Correlação e Regressão

Análise de dados. Tópico 1
Prof. Dr. Ricardo Primi & Prof. Dr. Fabian Javier Marin Rueda

Adaptado de
Gregory J. Meyer, University of Toledo, USA; Apresentação na Universidade e São Francisco, São Paulo, Brasil
31 Julho, 2007

Correlação e Regressão

- Mensura a relação entre variáveis

  Correlação = co-**r**elação = co-variância = *r*

- Geralmente examina variáveis bidimensionais

- Mas diferenças de média entre grupos também podem ser expressas por meio da co-relação
  - e.g.,

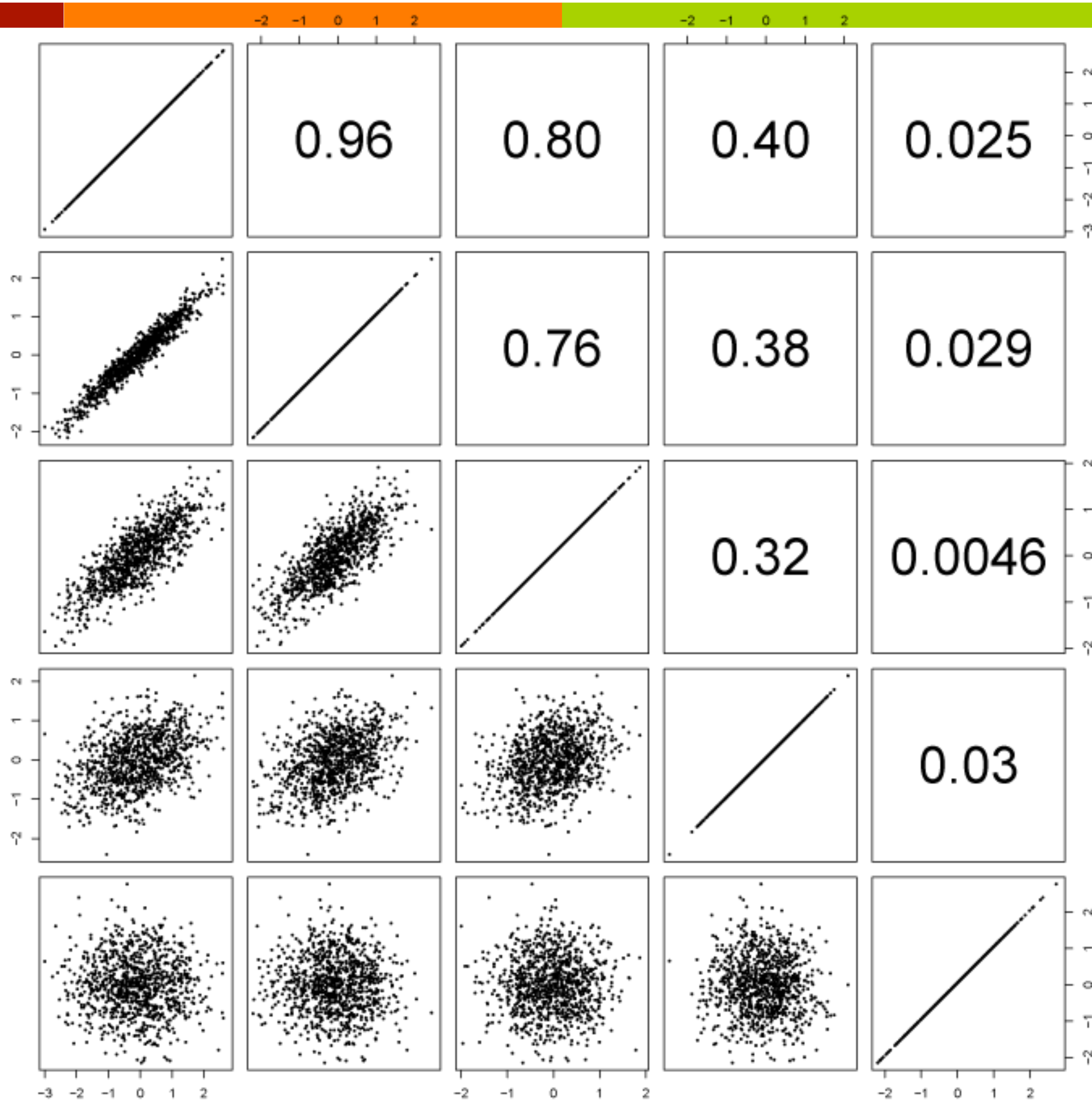
  VI = Diagnóstico: Transtorno Psicótico (codificado como 1) vs. outros transtornos (codificados como 0)

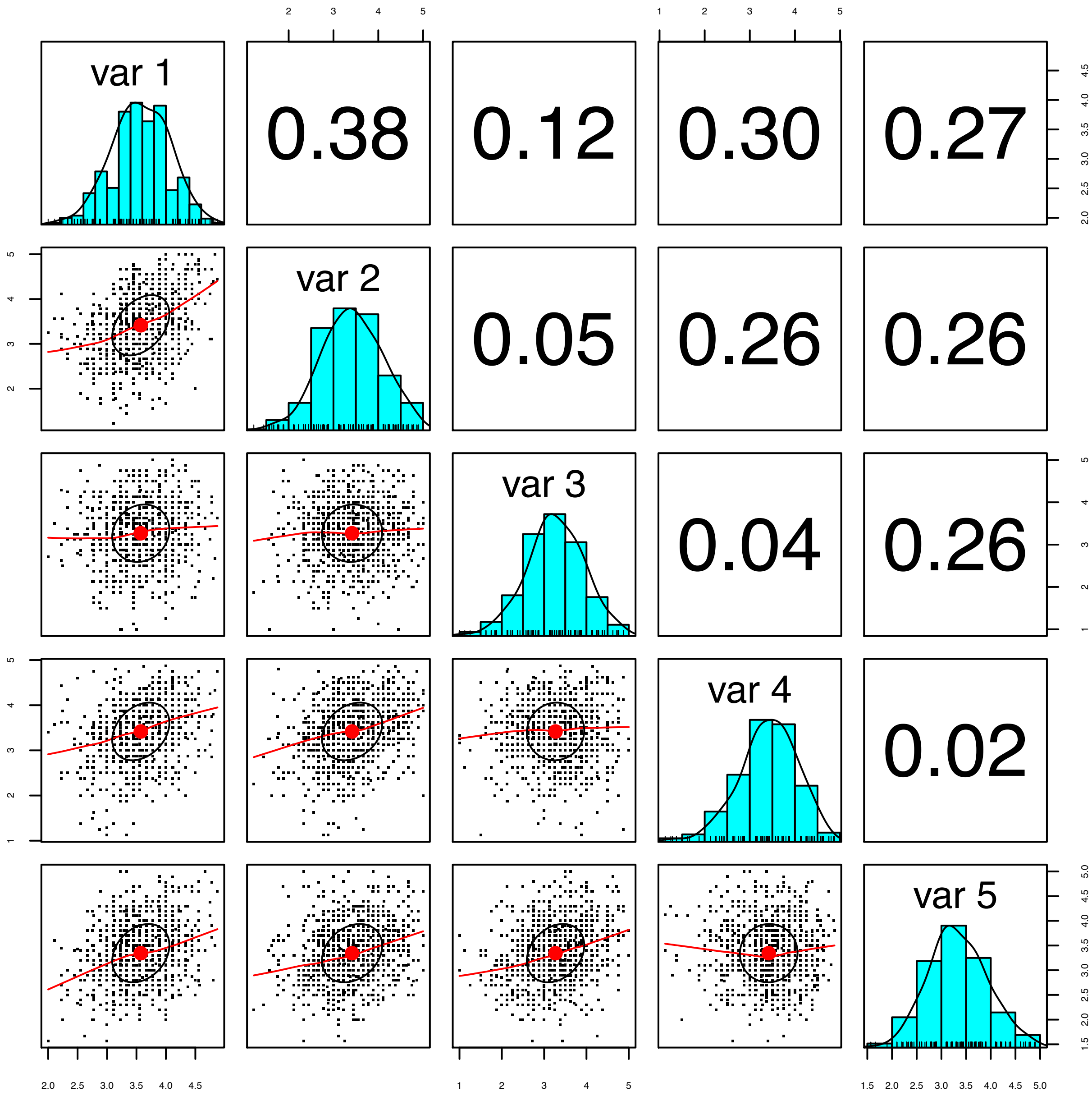
  VD = X-% como um índice de Acurácia Perceptual

  - *t*-test comparando $M_{\text{Psychotic}}$ vs. M_{Other} de X-% pode ser expressa como a *r do* Diagnóstico com X-%
    - i.e., $r_{\text{Diagnostico -X-\%}}$
  - *r* e *t* terão o mesmo valor de *p* ou de significância estatística

