




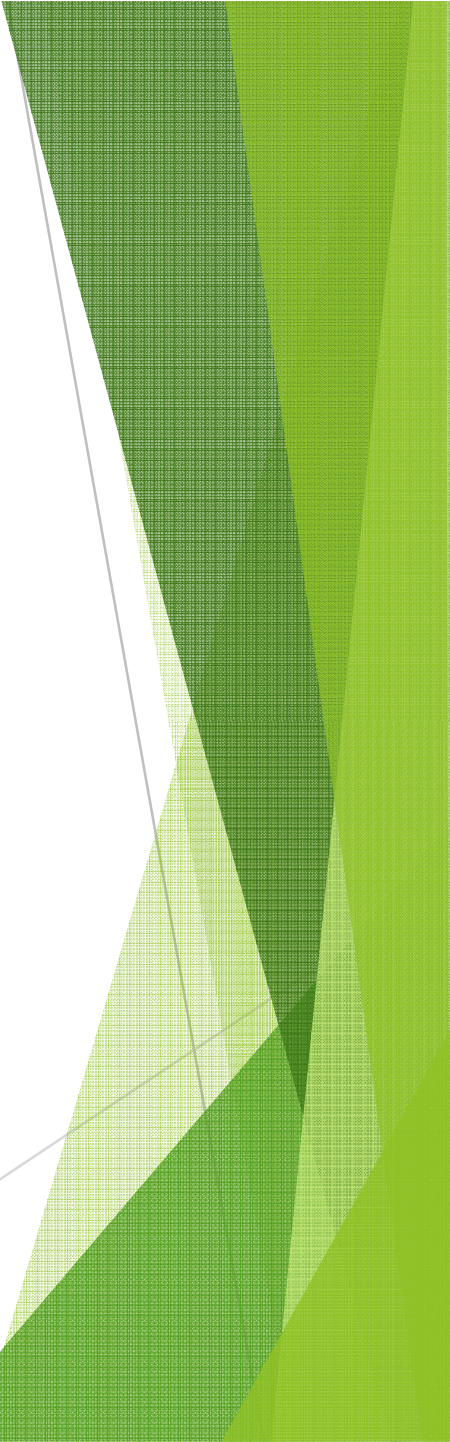
MINICURSO

Como planejar e otimizar experimentos ?

ARARANGUÁ - AGOSTO 2019



Como construir modelos empíricos

- 
- ▶ Nos modelos estudados, cada fator foi fixado em dois níveis
 - ▶ Por esta razão temos que nos contentar com uma visão limitada da equação que descreve a influência dos fatores na resposta
 - ▶ No exemplo analisado foram considerados rendimentos de 59% a 40°C e 90% a 60°C

Exemplo

- ▶ Se tivermos mais medidas para temperaturas intermediarias podemos ter uma estimativa melhor do modelo ajustado.
- ▶ Assim com mais três medidas:
- ▶

Temp	40	45	50	55	60
Rend %	60	70	77	86	91

- ▶ Considerando um modelo linear:
- ▶ “A melhor reta será a que passar ‘mais perto’ dos pontos experimentais”
- ▶ “ a maneira de conseguir esse resultado é **localizar a reta de tal maneira que a soma dos quadrados dos resíduos seja mínima...**”

$$\sum e_i^2$$

“no ajuste dos mínimos quadrados os valores ajustados de b_0 e b_1 são aqueles que tornam o somatório

