1:a(2)-1
    for j=a2(i)+1:a2(i+1)
        y(j)=ym(i);
    end
end
gama1x=skewness(x);
gama1y=skewness(y);
gama2x=kurtosis(x)-3;
gama2y=kurtosis(y)-3;
```

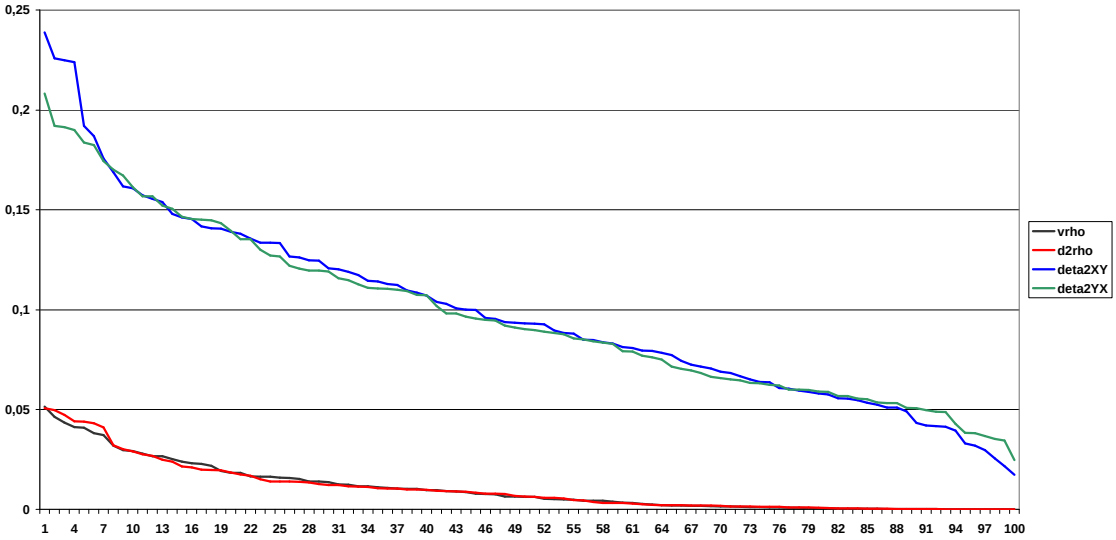
Ponto 1: Representações gráficas de Binormais ($\rho = 0$)

N=100 & m=100; deltah=0.5

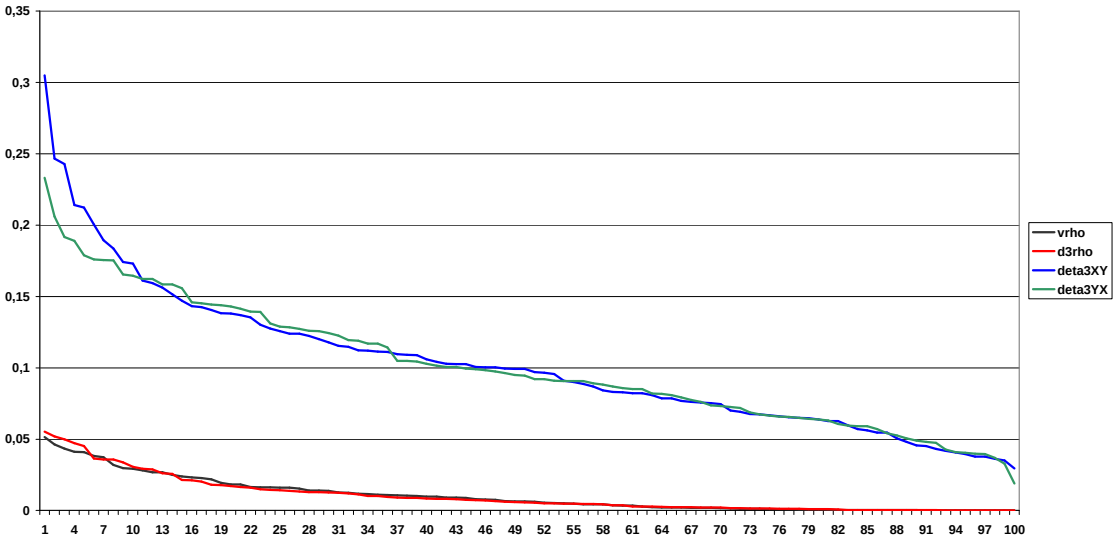
1. classes em que o zero é extremo duma classe