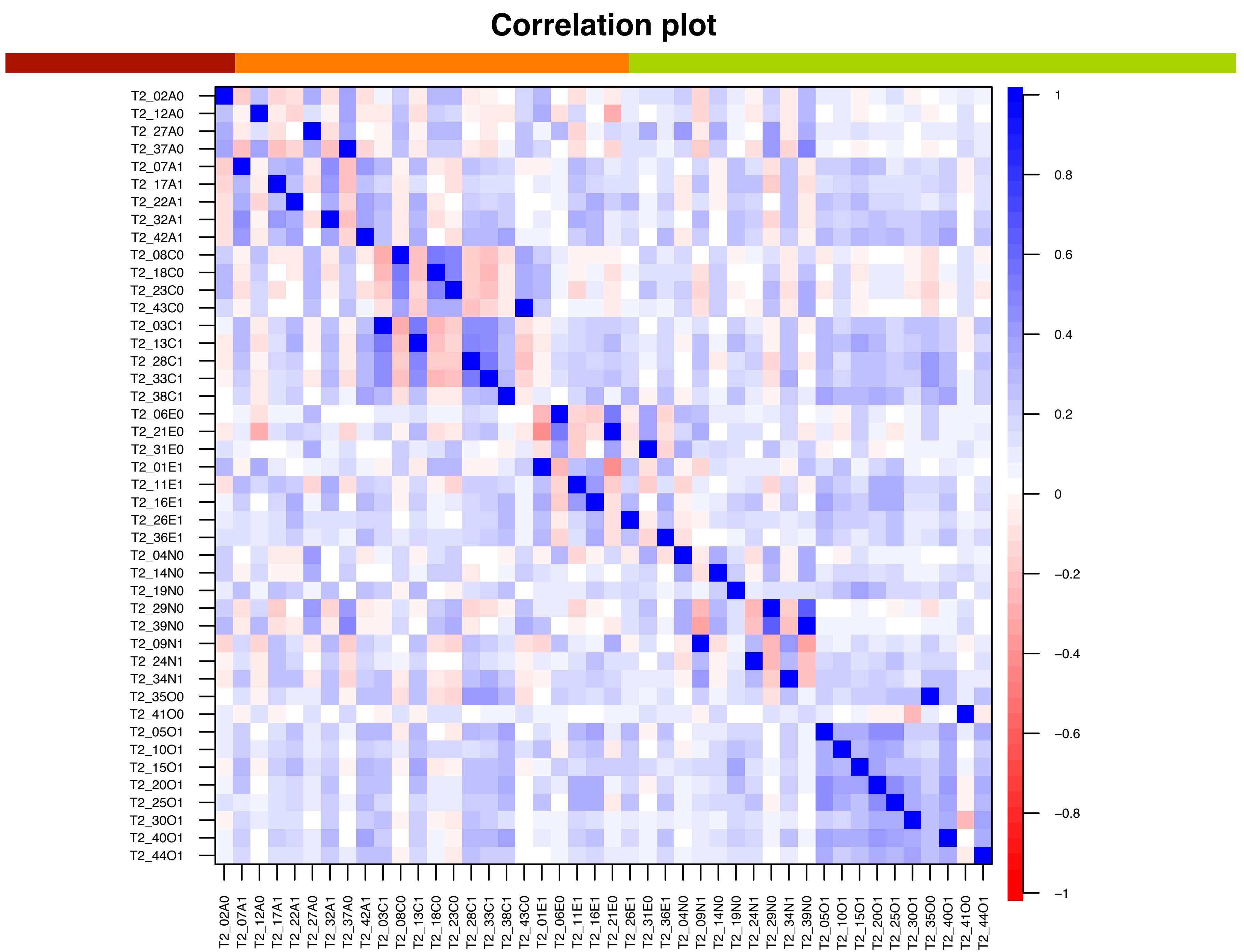
Correlação e Regressão

- Em geral as medidas estão associadas por relações lineares
  - Mas existem técnicas para correlações e regressões não lineares
- Correlação ≠ Causalidade
- *r's* assumem valores entre -1.0 e +1.0
  - O sinal mostra a direção das relações
  - Os valores absolutos mostram a magnitude da relação
    - 0.0 = ausência de relação
    - -1.0 or +1.0 = relação perfeita
- Visualizando as magnitudes das correlações
  - "ForcedDegreeofCorrelation.sps"