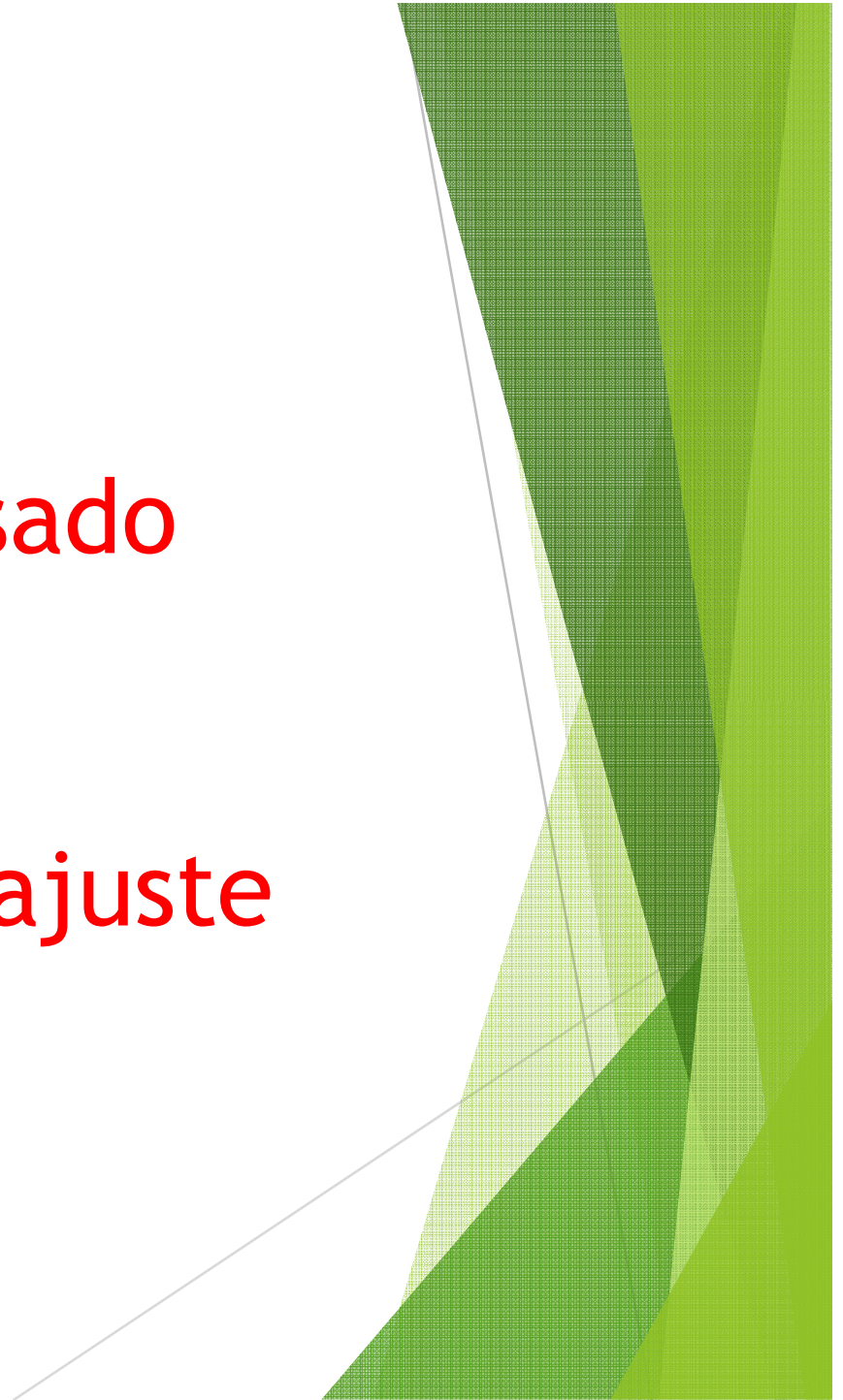
$$\sum e_i^2$$

O menor possível “

$$\hat{y}_i = b_0 + b_1 * X_1$$

Análise de variância

- ▶ O método mais usado para avaliar numericamente a qualidade de um ajuste de um modelo.



Análise de variância

- ▶ A qualidade de ajuste de um modelo depende da análise dos resíduos.
- ▶ Resíduo: Variação não explicada pelo modelo.
- ▶ O **desvio de uma resposta individual** em relação a média de todas respostas observadas pode ser decomposto em duas parcelas...

“Para fazer a ANOVA de um

- ... **começamos com uma decomposição algébrica dos desvios das respostas observadas em relação a resposta média global”**

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

$\bar{y} \rightarrow$ *valor · médio*

$y_i \rightarrow$ *valor · observado*

$\hat{y}_i \rightarrow$ *valor · predito*

$$\sum (y_i - \bar{y})^2 = \sum [(\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)]^2$$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + 2\sum (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) + \sum (y_i - \hat{y}_i)^2$$

- ▶ Estas somas de quadrados de desvios costumam ser chamadas de Somas Quadráticas (S.Q.), assim a equação abaixo pode ser lida como:

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

SQ em torno da média = SQ devido a regressão + SQ residual.

$$SQ_T = SQ_R + SQ_r$$

- ▶ Determina-se também R^2 que é o coeficiente de determinação (correlação) do modelo.

$$R^2 = \frac{SQ_R}{SQ_T} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

- ▶ Valor máximo = 1

Graus de liberdade

- ▶ O número total de graus de liberdade da soma quadrática (v_T) é $n-1$
- ▶ n = número de observações
- ▶ “num modelo com p parâmetros o número de graus de liberdade da soma quadrática **residual** (v_r) é dado por $n-p$ ”
- ▶ Se $v_T = v_R + v_r \dots$
... o número de graus de liberdade da soma quadrática da **Regressão** (v_R) é $p-1$

- ▶ Os resultados das considerações são reunidos na tabela chamada Tabela de Análise de variância ou simplesmente (ANOVA).