2. classes em que uma é centrada em zero

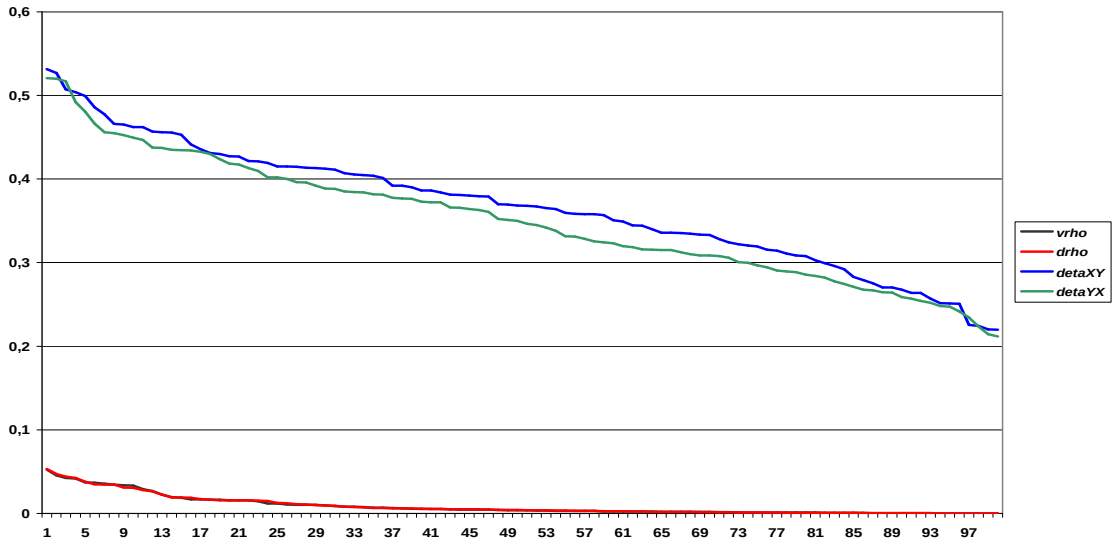


3. classes em que uma é centrada na média de x e de y

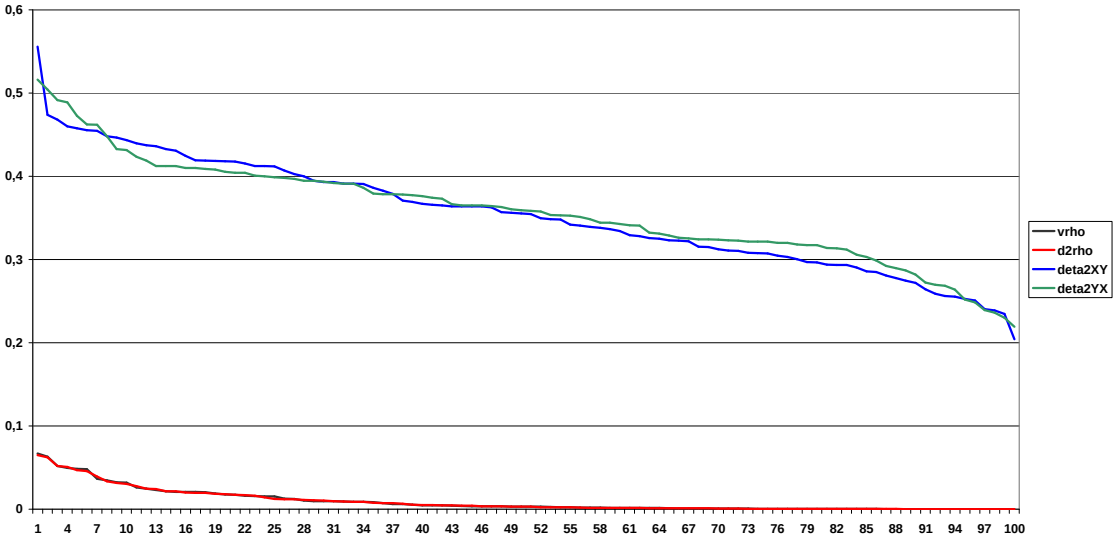


N=100 & m=100; deltah=0.1

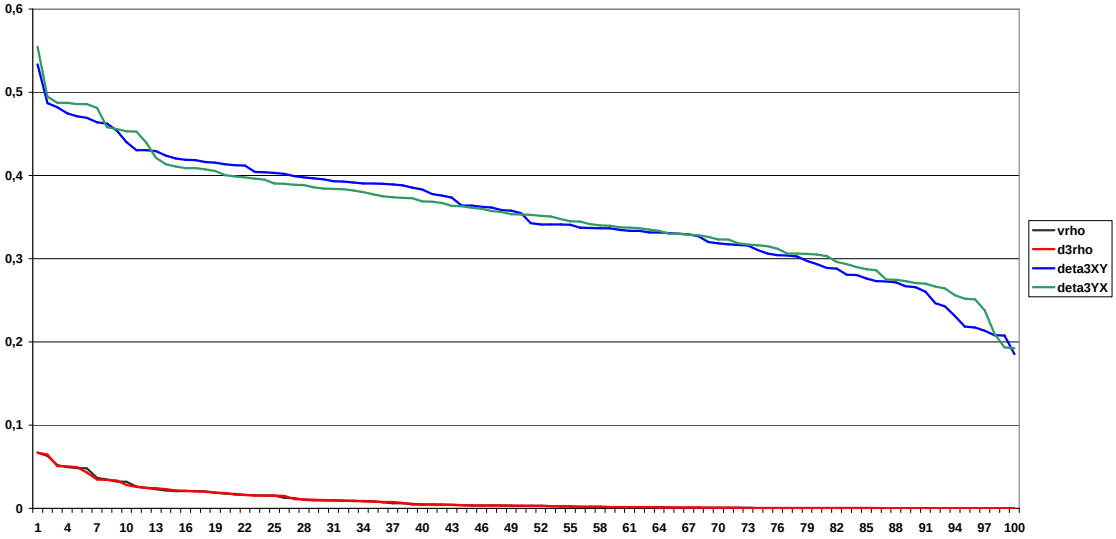
4. classes em que o zero é extremo duma classe



5. classes em que uma é centrada em zero