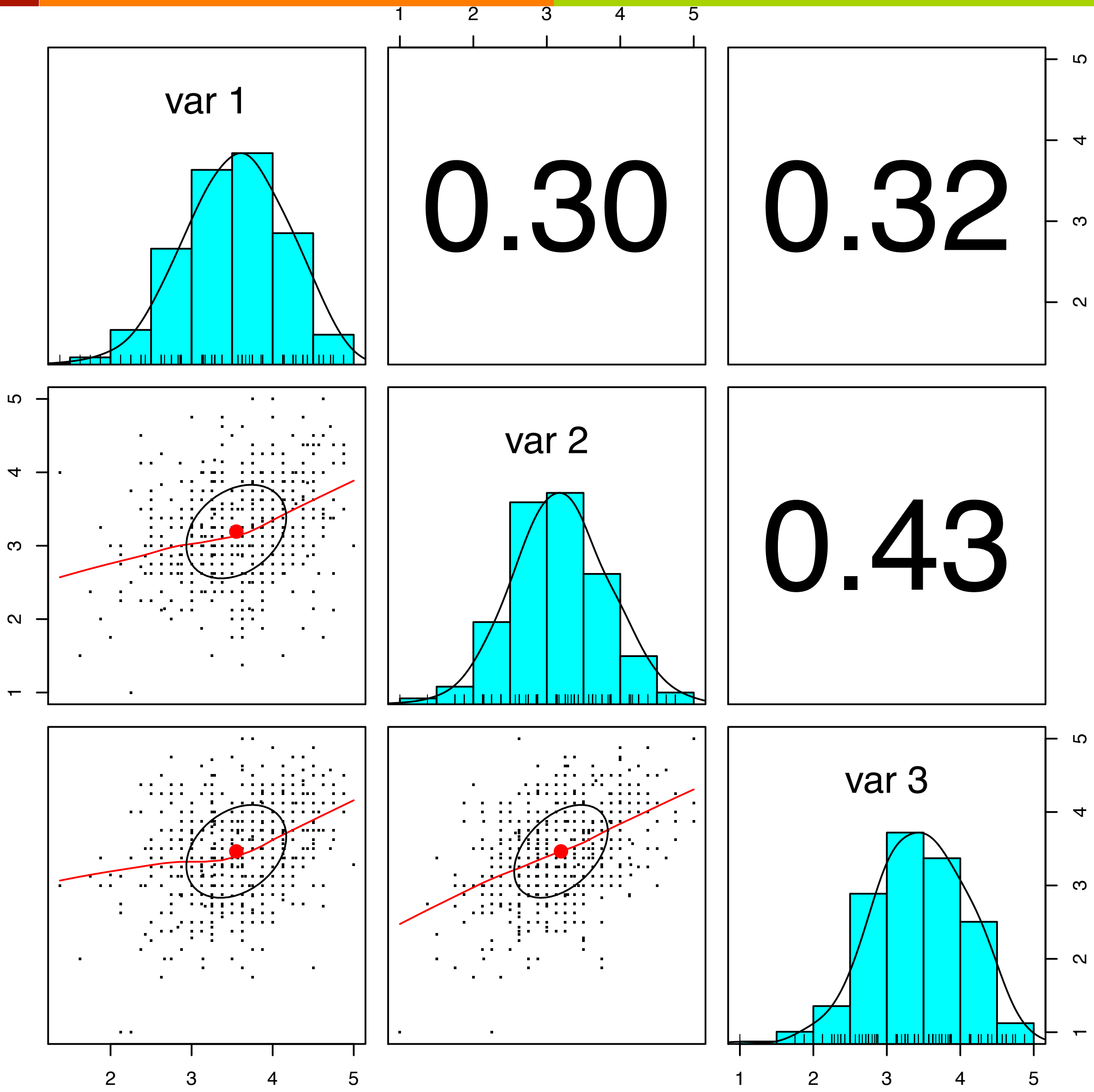

0.96
0.80
0.40
0.025
0.76
0.38
0.029
0.32
0.0046
0.03



var 1
var 2
var 3
var 4
var 5
0.38
0.12
0.30
0.27
0.05
0.26
0.26
0.04
0.26
0.02

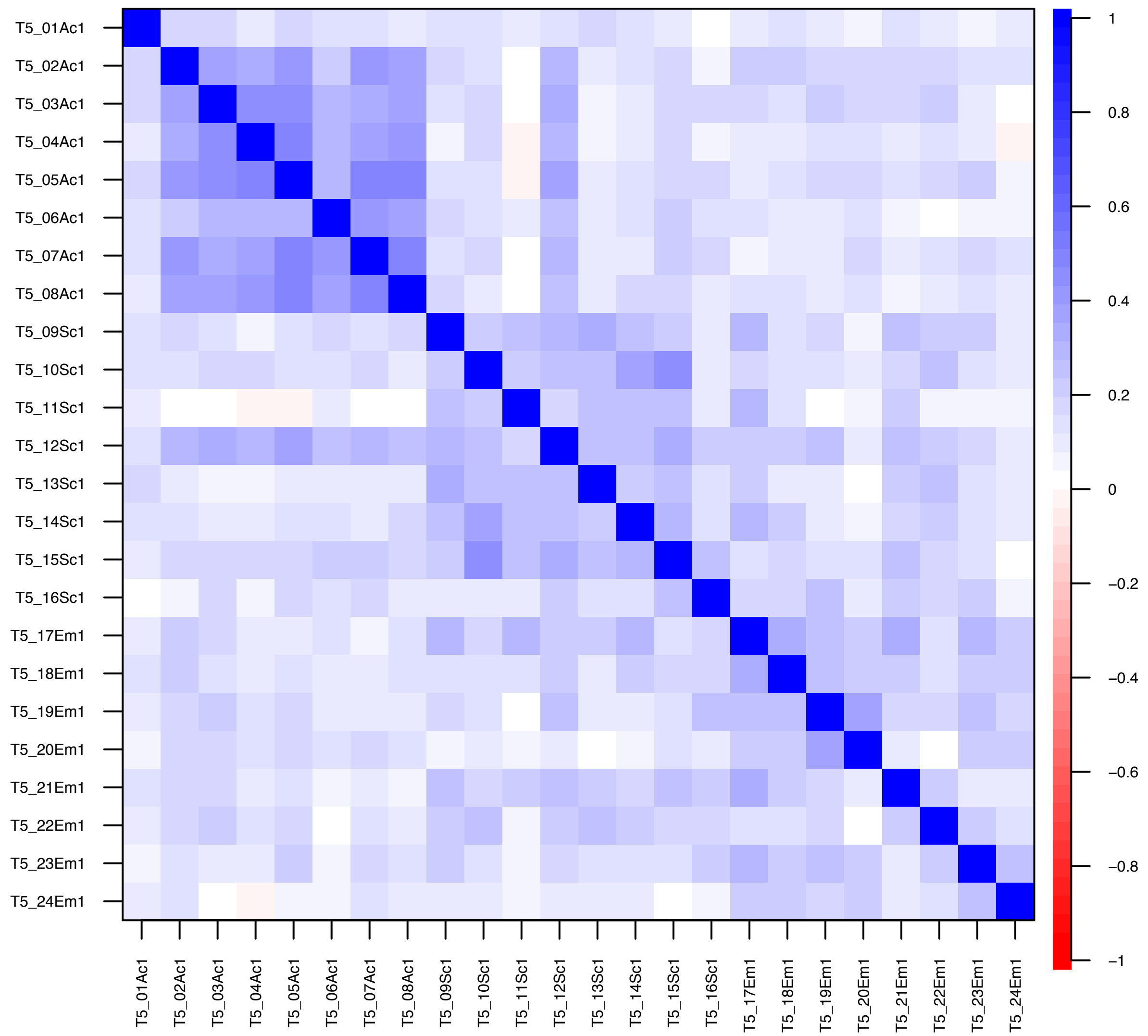


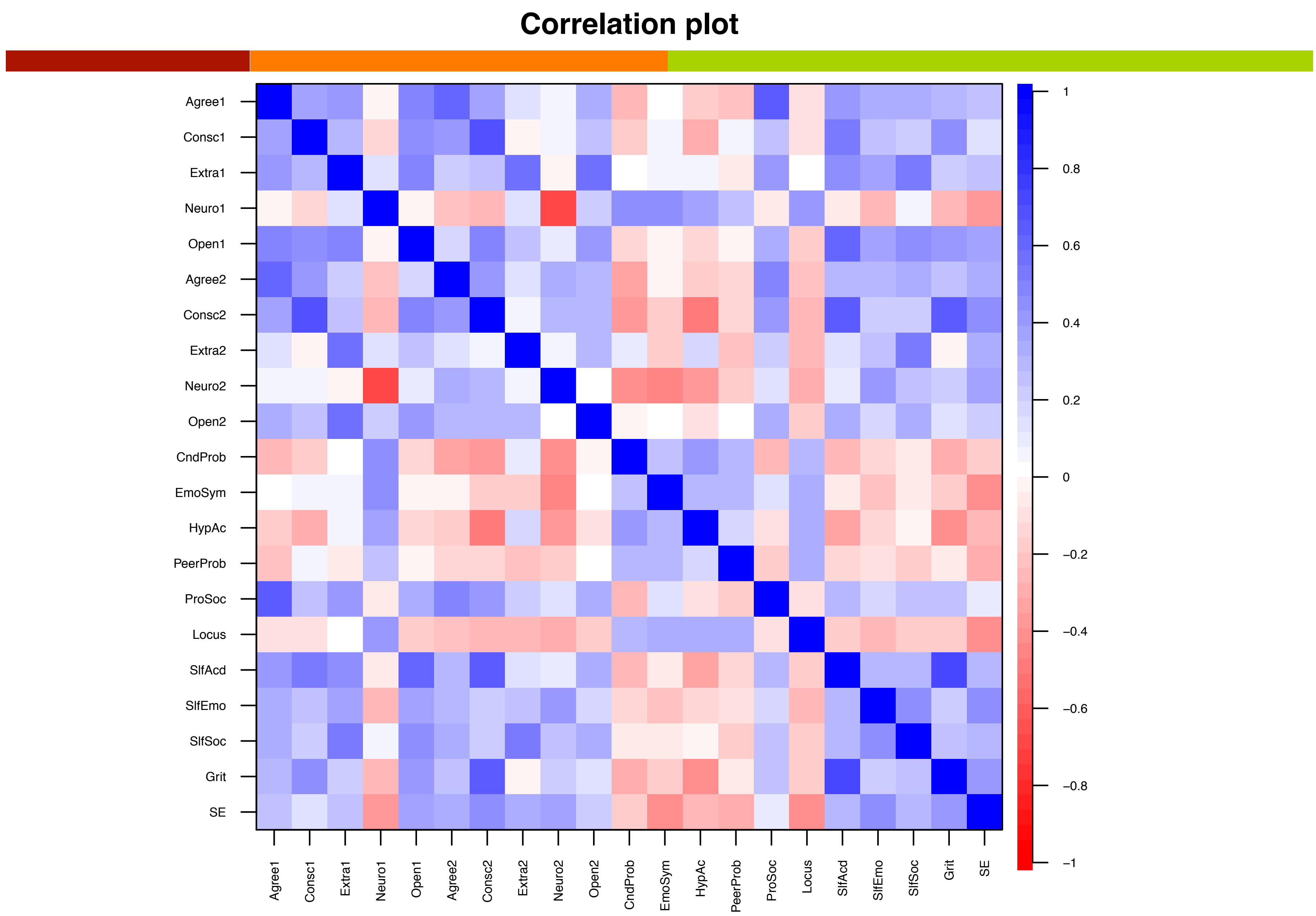
Correlation plot
T2_02A0
T2_12A0
T2_27A0
T2_37A0
T2_07A1
T2_17A1
T2_22A1
T2_32A1
T2_42A1
T2_08C0
T2_18C0
T2_23C0
T2_43C0
T2_03C1
T2_13C1
T2_28C1
T2_33C1
T2_38C1
T2_06E0
T2_21E0
T2_31E0
T2_01E1
T2_11E1
T2_16E1
T2_26E1
T2_36E1
T2_04N0
T2_14N0
T2_19N0
T2_29N0
T2_39N0
T2_09N1
T2_24N1
T2_34N1
T2_35O0
T2_41O0
T2_05O1
T2_10O1
T2_15O1
T2_20O1
T2_25O1
T2_30O1
T2_40O1
T2_44O1
T2_02A0
T2_07A1
T2_12A0
T2_17A1
T2_22A1
T2_27A0
T2_32A1
T2_37A0
T2_42A1
T2_03C1
T2_08C0
T2_13C1
T2_18C0
T2_23C0
T2_28C1
T2_33C1
T2_38C1
T2_43C0
T2_01E1
T2_06E0
T2_11E1
T2_16E1
T2_21E0
T2_26E1
T2_31E0
T2_36E1
T2_04N0
T2_09N1
T2_14N0
T2_19N0
T2_24N1
T2_29N0
T2_34N1
T2_39N0
T2_05O1
T2_10O1
T2_15O1
T2_20O1
T2_25O1
T2_30O1
T2_35O0
T2_40O1
T2_41O0
T2_44O1
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