Fonte de variação	Soma quadrática	Graus de liberdade	Média quadrática
Regressão	SQ_R	$p-1$	$MQ_R = SQ_R / p-1$
Resíduos	SQ_r	$n-p$	$s^2 = SQ_r / n-p = MQ_r$
Total	SQ_T	$n-1$	

n = número total de observações; p = número de parâmetros do modelo
% explicado pelo modelo = $(SQ_R / SQ_T) * 100$

Exemplo

- ▶ Equação a ser ajustada $Red = b_0 + b_1 * T$
- ▶ Possui dois termos a serem determinados:
 b_0 e b_1

Temp	40	45	50	55	60
Rend %	60	70	77	86	91

Graus de liberdade

- ▶ A cada SQ está associado um número de graus de liberdade, que indica quantos valores independentes envolvendo as n observações (no caso 5) são necessários para determiná-la.
- ▶ Para a SQ_T o número de graus de liberdade é $(n-1)$.
- ▶ A SQ_R tem apenas um grau de liberdade, pois o modelo ajustado tem dois termos.
- ▶ A SQ_r deve ter $(n-2)$ graus de liberdade.
- ▶ Assim satisfazemos a equação mostrada anteriormente.
- ▶ $(n-1) = 1 + (n-2)$

$$v_T = v_R + v_r \dots$$

Intervalos de confiança

- ▶ **Objetivo** “cálculo do erro padrão para b_0 e $b_1\dots$ ”
- ▶ Ao postular o modelo admitimos que cada observação (y_i) é constituída de uma parte sistemática e uma parte aleatória (erro).
- ▶ Admitimos também as seguintes hipóteses:
- ▶ A variância dos erros é constante.
- ▶ Os erros observados em valores diferentes da variável independente não estão correlacionados.
- ▶ Os erros seguem uma distribuição normal.

Significância estatística da regressão

- Pode-se demonstrar que a razão entre as médias quadráticas MQ_R e MQ_r segue uma **distribuição F**.

$$F_{p-1, n-p} \approx \frac{MQ_R}{MQ_r} = \frac{[\sum (\hat{y}_i - \bar{y})^2] / (p-1)}{[\sum (y_i - \hat{y}_i)^2] / (n-p)}$$

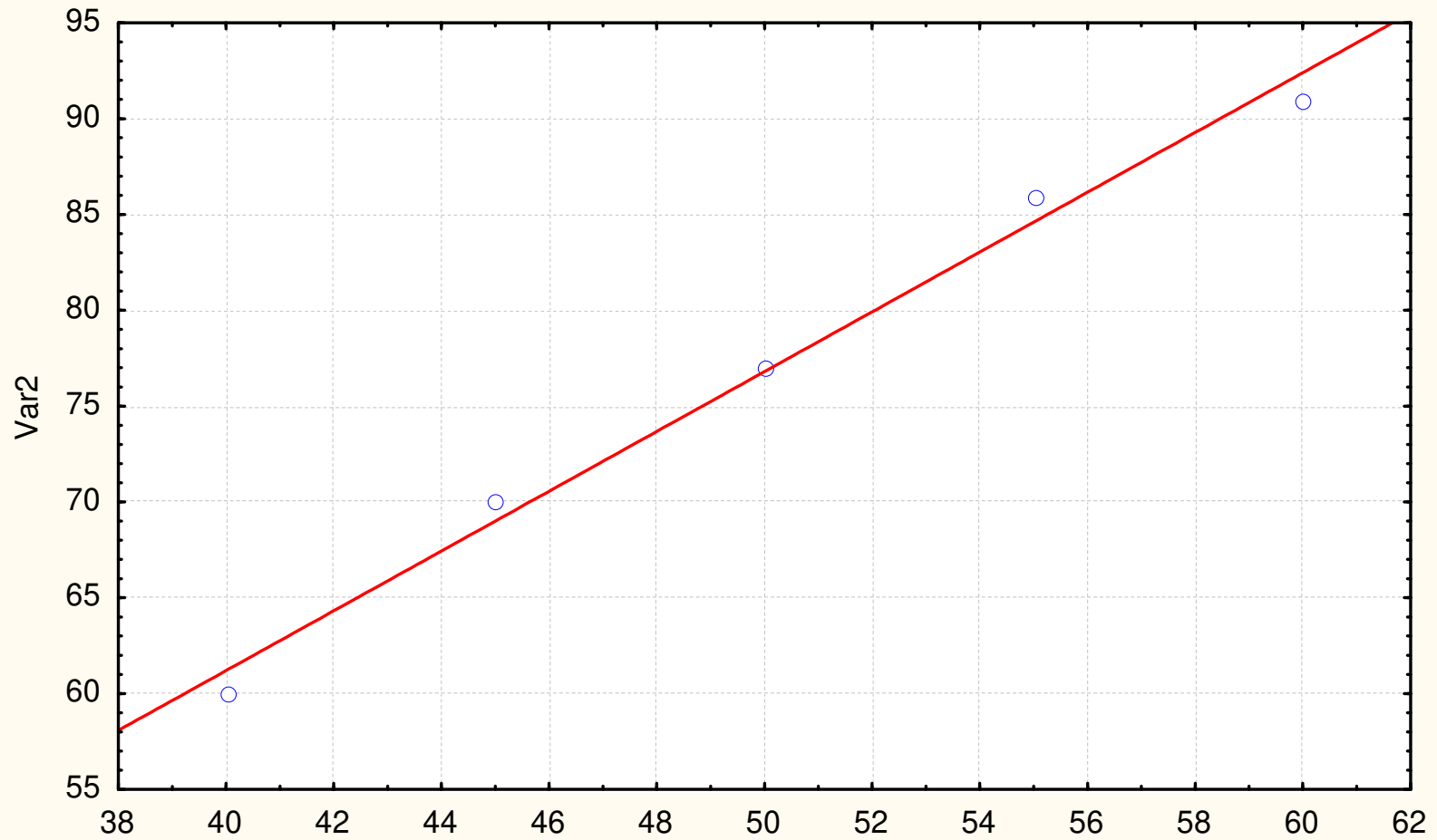
- Para o exemplo...