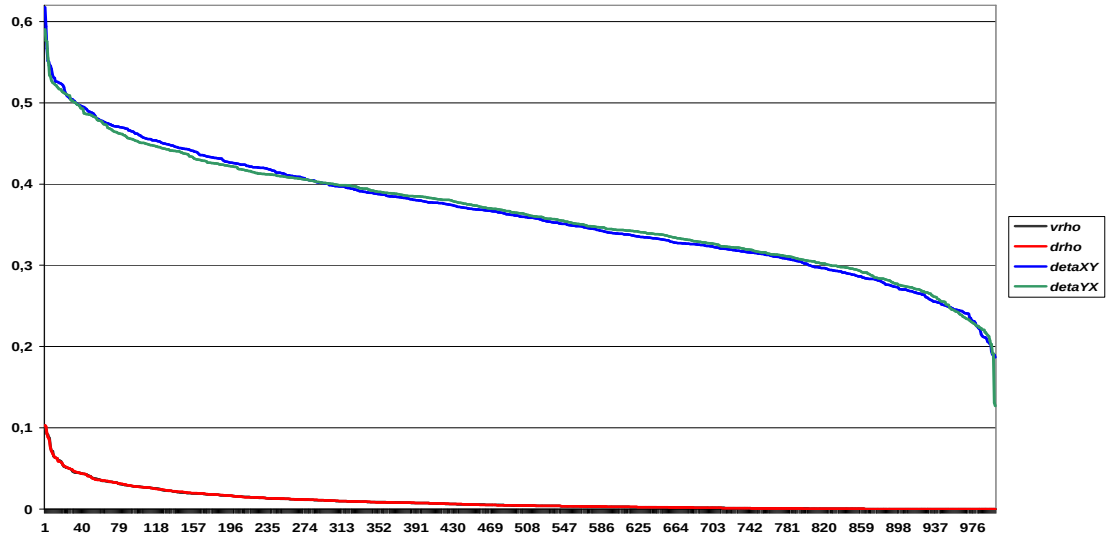
6. classes em que uma é centrada na média de x e de y



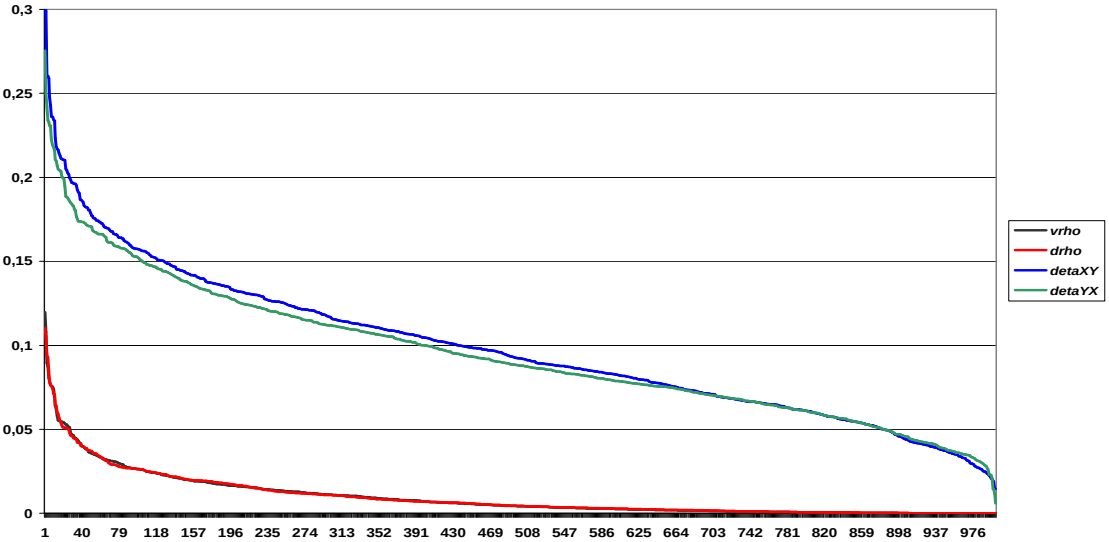
Ponto 2: Outras representações gráficas de Binormais (classes em que o zero é extremo duma classe)

- $\rho = 0$

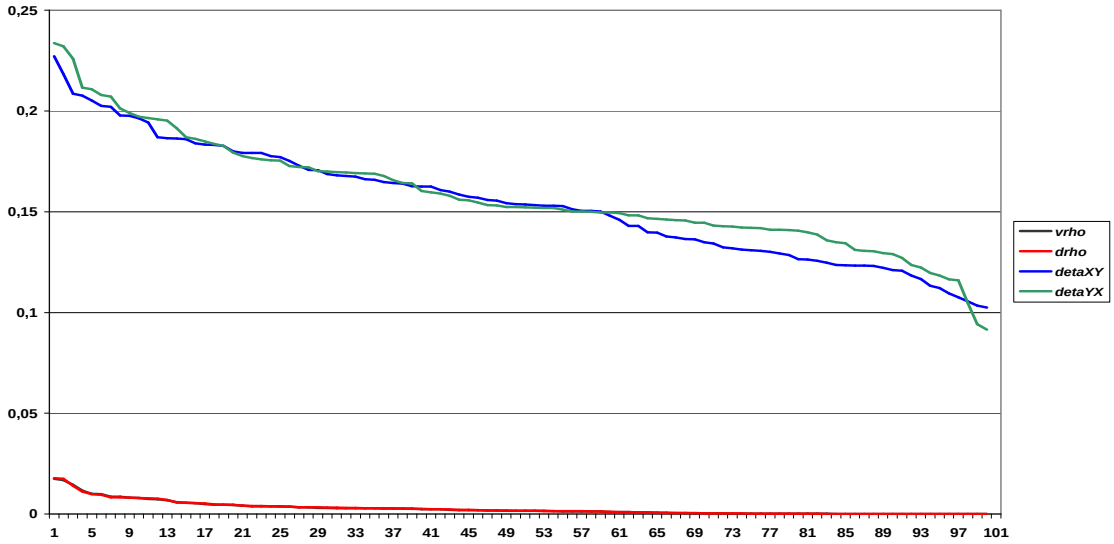
N=100 & m=1000; deltah=0.1



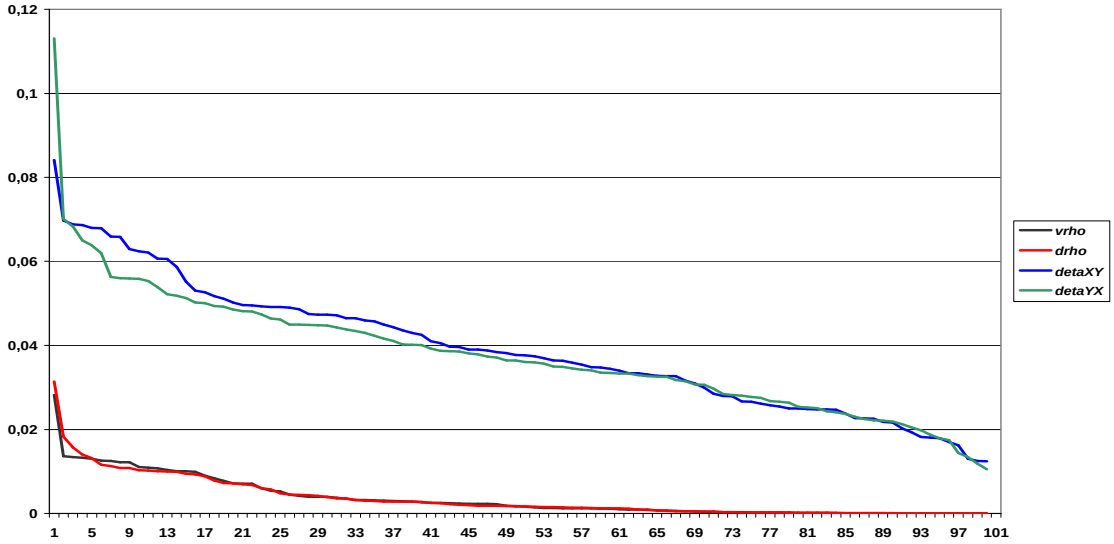