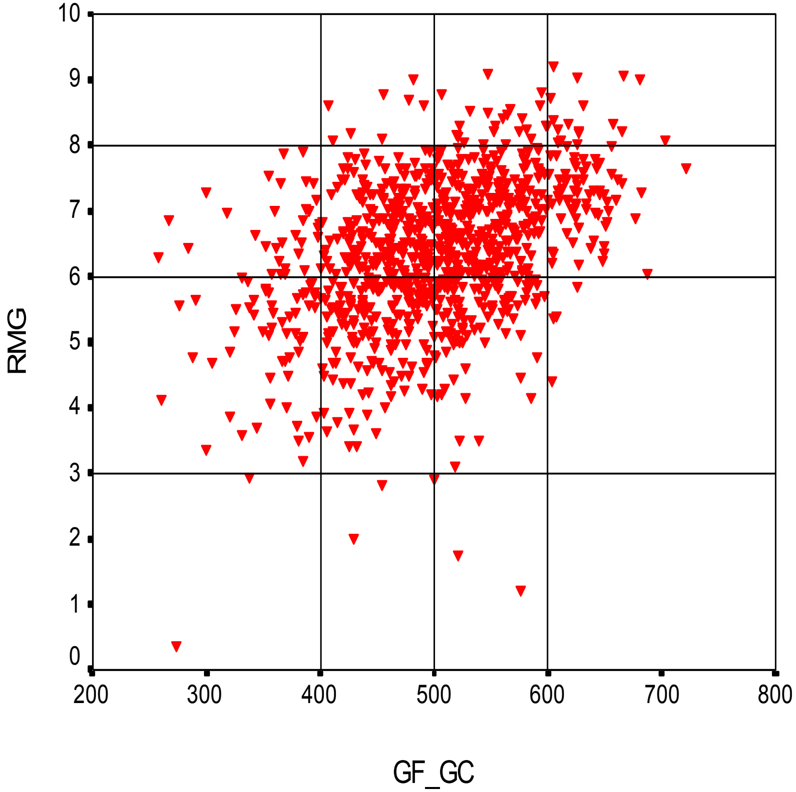


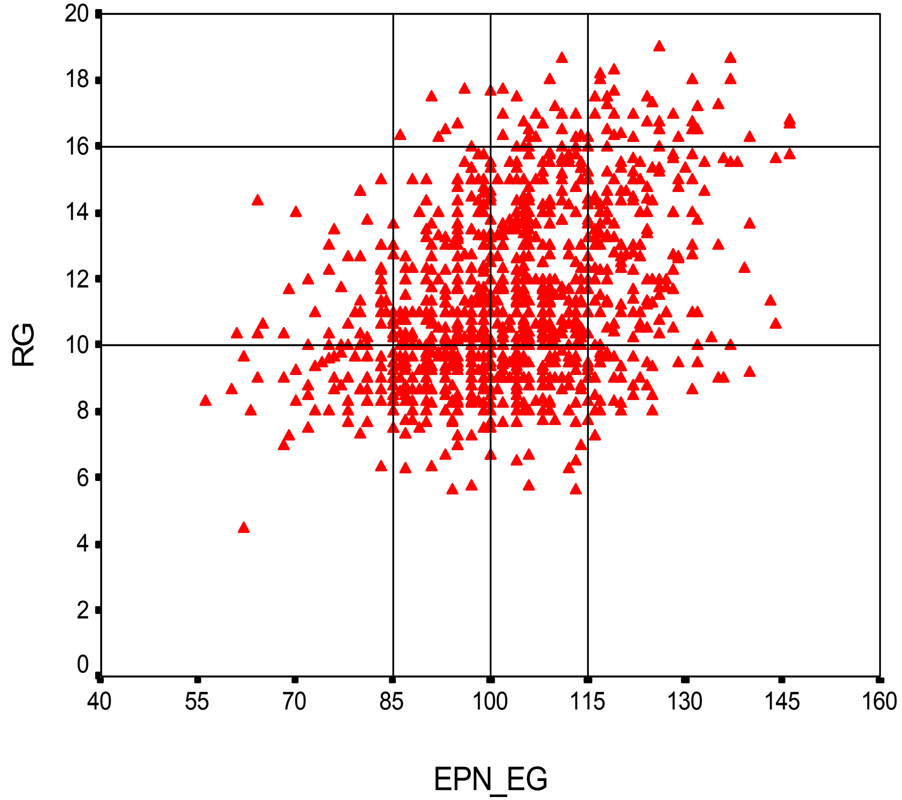
var 1
0.30
0.32
var 2
0.43
var 3
1
2
3
4
5

Correlation plot




Correlation plot
Agree1
Consc1
Extra1
Neuro1
Open1
Agree2
Consc2
Extra2
Neuro2
Open2
CndProb
EmoSym
HypAc
PeerProb
ProSoc
Locus
SlfAcd
SlfEmo
SlfSoc
Grit
SE
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1


RMG
GF_GC
0
1
2
3
4
5
6
7
8
9
10
200
300
400
500
600
700
800


RG
EPN_EG
0
2
4
6
8
10
12
14
16
18
20
40
55
70
85
100
115
130
145
160

Correlação e Regressão

- Distinção estatística
  - Correlação
    - 2 variáveis randômicas, *X* e *Y*
      - Nenhuma delas está sob controle do experimentador
  - Regressão:
    - 1 variável fixa (*X*), 1 variável randômica (*Y*)
      - se presume *X* que está sobre controle do experimentador e somente os valores que interessam são estudados
- Essa distinção é precisa tecnicamente mas não necessária no utilização prática das correlações e regressões

Fórmula

$$r = \frac{\sum_{i=1}^{N}\left(\frac{\left(x_i - \bar{x}\right)}{s_x}\right)\left(\frac{\left(y_i - \bar{y}\right)}{s_y}\right)}{\left(N-1\right)}$$

$$r = \frac{\sum_{i=1}^{N} z_{x_i} z_{y_i}}{N-1}$$

Produto-momento!

- A média do produto de dois momentos indicando co-relação
- Produto : multiplicação de duas variáveis (X, Y)
- Momento: função aplicada a média de desvios
- Momentos centrais: : 1o = Média, 2º = Variancia, 3º = Assimetria, 4o = Kurtose
- Os escores z são momentos
- Desvios da média em unidades de desvio padrão
- Co-relação: ocorrência simultânea together
- z para X pareado copm z para Y
- Então a correlação Produto-Momento de Pearson (r) é a magnitude média em que pares de escores (X, Y) se correlacionam por desviarem simultaneamente de suas respectivas médias

$$z = \frac{(X - \bar{X})}{\sigma}$$

$$r = \frac{\sum z_X z_Y}{N}$$

Coeficiente de Correlação de Pearson - Produto Momento

n	Nota A	$x = X - \overline{X}$	$\lvert x \rvert$	x^2	$z = \frac{(X - \overline{X})}{S}$	Nota B	$y = Y - \overline{Y}$	$\lvert y \rvert$	y^2	$z = \frac{(Y - \overline{Y})}{S}$	z^2
1	2					3					
2	2					2					
3	4					5					
:	:					:					
9	9					7					
10	10					7					
11	9					6					
12	9					8					

$$r = \frac{\sum z_x z_y}{N - 1}$$

$$z = \frac{(X - \overline{X})}{S} \qquad z = \frac{(Y - \overline{Y})}{S}$$

$$x = X - \overline{X} \qquad y = Y - \overline{Y}$$

$$S = \sqrt{\frac{\sum (X - \overline{X})^2}{N - 1}}$$

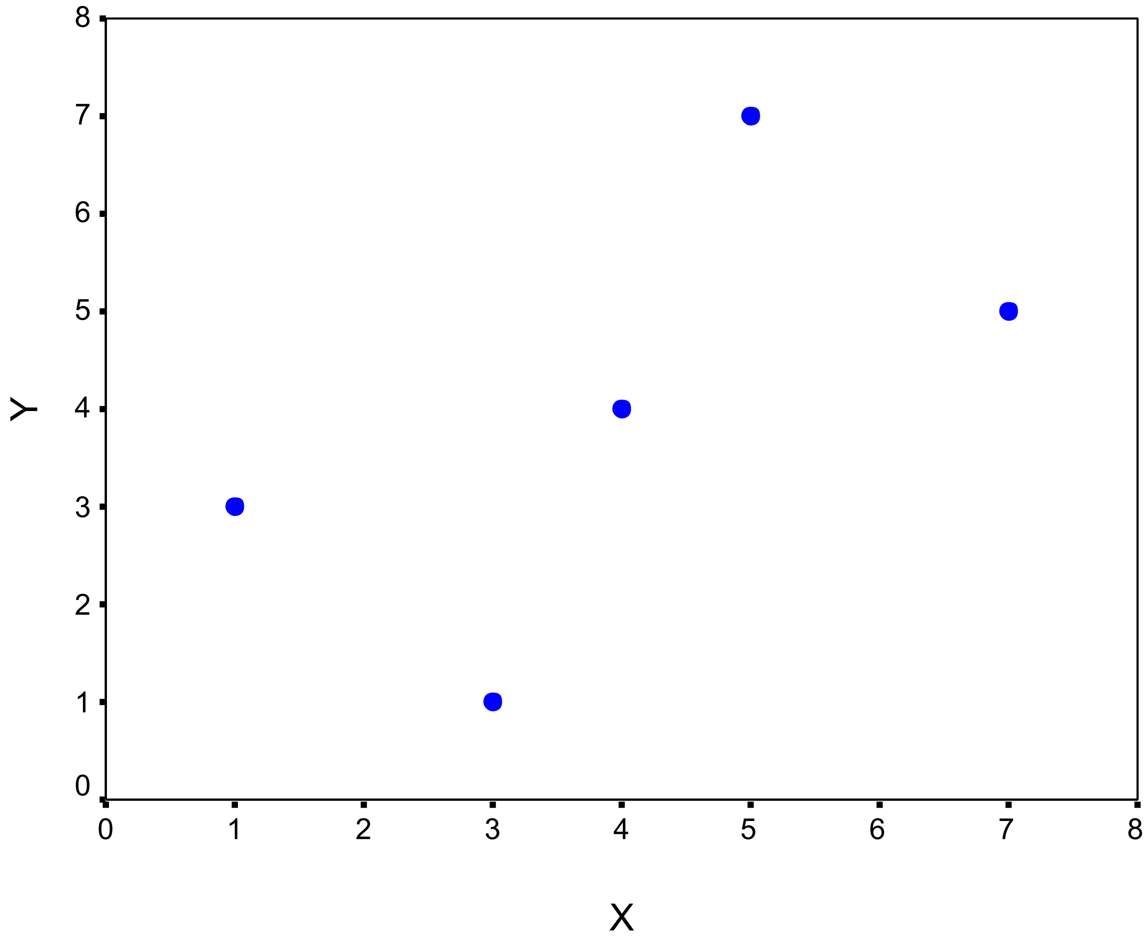
$$x^2 = (X - \overline{X})(X - \overline{X})$$

Correlação e Regressão

- Distinção Prática
  - Correlação:
    - Magnitude de associação entre *X* e *Y*
      - Ambas podem ser desenhadas no eixo horizontal (abscissa) porque ambas poderiam ser designadas como *X*
  - Regressão:
    - Usa a associação para desenvolver uma equação que prevê *Y* a partir de *X* a partir da melhor forma possível.
      - "Regride *Y* em *X*"
      - Desenha *X* no eixo horizontal; *Y* no eixo vertical (ordenada)

r Produto Momento: Exemplo

ID	X	Y
1	5	7
2	1	3
3	4	4
4	3	1
5	7	5
M	4.00	4.00
σ	2.00	2.00
Sum		



- Insira *X* and*Y no* SPSS e calcule o *r*
  - r_{XY} = .60

Reta de regressão

Equação preditora:

$\hat{Y} = b_0 + b_1X$

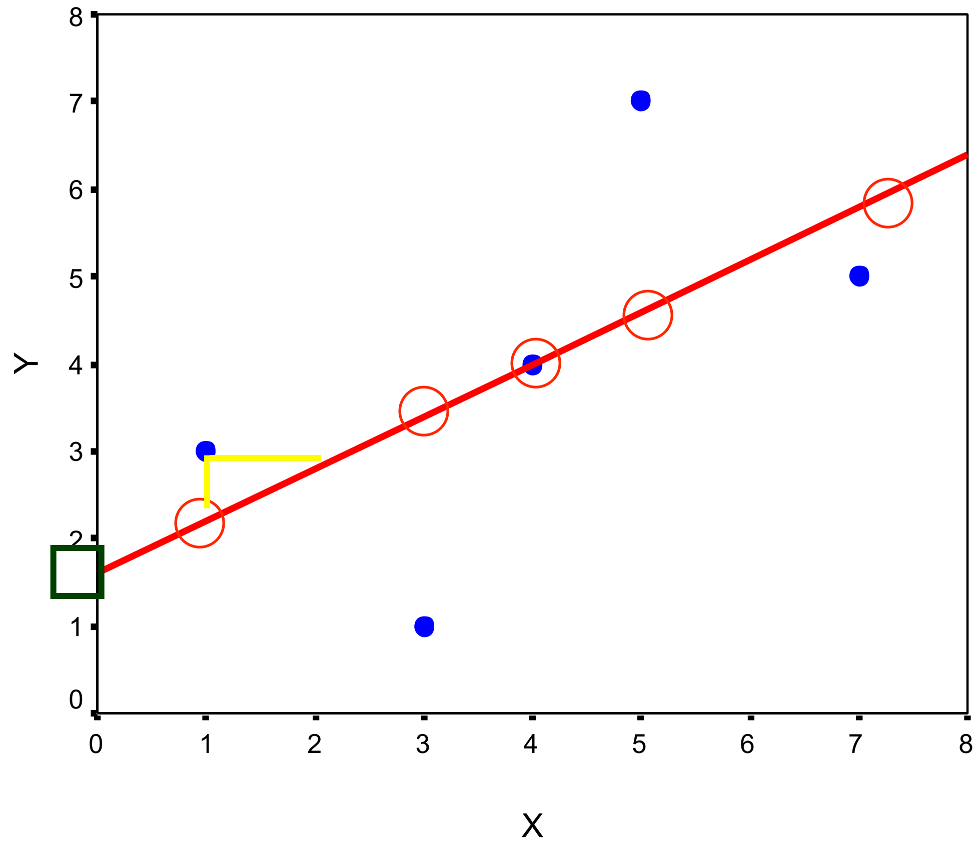
X = valor do preditor

Y = valor do critério

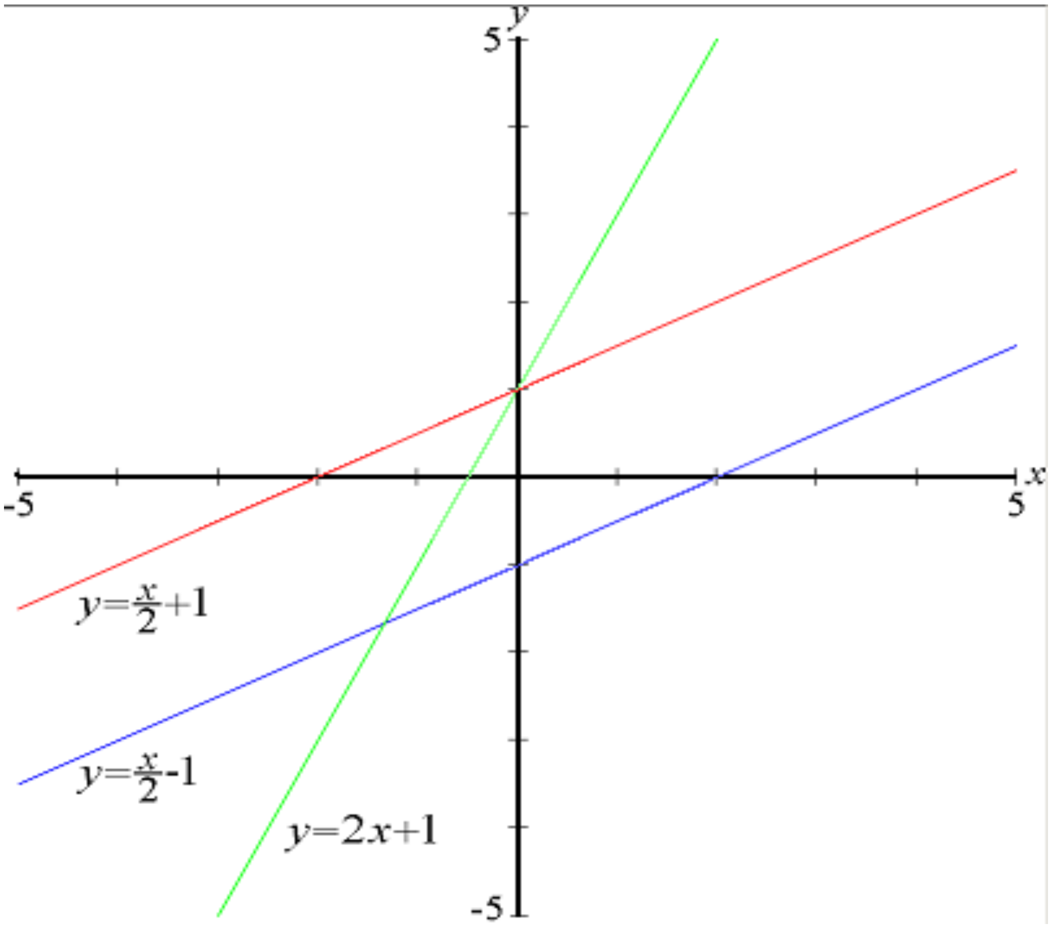
$\hat{Y}$= valor predito

b_1= inclinação da linha

b_0= constante



Equação da reta



Reta de regressão

- Melhor previsão de *Y* em relação aos valores de *X*
- Equação de previsão:

$$\hat{Y} = b_0 + b_1X$$

Na qual:

X = valor do preditor (variável preditora ou VI)

$\hat{Y}$ = valor previsto de *Y* (variável resposta ou VD ou critério)

i.e., valor de *Y na linha*, dado *X*

b_1 = inclinação (*slope*) da linha, Mudança em $\hat{Y}$ para uma mudança de 1-unidade de mudança em *X*

$b_1 = r_{XY}(S_Y/S_X)$

b_0 = constante (*intercept*)

$\hat{Y}$ quando $X = 0.0$

$b_0 = M_Y - b_1M_X$

Reta de regressão

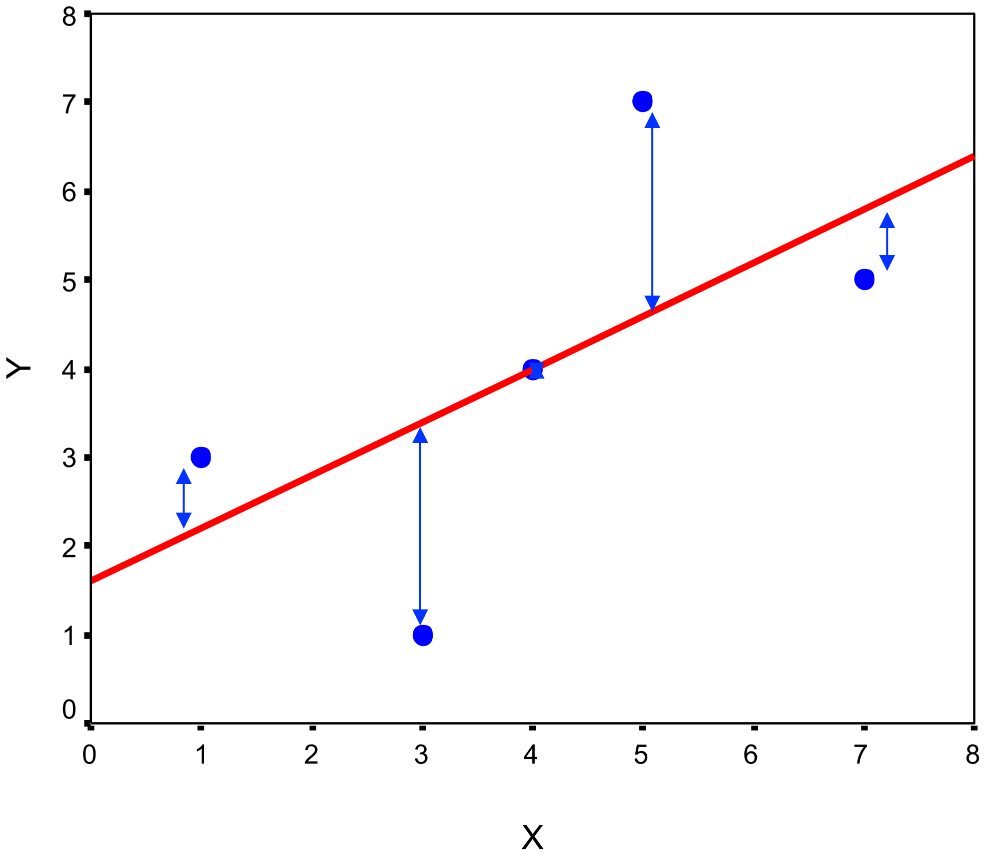
- Gere a equação de regressão do exemplo no SPSS
- A equação de previsão minimiza os erros, definidos como

$$SS_{Residual} = \sum(Y - \hat{Y})^2$$

- Em que :
  - Y = valores observados (i.e., os valores do gráfico de dispersão)
  - $\hat{Y}$= valores preditos na reta de regressão
- Portanto, $SS_{Residual}$ indica a extensão em que a linha não consegue prever os dados observados
  - SS_{Resid} na regressão é análogo ao SS_{WG} na ANOVA
    - i.e., variabilidade nas células que não pode ser explicada pelo modelo.