$$F_{1,3} \approx \frac{MQ_R}{MQ_r} = \frac{608,4 / 1,0}{6,4 / 3} = 285,2$$

Scatterplot of Var2 against Var1

Spreadsheet1 10v*10c

$$\text{Var2} = -1,2 + 1,56 * x$$



Var1:Var2: $y = -1,2 + 1,56 * x$; $r^2 = 0,9896$

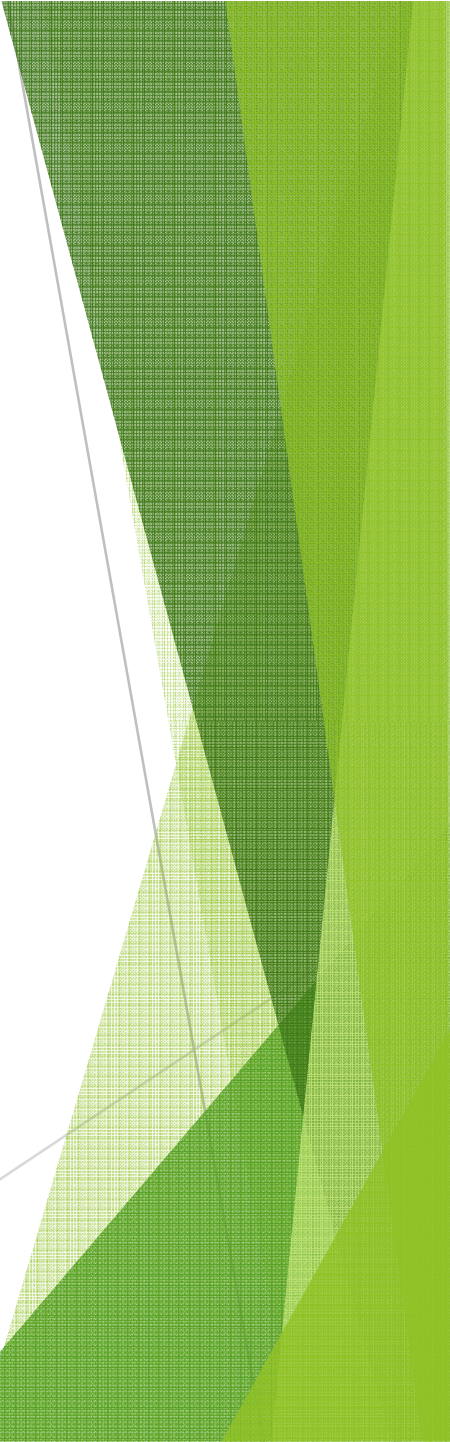
Var1

Analysis of Variance; DV: Var2 (Spreadsheet10.sta)

Effect	Sums of Squares	df	Mean Squares	F	p-level
Regress.	608,4000	1	608,4000	285,1875	0,000452
Residual	6,4000	3	2,1333		
Total	614,8000				

Regression Summary for Dependent Variable: Var2 (Spreadsheet10):
 $R = ,99478144$ $R^2 = ,98959011$ Adjusted $R^2 = ,98612015$
 $F(1,3) = 285,19$ $p < ,00045$ Std.Error of estimate: 1,4606