N=100 & m=1000; deltah=0.5



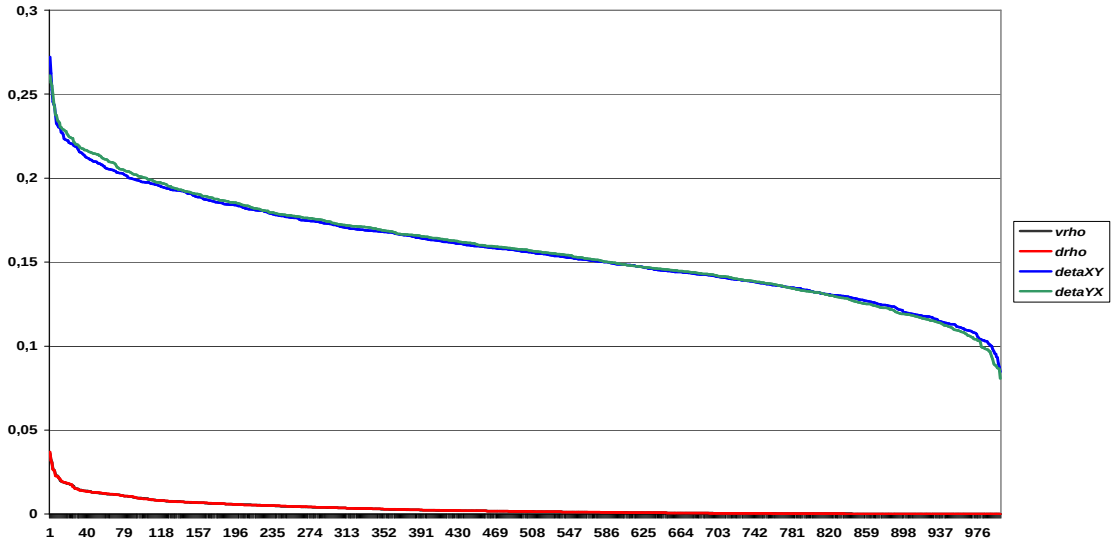
N=300 & m=100; deltah=0.1



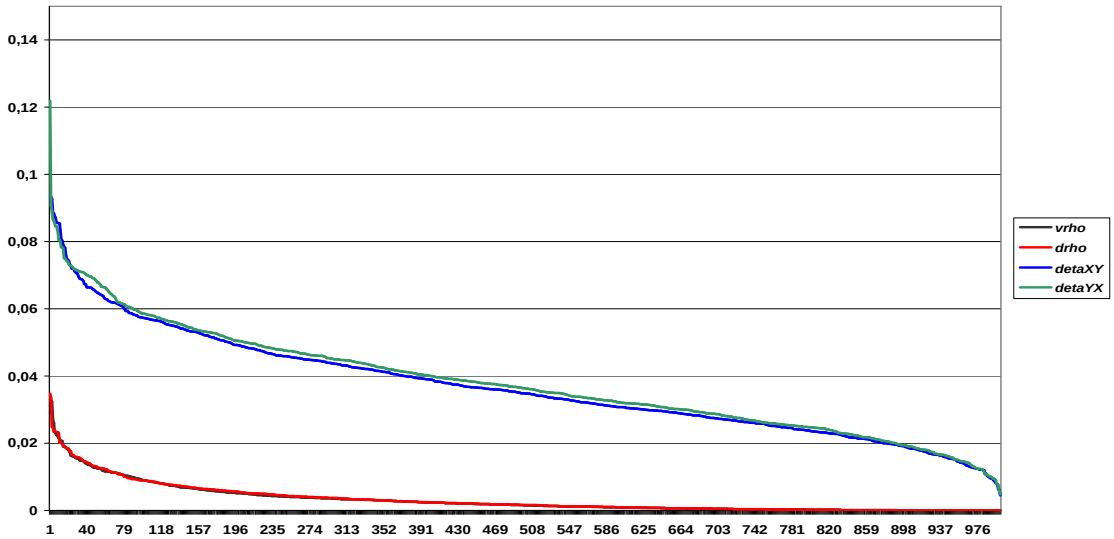
N=300 & m=100; deltah=0.5



N=300 & m=1000; deltah=0.1

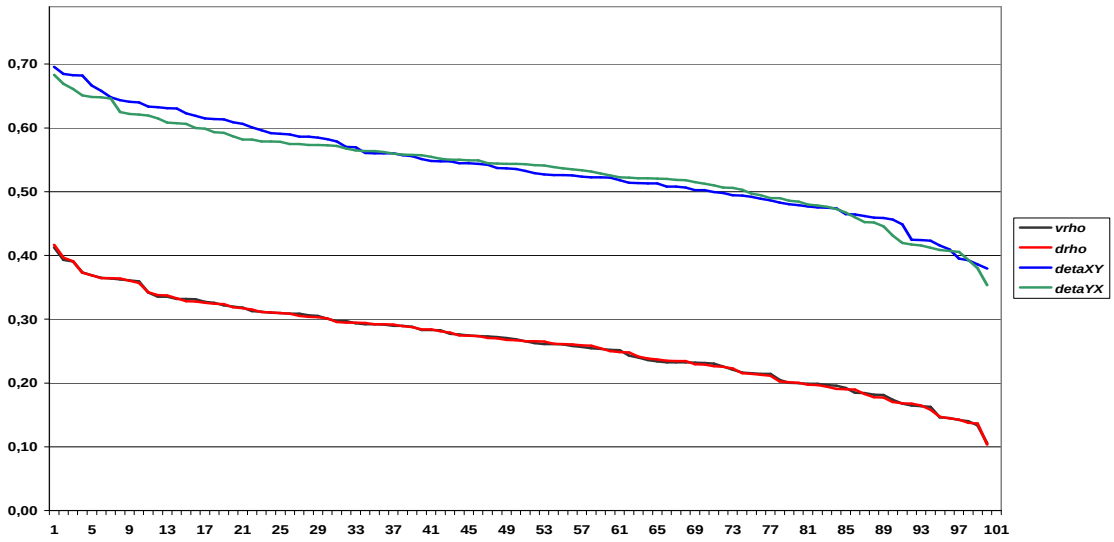


N=300 & m=1000; deltah=0.5

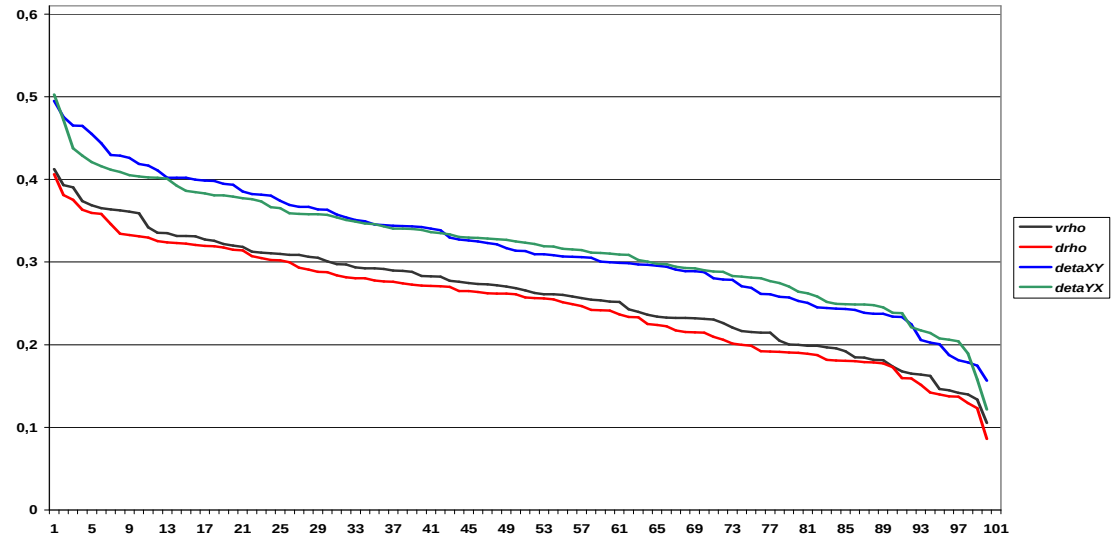


- $\rho = 0,5$

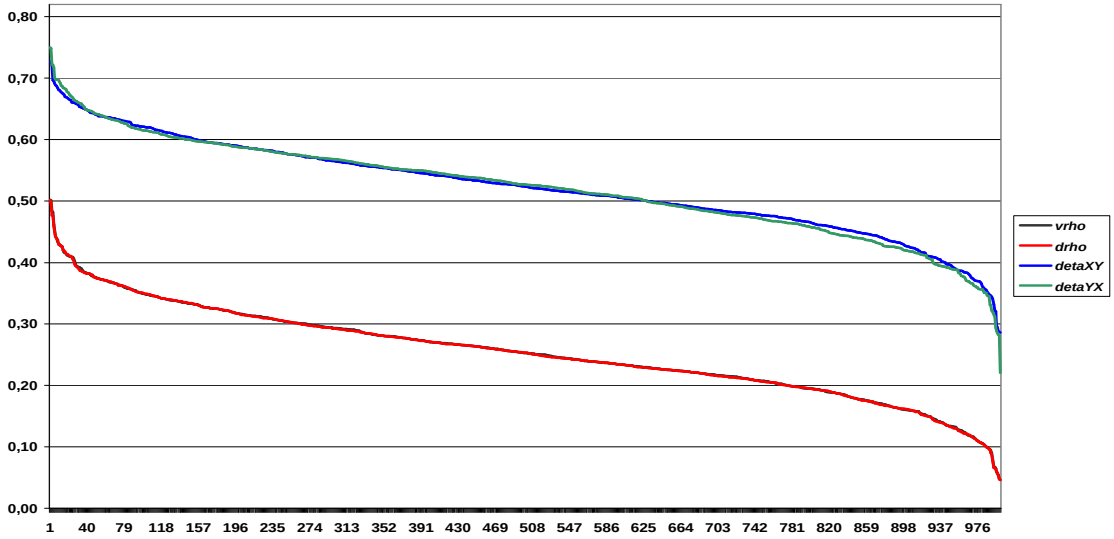