Reta de regressão: Exemplo

- $SS_{Resid} = \sum(Y - \hat{Y})^2$
  - Sum of squared distances of each person's Y score from the line of prediction



Reta de regressão : Exemplo

Equação de previsão: $\hat{Y} = b_0 + b_1 X$

$b_1 = r_{XY}(S_Y/S_X)$

$= .60(2.236/2.236) = .60$

(Porque $S_Y = S_X$, $r = b_1$; isso é raro acontecer)

$b_0 = M_Y - b_1 M_X$

$= 4 - .60(4) = 4 - 2.4 = 1.6$

$\hat{Y} = 1.6 + (.60)(X)$

Calcule $\hat{Y}$ para cada X

e.g., for $X = 5$

$\hat{Y} = 1.6 + (.60)(5) = 4.6$

ID	X	Y	$\hat{Y}$
1	5	7	
2	1	3	
3	4	4	
4	3	1	
5	7	5	
M	4.0	4.0	
S	2.236	2.236	

Reta de regressão : Exemplo

Equação de previsão: $\hat{Y} = b_0 + b_1 X$

$b_1 = r_{XY}(S_Y/S_X)$

$= .60(2.236/2.236) = .60$

(Porque $S_Y = S_X$, $r = b_1$; isso é raro acontecer)

$b_0 = M_Y - b_1 M_X$

$= 4 - .60(4) = 4 - 2.4 = 1.6$

$\hat{Y} = 1.6 + (.60)(X)$

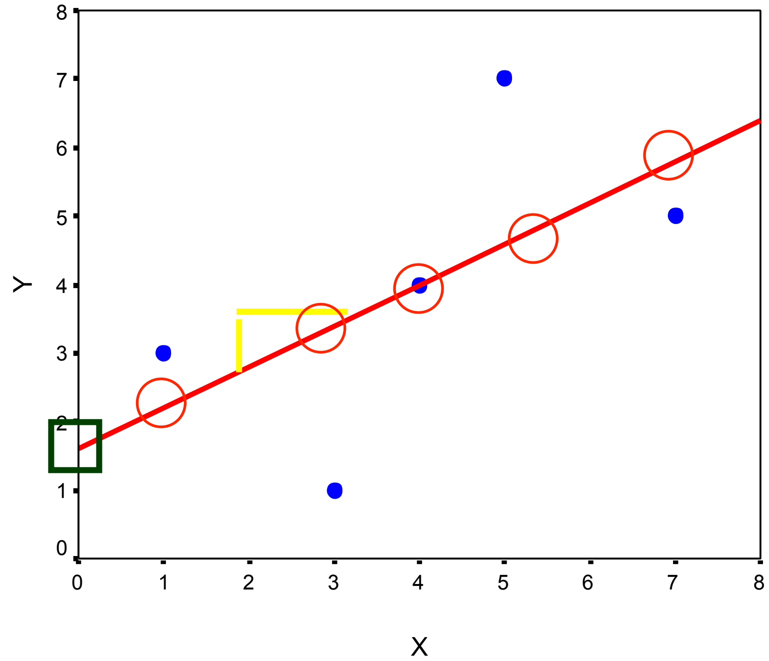
Calcule $\hat{Y}$ para cada X

e.g., for $X = 5$

$\hat{Y} = 1.6 + (.60)(5) = 4.6$

ID	***X***	***Y***	$\hat{Y}$
1	5	7	4.60
2	1	3	2.20
3	4	4	4.00
4	3	1	3.40
5	7	5	5.80
M	4.0	4.0	
S	2.236	2.236	

Reta de regressão exemplo



$\hat{Y} = b_0 + b_1 X = 1.6 + (.60)(X)$

b_1 = .60; mudança no escore bruto de $\hat{Y}$ *para* 1a unidade de mudança em X

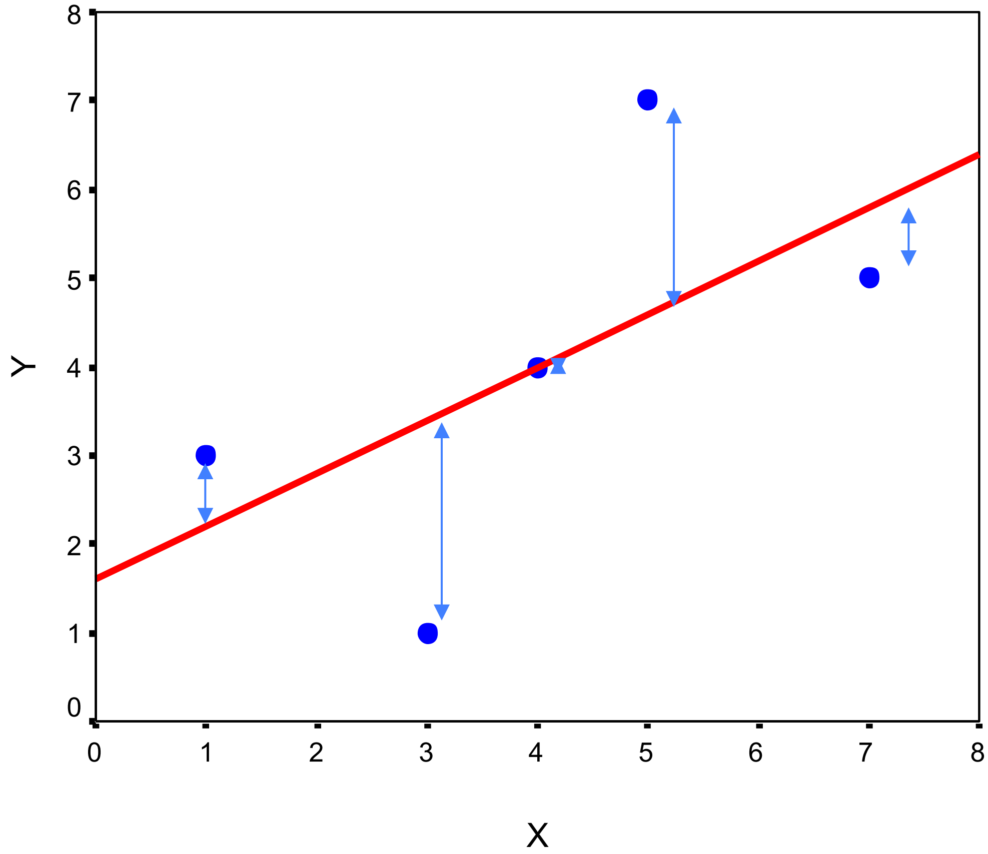
b_0 = 1.6; constante; valor de $\hat{Y}$ quando X = 0

- Como na ANOVA, a análise de regressão particiona a variância da variável dependente em componentes mutuamente exclusivos e exaustivos:
  - 1. Aquilo que é explicado pelo modelo
    - i.e., pela VI ou VIs
  - 2. Aquilo que não pode ser explicado pelo modelo
    - i.e., variância residual

- $SS_{Total} = SS_{Modelo} + SS_{Residual}$
  - Como definido, $SS_{Residual} = \sum(Y - \hat{Y})^2$
    - Soma das distâncias ao quadrado de cada escore Y das pessoas em relação a reta de regressão (linha de predição)