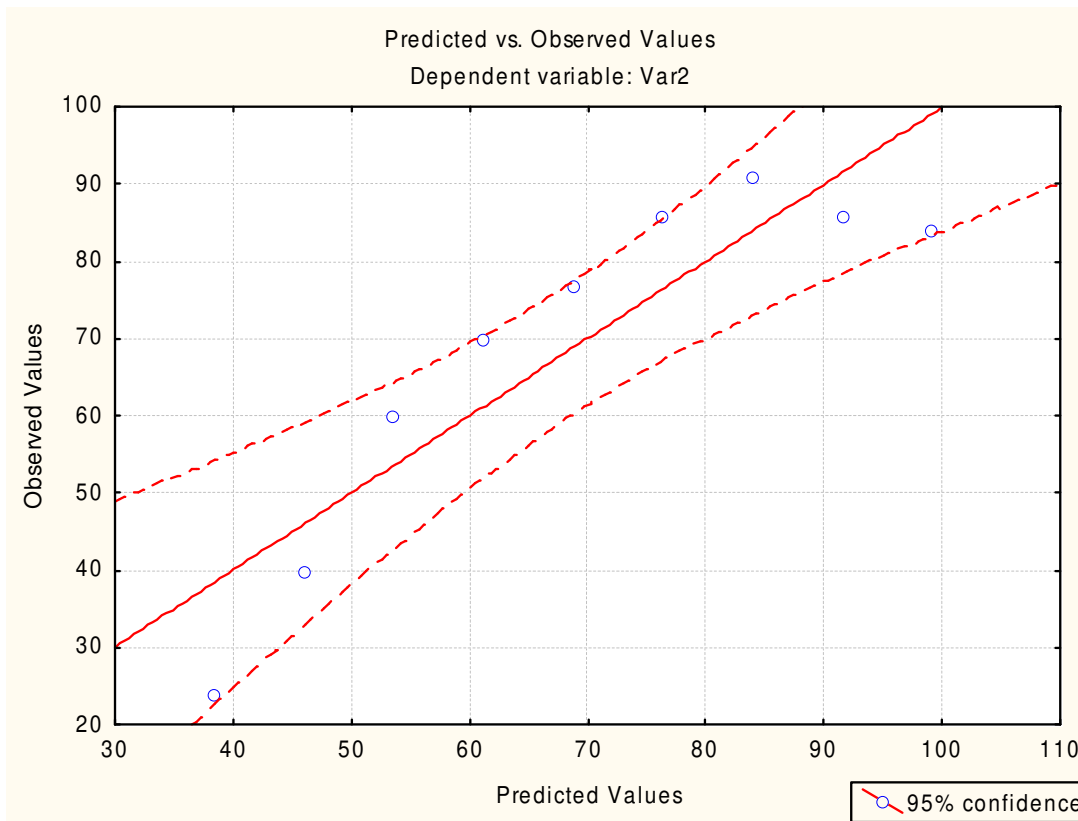
	Beta	Std. Err. of Beta	B	Std. Err. of B	t(3)	p-level
=5						
intercept			-1,20000	4,664762	-0,25725	0,813623
var1	0,994781	0,058906	1,56000	0,092376	16,88750	0,000452

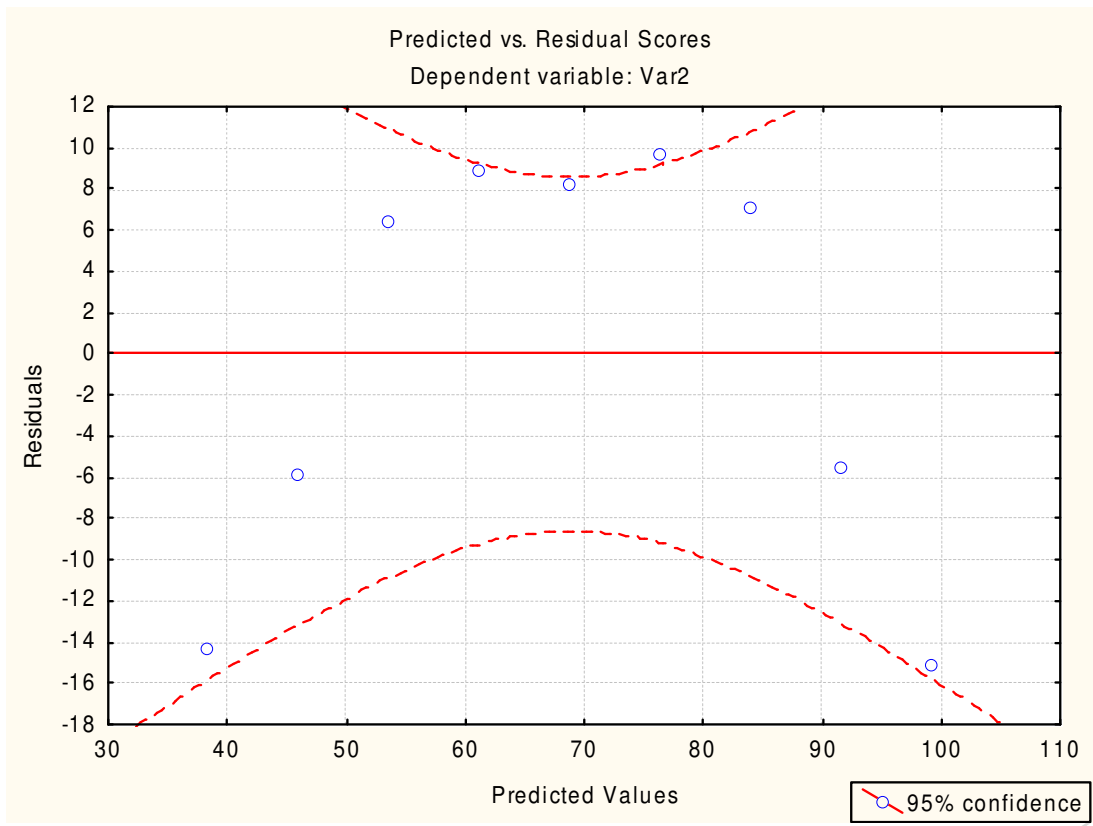
- 
- ▶ Para o exemplo, no nível de 95% de confiança temos que $F_{1,3} = 10,13$.
 - ▶ Assim a regressão calculada é estatisticamente significativa se
 - ▶ $MQ_R / MQ_r > 10,13$
 - ▶ Como o valor calculado é 285,6 temos que a **nossa equação é altamente significativa.**