N=100 & m=100; deltah=0.1



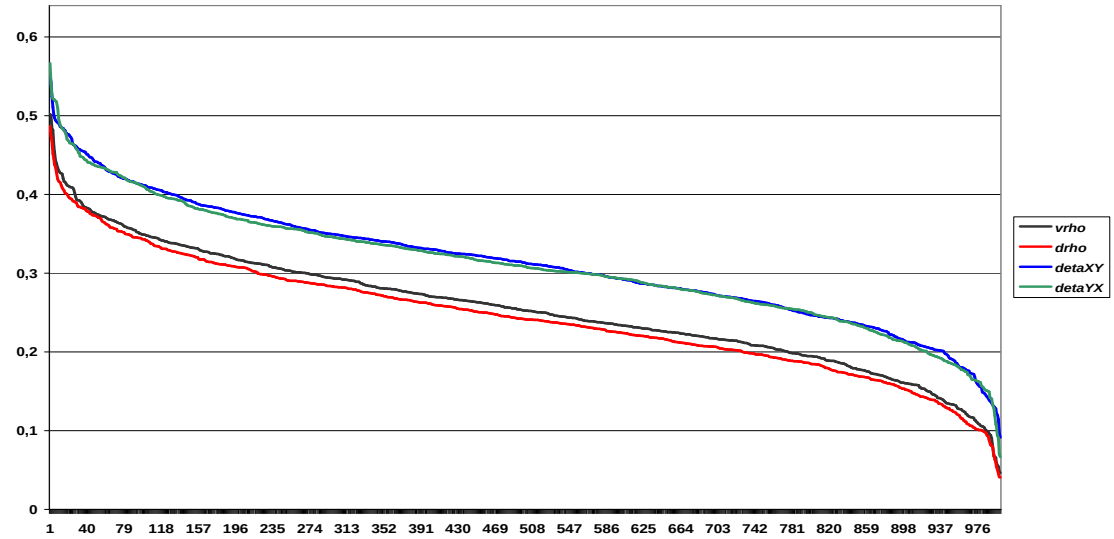
N=100 & m=100; deltah=0.5



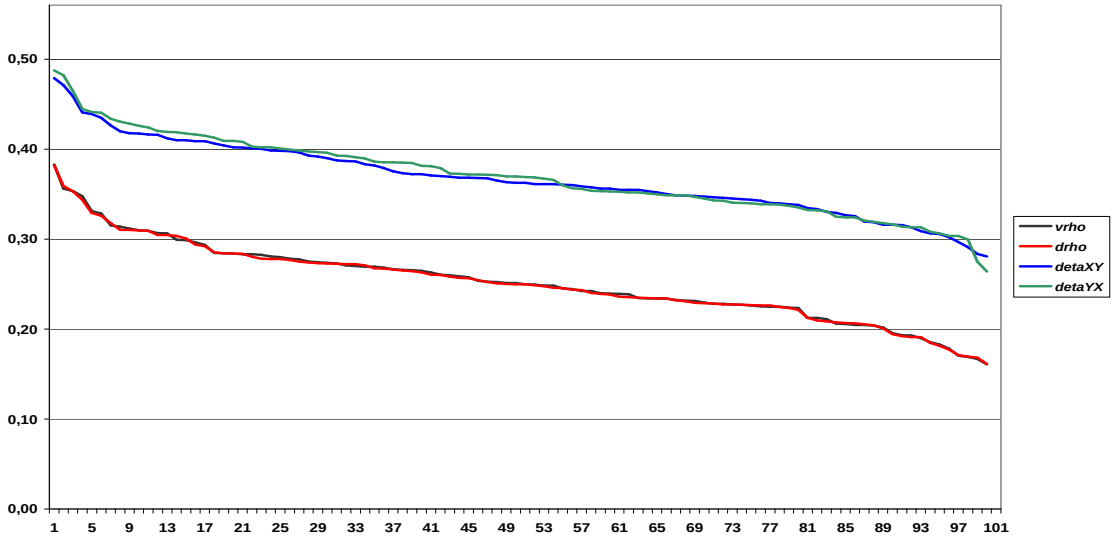
N=100 & m=1000; deltah=0.1



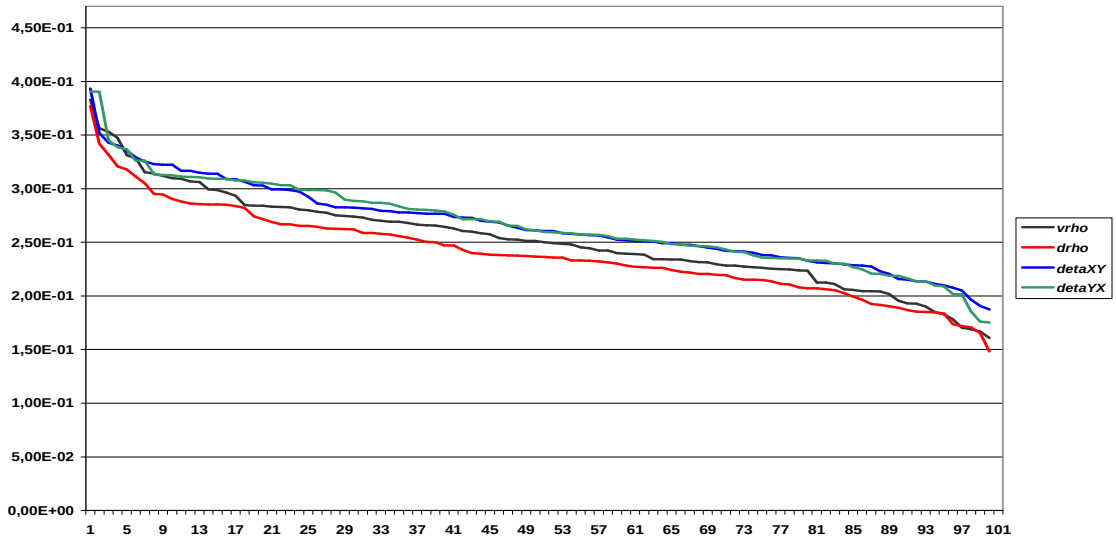
N=100 & m=1000; deltah=0.5



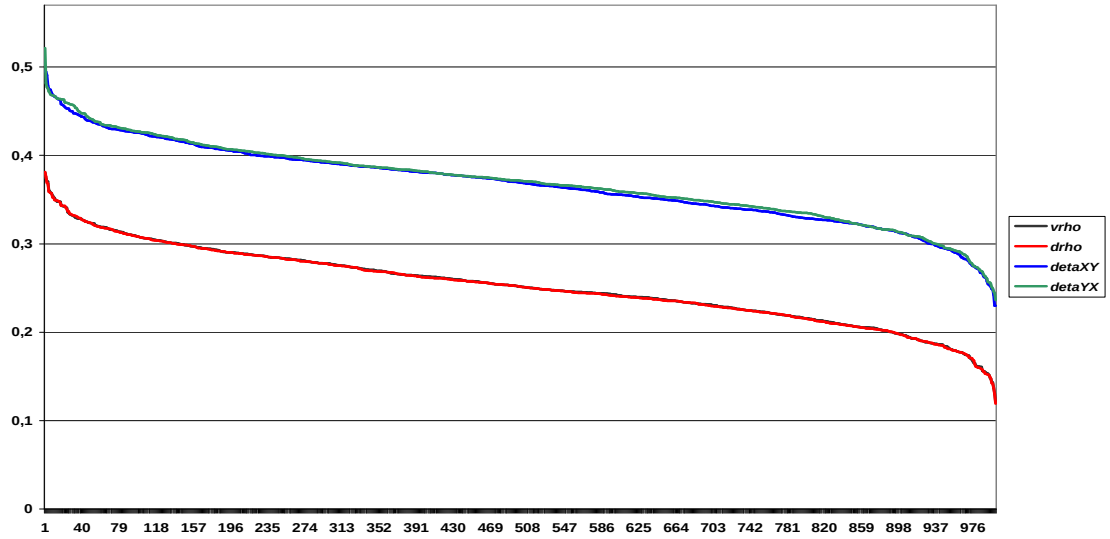
N=300 & m=100; deltah=0.1



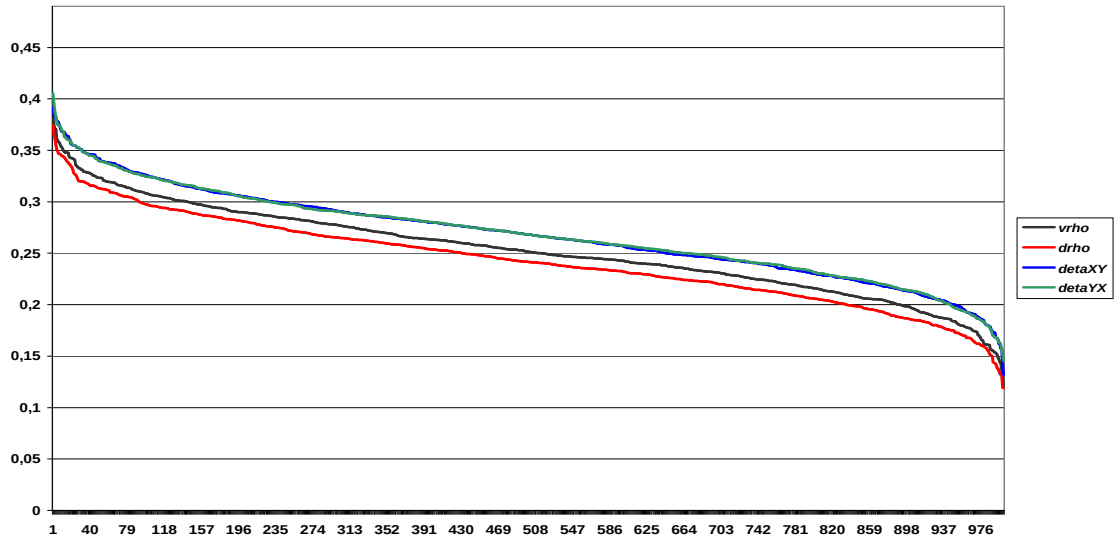
N=300 & m=100; deltah=0.5