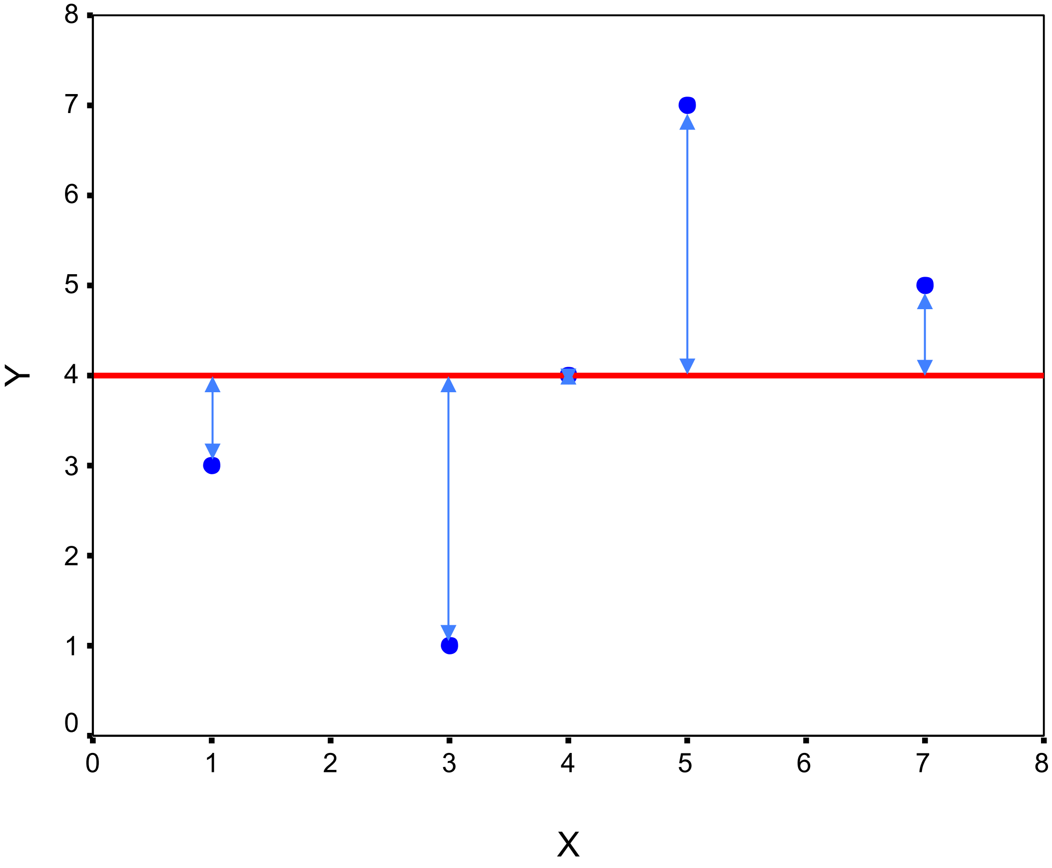
Soma de Quadrados da Regressão

- $SS_{Residual} = \sum(Y - \hat{Y})^2$
  - Soma das distâncias ao quadrado de cada escore Y das pessoas em relação a reta de previsão



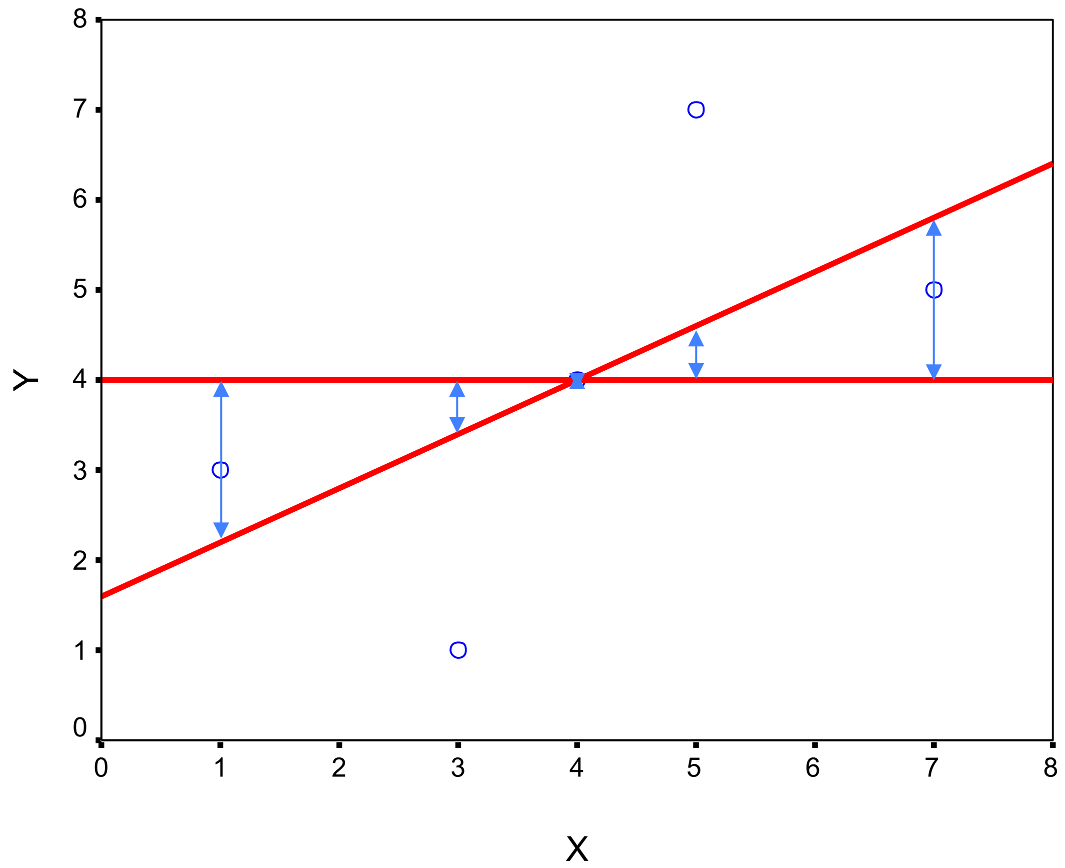
Soma de Quadrados da Regressão

- $SS_{Total} = \sum(Y - M_Y)^2$
  - Soma das distâncias ao quadrado de cada escore Y das pessoas em relação à média geral de *Y*
  - i.e., numerador da variância de Y; total de variância na VD



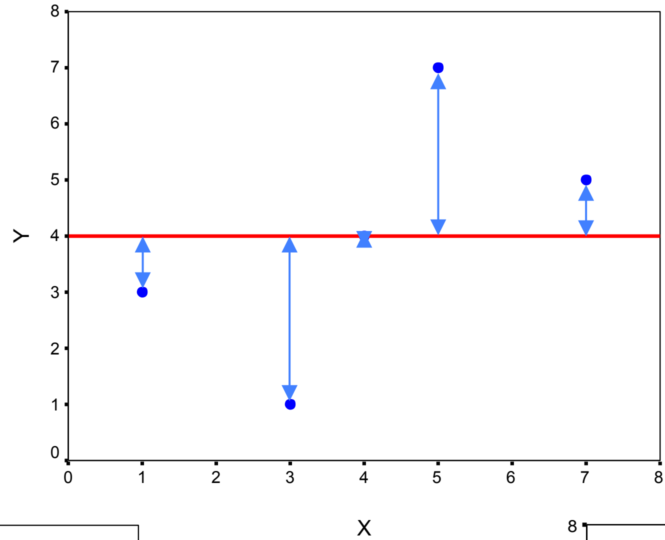
Soma de Quadrados da Regressão

- $SS_{Modelo} = \sum(\hat{Y} - M_Y)^2$
  - Soma das distâncias ao quadrado de cada escore predito Y (i.e., a linha) da média de *Y*
  - Indica a variação na VD que pode ser explicada pelo modelo
  - Os pontos observados de dados não são considerados; somente a comparação do modelo de Y tem relação à média de Y



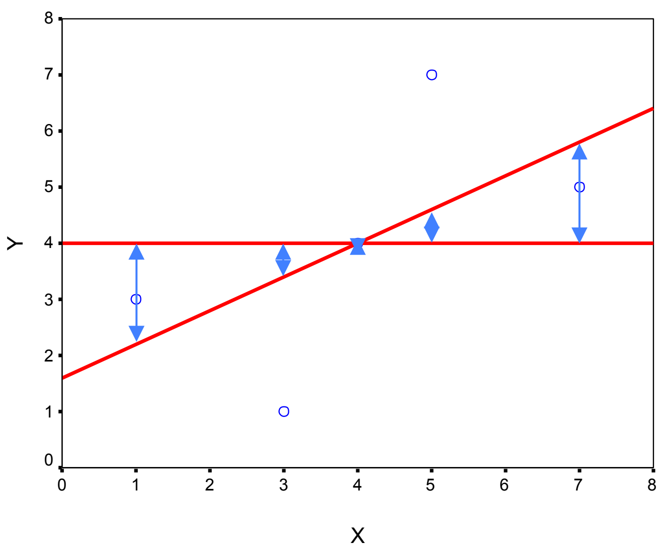
Soma de Quadrados da Regressão

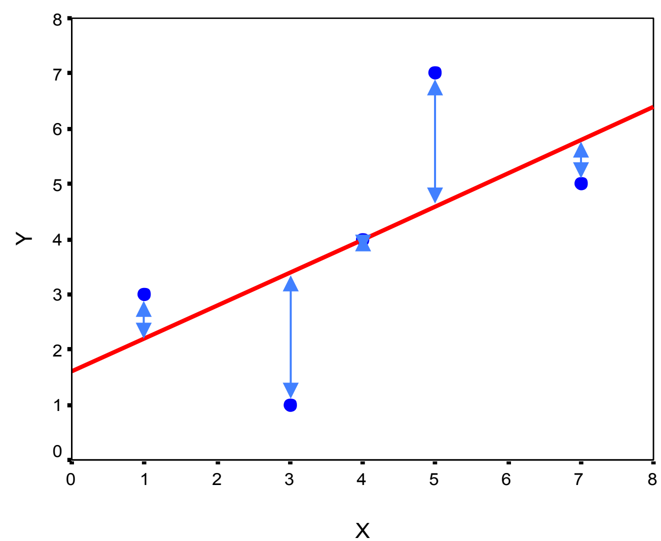
$SS_{Total} = \sum(Y - MY)^2$



$SS_{Model} = \sum(\hat{Y} - MY)^2$

$SS_{Residual} = \sum(Y - \hat{Y})^2$





Saídas do SPSS (Output) Dados do Modelo: 2as Tabelas

- Tabelas da ANOVA
  - SS_{Modelo} é chamado $SS_{Regression}$
  - Variação (i.e., SS) é dividida pelo *gl* (*df*) correspondente para calcular-se a a variância (i.e., MS)
    - MS = SS / *df*
  - *F* é calculado por $MS_{Modelo}/MS_{Residual}$
    - i.e., A variância explicada pelo modelo é maior do que a variância não explicada?
    - Avaliar a significância usando $df_{Numerador}$ e $df_{Denominador}$