Novo modelo

- ▶ Novo conjunto de dados com mais quatro ensaios

Temp	30	35	40	45	50	55	60	65	70
Rend %	24	40	60	70	77	86	91	86	84





- ▶ Para o modelo linear temos:
- ▶ $MQ_R/MQ_r = 29,14$
- ▶ Valor de $F_{1,7} = 5,59$ (nível de 95%)
- ▶ **Isso indicaria uma regressão significativa!**
- ▶ A percentagem de variação explicada pelo modelo é de 80,63%.
- ▶ Entretanto só poderíamos usar F se **não** houvesse anormalidade na distribuição de resíduos.

Um novo modelo para $Y=f(T)$

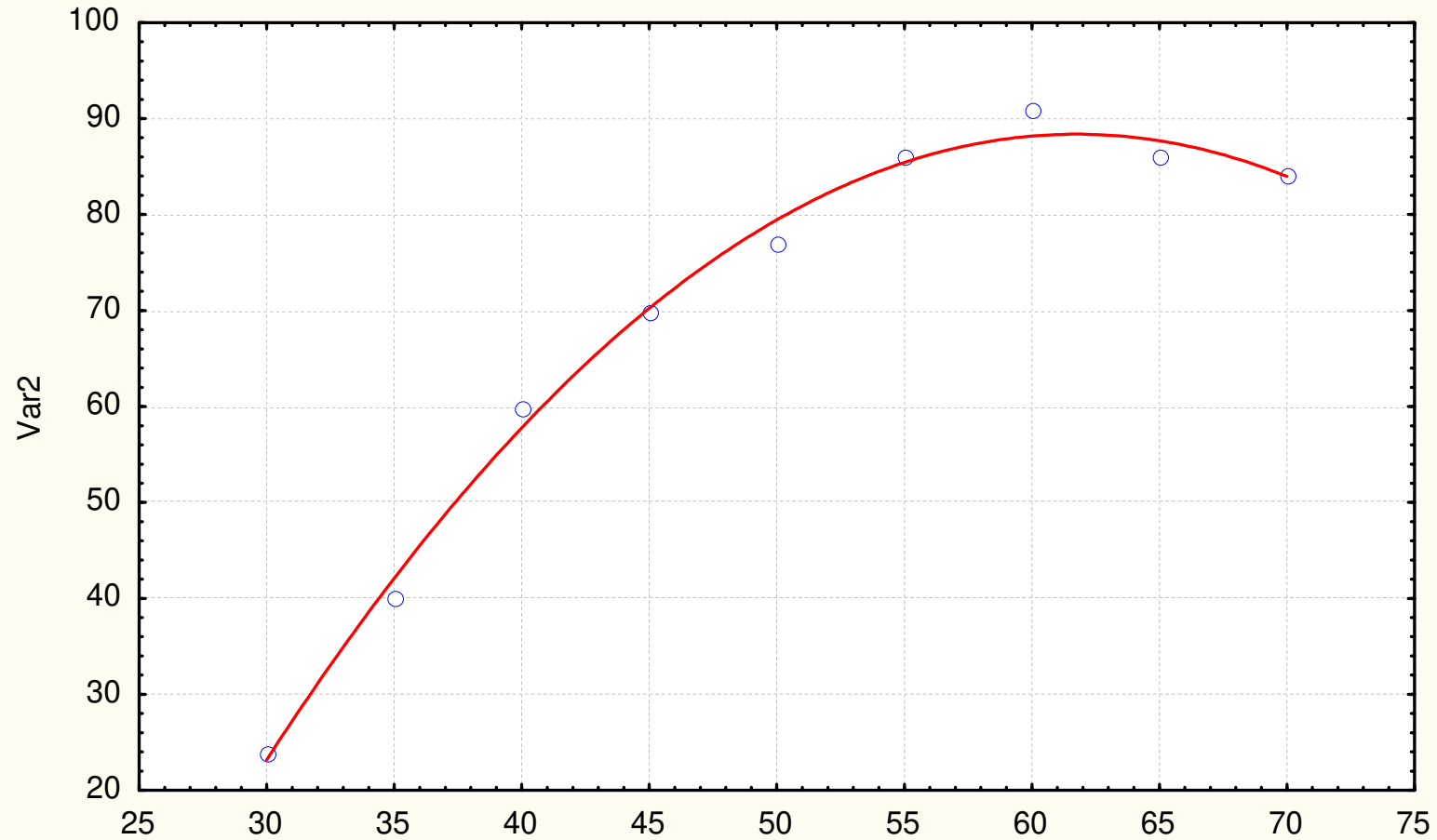
- ▶ Como o modelo linear é insatisfatório é necessário acrescentar mais um termo na equação.