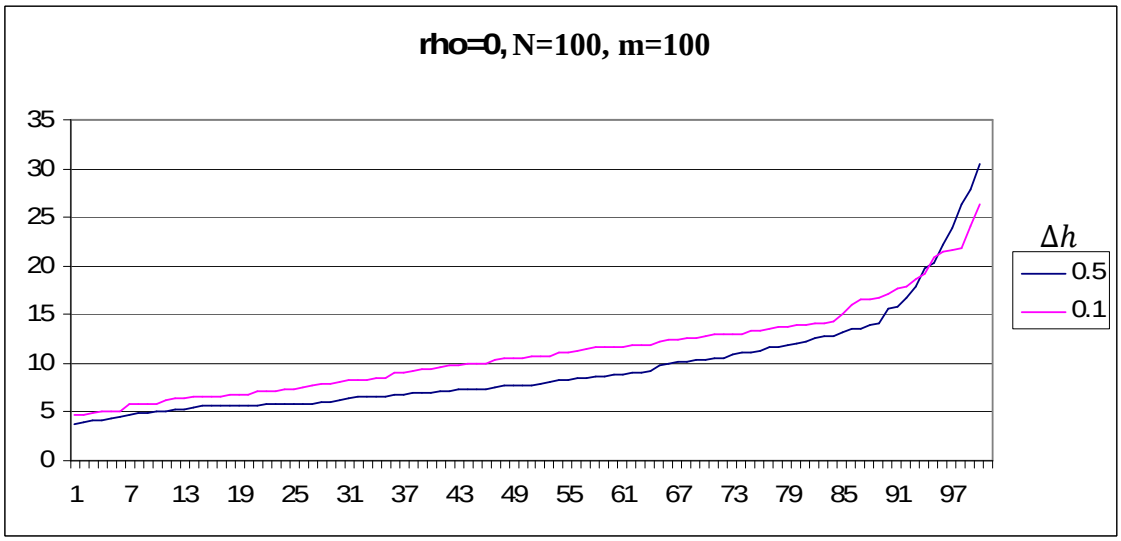
N=300 & m=1000; deltah=0.1

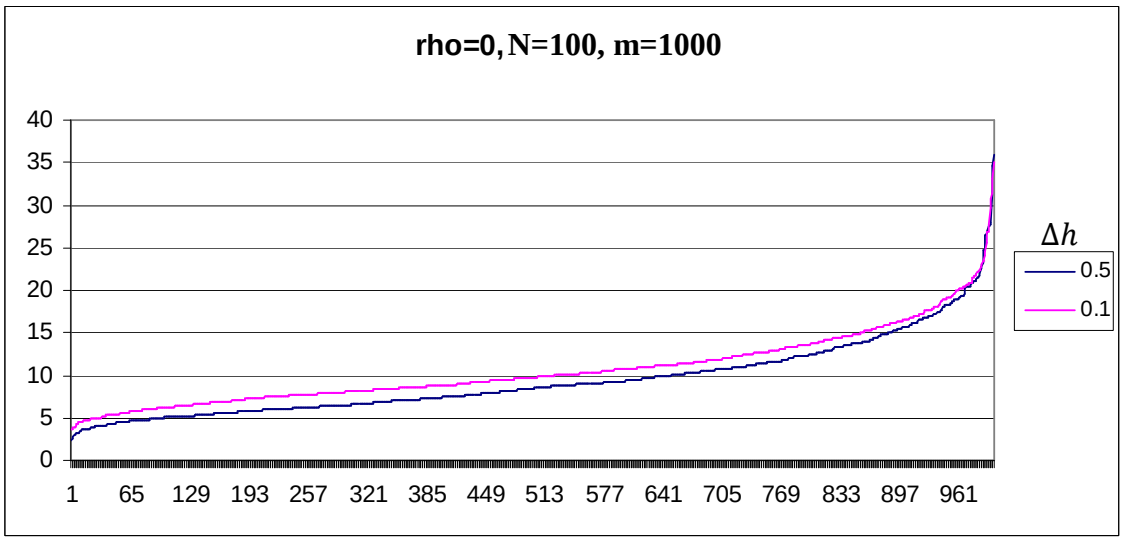


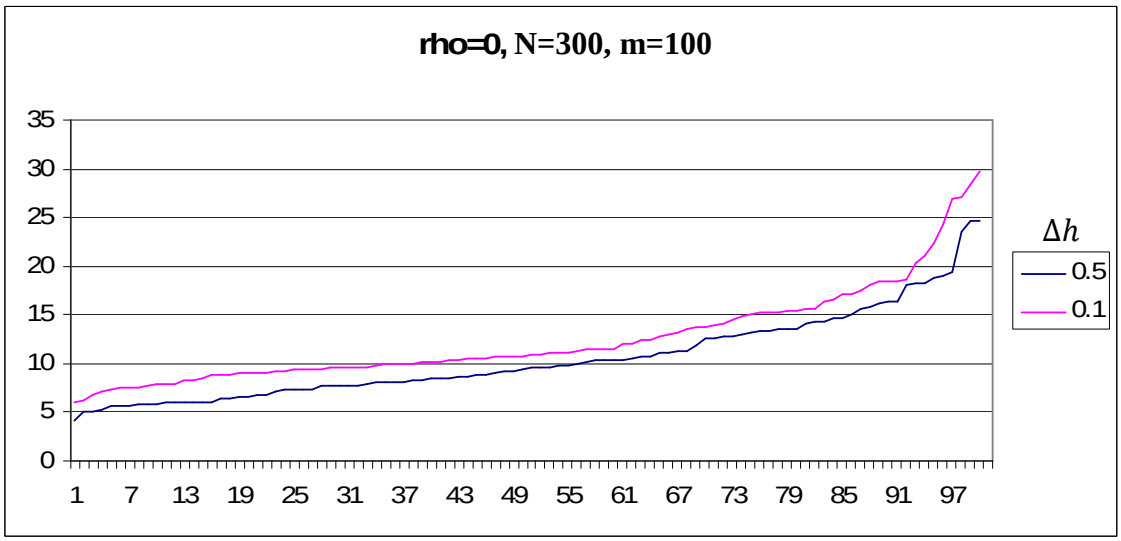
N=300 & m=1000; deltah=0.5

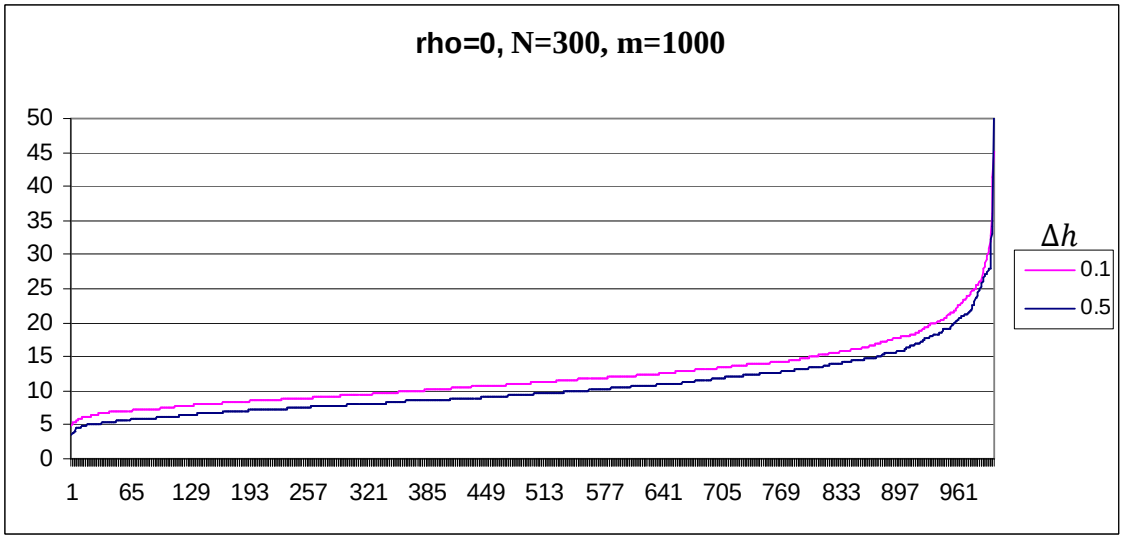


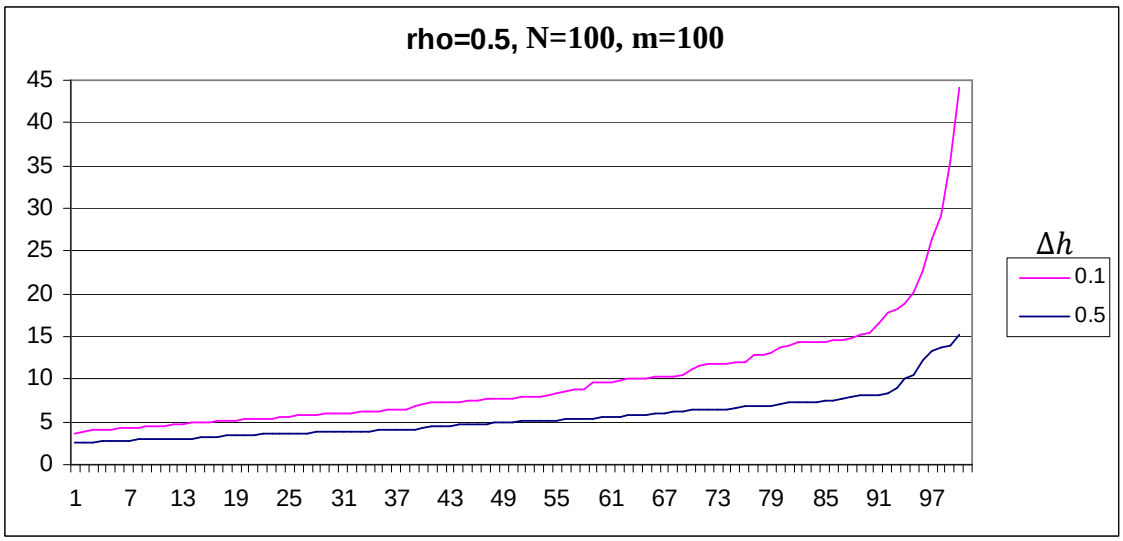
Ponto 3: Gráficos do quociente dos coeficientes de correlação $\left(\frac{\eta_{**}^2}{r^2}\right)$