Saídas do SPSS (Output) Dados do Modelo: 2as Tabelas

- ModeloTabelaSumária(*Summary table*)
  - R^2 =proporçãodavariânciatotal explicadapelomodelode regressão
    - $R^2 = SS_{Model} / SS_{Total}$
    - No exemplo: $R^2 = 7.2 / 20.0 = .36$
  - $R = \sqrt{R^2} = r_{Y\hat{Y}}$
    - No exemplo: $R = \sqrt{.36} = .60$
    - $R = r_{Y\hat{Y}} = .60$

Saídas do SPSS (Output) Dados do Modelo: 2as Tabelas

- Tabela sumária (cont.)
  - R^2 ajustado = Estimativa do parâmetro populacional , ρ (rho)
  - Em que p = # de preditores
    - No exemplo a: Adj. R^2 = 1-[(1-.36)(5-1)/(5-1-1)]
    - = 1-[(.64*4)/3] = .14667
  - Erro padrão da Estimativa (SEE)
    - Indica o grau de incerteza na previsão
    - SEE = $\sqrt{MS_{Residual}}$
    - No exemplo: SEE = $\sqrt{4.26667}$ = 2.06559

Coeficientes de regressão: *b*, *B*, &*β*

- Indicam a mudança em $\hat{Y}$ para cada 1a-unidade de mudança em *X*
  - *b* = coeficiente de regressão não padronizado
    - Mudança expressa em unidades do escore bruto
    - Mudança no escore bruto $\hat{Y}$para uma unidade de mudança no escore bruto de *X*
  - *B* = SPSS notação para *b*
  - *β* (Beta) = coeficiente de regressão padronizado.
    - Mudança expressa em unidades de desvio padrão (SD)
    - Número de mudanças em SD em $\hat{Y}$ para uma mudança de 1 SD em *X*

Coeficientes de regressão: *b*, *B*, & β

- *b* pode ser muito maior que β (& vice versa)
  - Isso dependerá da unidade de medida
- *b* = β quando:
  - $S_X = S_Y$
  - e.g., quando *X* e *Y* são escores *z*
- Interpretação
  - *b* é empregado quando as unidades tenham um sentido inerente
  - e.g., renda, altura, peso
  - β é empregado quando as métricas são arbitrárias
    - e.g., maioria das escalas psicológicas
- Note: *b* para (constante) é o termo constante (*intercept*, b_0)

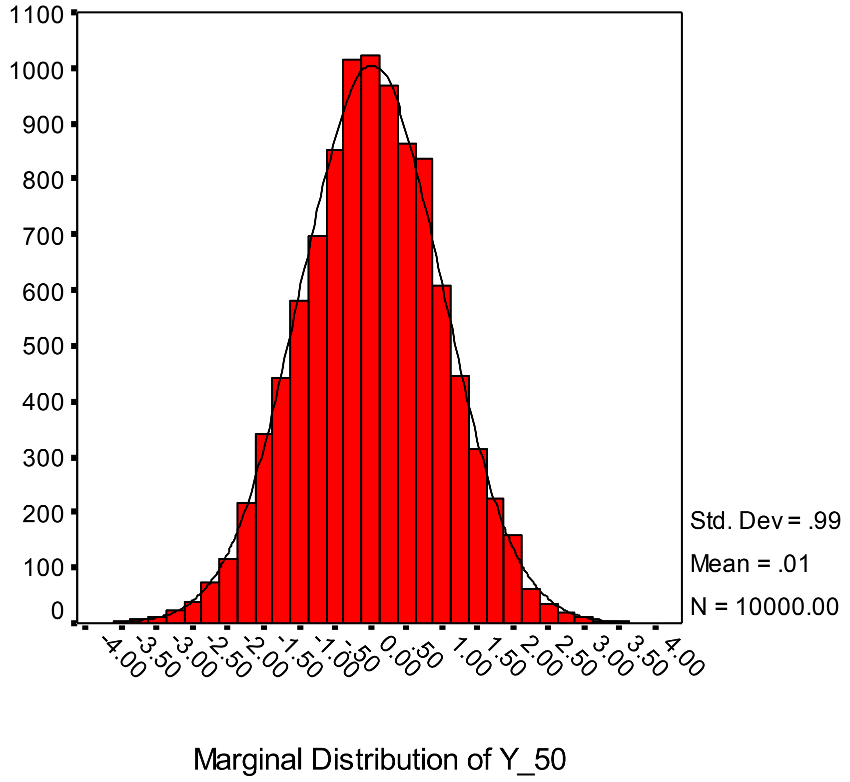
Pressupostos

- Correlação: como uma estatística descritiva não há assunções prévias
- Regressão: requer 2 assunções
  - Ambas relacionadas às distribuições condicionais

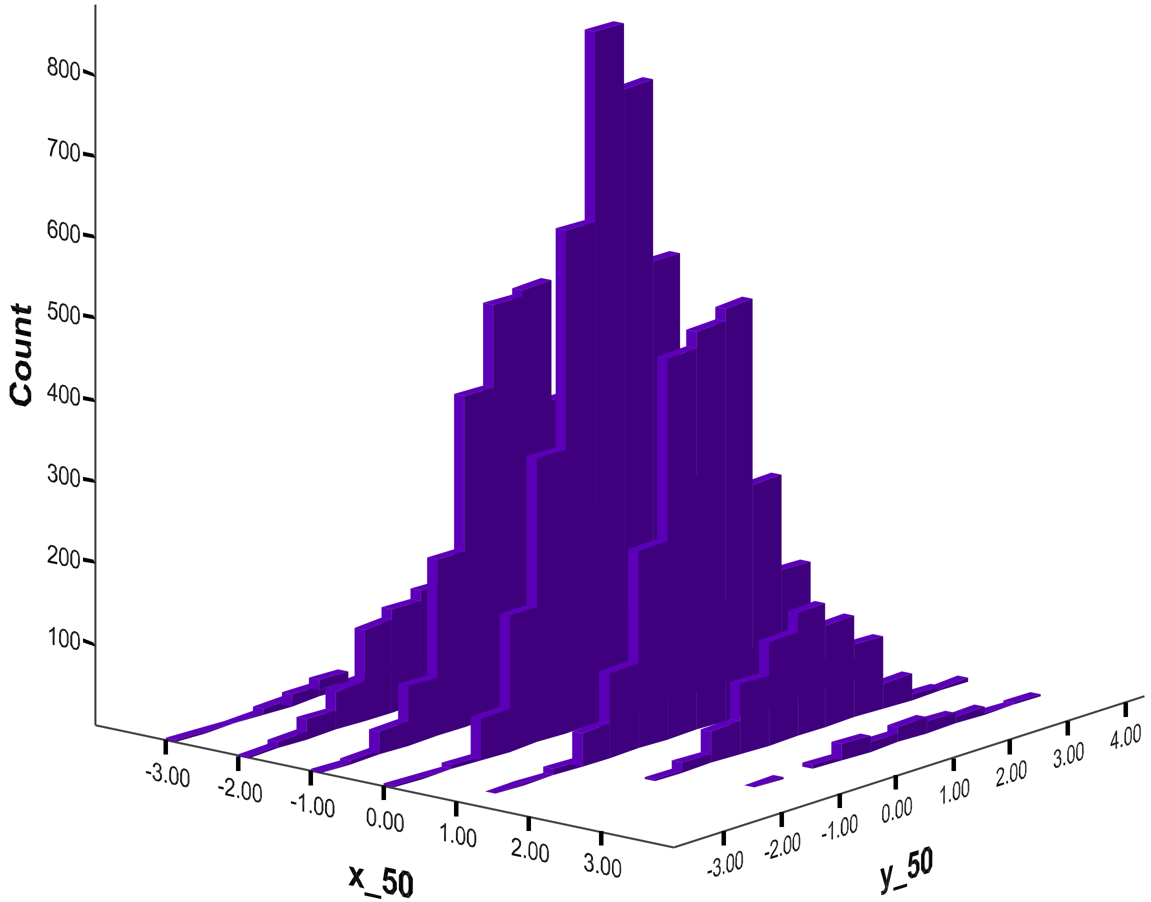
Distribuições Marginais e Condicionais

- Distribuições marginais:
  - Distribuição de *Y* desconsiderando *X*
  - i.e., Variância de *Y* transpassando todos os níveis de *X*: S^2_Y
- Distribuição condicional:
  - Distributição de *Y* condicionada em *X*
  - i.e, Variância de *Y em um dado valor de X*: $S^2_{Y \cdot X}$
- Considere as diferenças usando a sintaxe do SPSS
  - "Marginal andConditionalDistributions.sps"

Distribuição marginal de Y_50



Distribuiçãocondicional de Y_50



Duas assunções na regressão

- Normalidade:
  - Y é normalmente distribuído para cada valor de X
    - Procure por assimetria < 2.0 e curtose < 7.0

- Homocedasticidade:
  - S^2 de *Y é* constante quando calculado separadamente para cada valor específico de X; i.e., cada distribuição condicional
    - Análogo à assunção de variâncias iguais dentro dos grupos no *t*-test ou ANOVA
  - Justifica o uso de um único $MS_{Residual}$ para:
    - Calcular a significância estatística
    - Determinar o erro padrão da estimativa (SEE)
      - SEE é simplesmente o SD das distribuições condicionais
      - SEE = $S_{Y \cdot X}$

$$S_{Y \cdot X} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{N - 2}} = S_Y \sqrt{(1 - R^2)}$$

use $N - 2$ because estimated a and b from data