$$\hat{y}_i = b_o + b_1 * X_i + b_2 * X_i^2$$

Scatterplot of Var2 against Var1

AutoRecovery save of 000008DCSpreadsheet1.sta 10v*10c

$$\text{Var2} = -158,2424 + 7,9875 * x - 0,0647 * x^2$$



Var1:Var2: $y = -7,3333 + 1,52 * x$; $r = 0,8980$; $p = 0,0010$

Anova e coeficientes

Analysis of Variance; DV: Var2 (Spreadsheet1)					
Effect	Sums of Squares	df	Mean Squares	F	p-level
Regress.	4270,808	2	2135,404	471,1800	0,000000
Residual	27,192	6	4,532		
Total	4298,000				

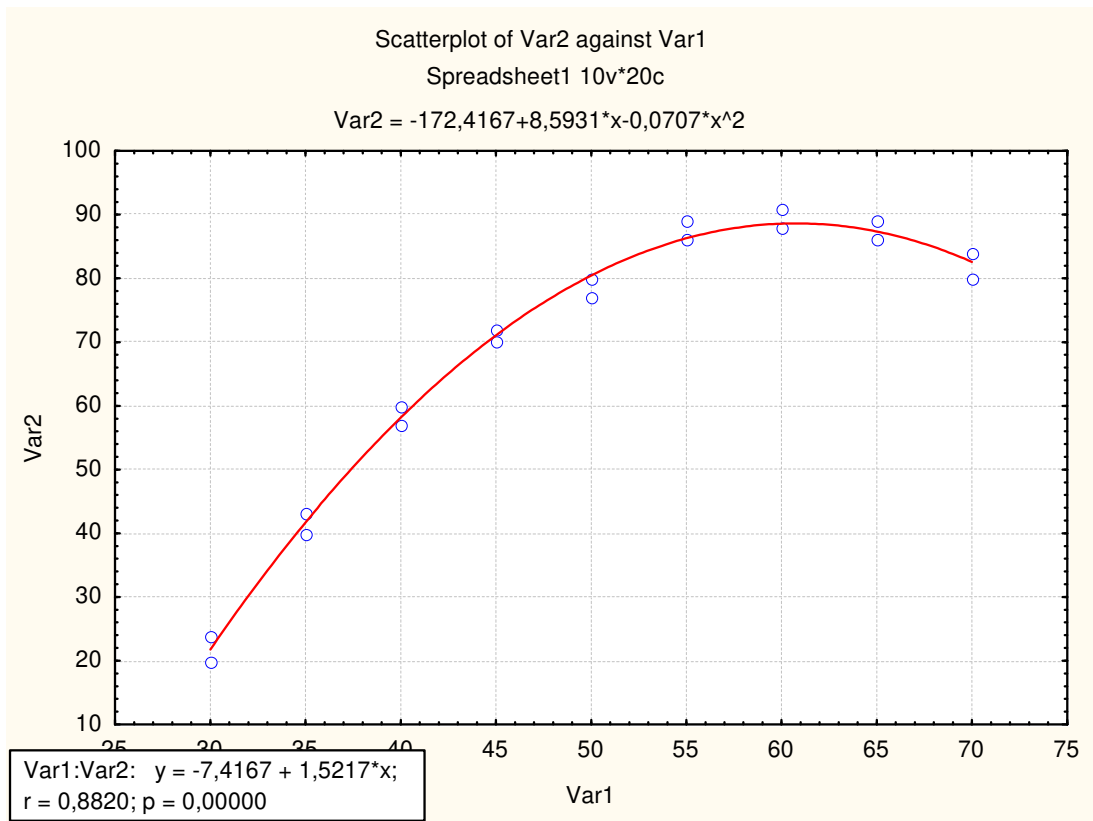
Regression Summary for Dependent Variable: Var2 (Spreadshe						
R= ,99683162 R ² = ,99367329 Adjusted R ² = ,99156438						
F(2,6)=471,18 p<,00000 Std.Error of estimate: 2,1289						
N=9	Beta	Std.Err. of Beta	B	Std.Err. of B	t(6)	p-level
Intercept			-158,242	11,67201	-13,5574	0,000010
Var1	4,71873	0,288478	7,988	0,48832	16,3573	0,000003
V1**2	-3,84521	0,288478	-0,065	0,00485	-13,3293	0,000011