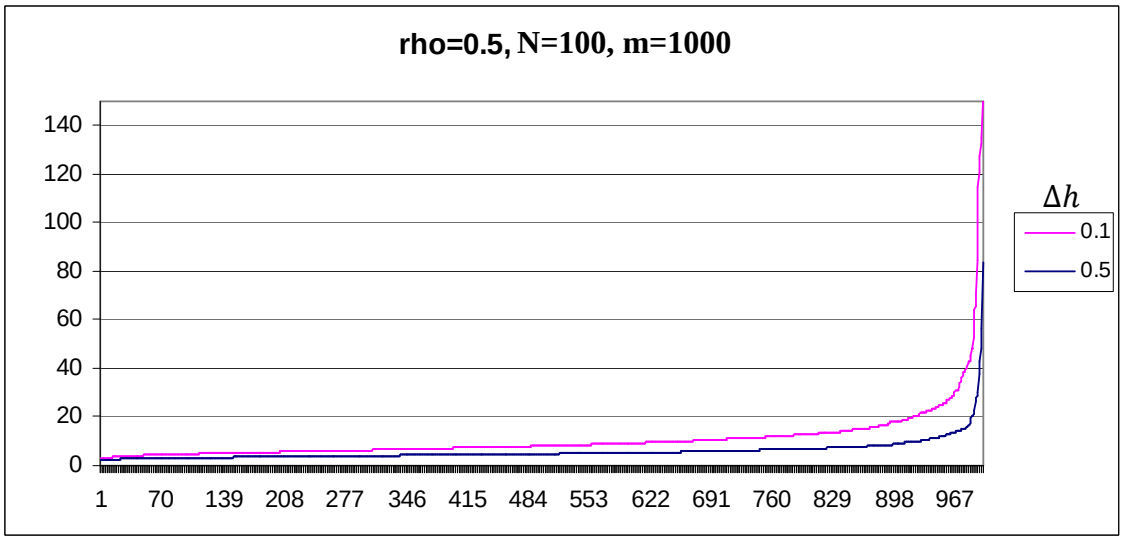


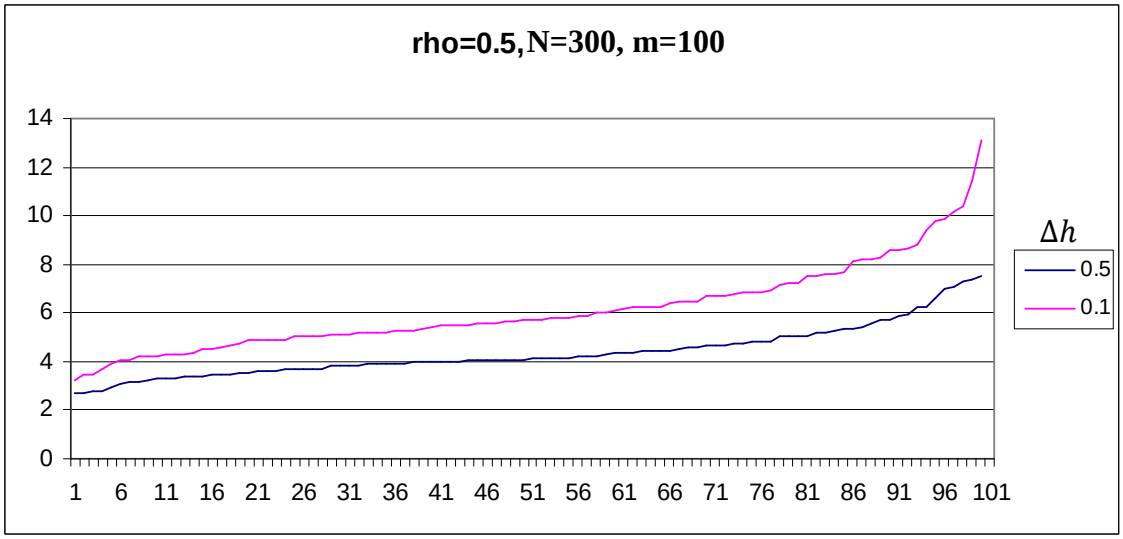


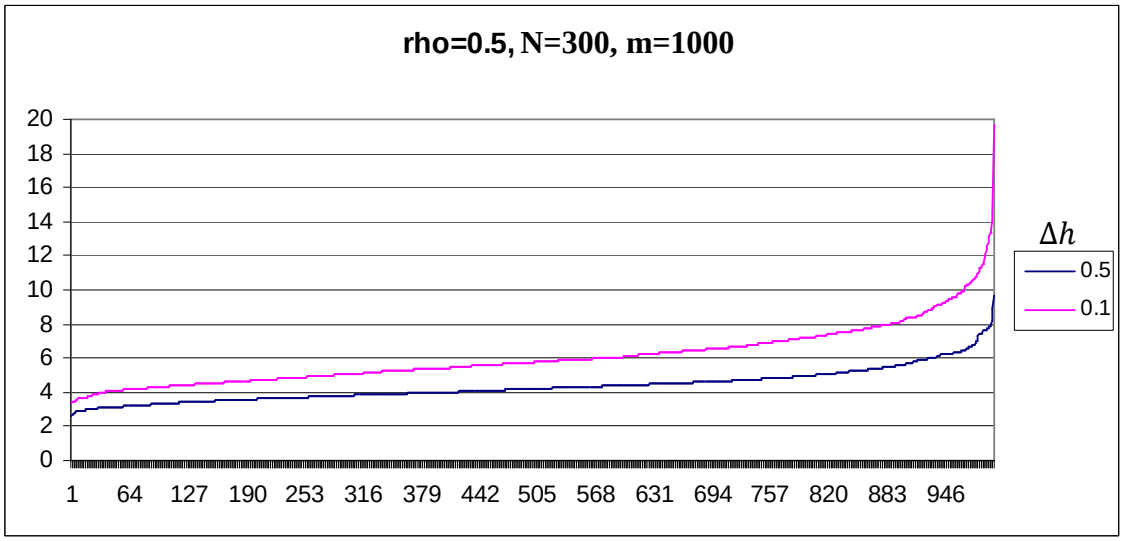



rho=0, N=300, m=1000
50
45
40
35
30
25
20
15
10
5
0
1
65
129
193
257
321
385
449
513
577
641
705
769
833
897
961
Δh
0.1
0.5


rho=0.5, N=100, m=100
45
40
35
30
25
20
15
10
5
0
1
7
13
19
25
31
37
43
49
55
61
67
73
79
85
91
97
Δh
0.1
0.5


rho=0.5, N=100, m=1000
140
120
100
80
60
40
20
0
1
70
139
208
277
346
415
484
553
622
691
760
829
898
967
Δh
0.1
0.5


rho=0.5,N=300, m=100
14
12
10
8
6
4
2
0
1 6 11 16 21 26 31 36 41 46 51 56 61 66 71 76 81 86 91 96 101
Δh
0.5
0.1


rho=0.5, N=300, m=1000
20
18
16
14
12
10
8
6
4
2
0
1 64 127 190 253 316 379 442 505 568 631 694 757 820 883 946
Δh
0.5
0.1