- ▶ A avaliação foi feita na aparência do gráfico de resíduos.
- ▶ O novo modelo reproduz 99,37% da variação total.
- ▶ A razão MQ_R/MQ_r sobe para 471,4 (contra 29,14 do modelo linear) valor que deve ser comparado com $F_{2,6} = 5,14$

Novo experimento com respostas em duplicata

► Dados experimentais com duplicata

Temp	30	35	40	45	50	55	60	65	70
Rend %	24	40	60	70	77	86	91	86	84
	20	43	57	72	80	89	88	89	80



ANOVA

Fonte de variação	Soma quadrática	Graus de liberdade	Média quadrática
Regressão	8.871,61	2	4.435,80
Resíduos	58,40	15	3,89
Total	8.930	17	

Analysis of Variance; DV: Var2 (Spreadsheet1)					
Effect	Sums of Squares	df	Mean Squares	F	p-level
Regress.	8871,605	2	4435,802	1139,426	0,000000
Residual	58,395	15	3,893		
Total	8930,000				

► COEF.

Regression Summary for Dependent Variable: Var2 (Spreadshe						
R= ,99672503 R²= ,99346078 Adjusted R²= ,99258888						
F(2,15)=1139,4 p<,00000 Std.Error of estimate: 1,9731						
N=18	Beta	Std.Err. of Beta	B	Std.Err. of B	t(15)	p-level
Intercept			-172,417	7,649393	-22,5399	0,000000
Var1	4,98063	0,185488	8,593	0,320023	26,8515	0,000000
V1**2	-4,12488	0,185488	-0,071	0,003180	-22,2379	0,000000

► Logo podemos escrever o modelo (+/-)

► $Y = -172,42 + 8,59 * T + - 0,07 * T^2$

Falta de ajuste e erro puro

- ▶ O procedimento utilizado até agora está baseado na aparência do gráfico de resíduos.
- ▶ Se os experimentos apresentarem duplicatas, podemos utilizá-las para obter uma estimativa do **erro aleatório**.
- ▶ Assim a soma quadrática dos resíduos (SQ_r) pode ser decomposta em duas partes. (erro puro e falta de ajuste)

- ▶ Um termo nos dá a medida do **erro aleatório** é chamado de soma quadrática devida ao **erro puro** (SQ_{ep})
- ▶ O outro termo fornece a medida da **falta de ajuste** do modelo às respostas observadas, sendo chamado de soma quadrática devida a falta de ajuste (SQ_{faj})
- ▶ $SQ_r = SQ_{ep} + SQ_{faj}$
- ▶ Assim a tabela de ANOVA ganha duas novas linhas.

Versão completa da ANOVA

Fonte de variação	Soma quadrática	Graus de liberdade	Média quadrática
Regressão	SQ_R	$p-1$	$MQ_R = SQ_R / (p-1)$
Resíduos	SQ_r	$n-p$	$MQ_r = SQ_r / (n-p)$
Falta de ajuste	SQ_{faj}	$m-p$	$MQ_{faj} = SQ_{faj} / (m-p)$
Erro Puro	SQ_{ep}	$n-m$	$MQ_{ep} = SQ_{ep} / (n-m)$
Total	SQ_T	$n-1$	

n = número total de observações;
p = número de parâmetros do modelo

m = número de níveis distintos da variável independente;
% explicado pelo modelo = SQ_R / SQ_T

ANOVA

Fonte de variação	Soma quadrática	Graus de liberdade	Média quadrática
Regressão	8.871,61	2	4.435,80
Resíduos	58,40	15	3,89
Falta de ajuste	13,39	6	2,23
Erro Puro	45,00	9	5,00
Total	8.930	17	