Diagnóstico do Modelo e Tratamento

Frases

"Por serem mais precisos que as palavras, os números são particularmente adequados para transmitir conclusões científicas"

Pagano e Gauvre, 2004

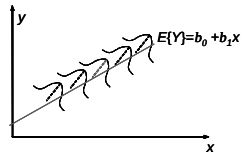
Roteiro

1. Suposições do Modelo
2. Análise de Resíduos
3. Aplicação
4. Tratamento de *Outliers*
5. Transformações Estabilizadoras de Variância
6. Omissão de Variáveis
7. Referências

Suposições do Modelo

$$Y_i = b_0 + b_1 X_i + e_i$$

- os termos de erro (e_1, e_2, ..., e_n) são variáveis aleatórias independentes;
- $E\{e_i\} = 0$;
- $Var\{e_i\} = s^2$; e
- $e_i \sim$ N,



Verificação da Adequação do Modelo (1)

- O modelo é adequado?
  A relação entre a resposta e o regressor é linear, pelo menos aproximadamente?
- Os erros têm distribuição normal?
- Os erros são independentes (não-correlacionados)?

Verificação da Adequação do Modelo (1)

- Os erros têm variância constante (homocedasticidade)?
- Existem valores discrepantes (*outliers*)?
- Foram omitidas variáveis importantes do modelo?

Verificação da Adequação do Modelo (3)

- Considerar a validade das hipóteses no exame da adequação do modelo
- As violações às hipóteses podem levar a:
  - √ modelos instáveis
  - √ estimação imprecisa

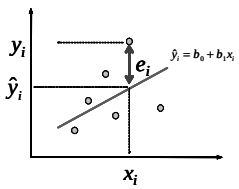
Análise de Resíduos

Análise dos Resíduos

- Um diagnóstico das suposições do modelo
  - Valor ajustado

    $\hat{y}_i = b_0 + b_1 x_i$
  - Resíduo

    $\hat{e}_i = y_i - \hat{y}_i$



Gráficos Utilizados no Diagnóstico

- Resíduos versus variáveis preditoras;
- Resíduos versus valores ajustados;
- Resíduos versus tempo (ou outro tipo de seqüência).
- Gráfico de normalidade dos resíduos.
- Histograma dos resíduos

Gráficos Complementares

- Gráfico dos resíduos absolutos ou quadráticos versus variáveis explicativas.
- Gráfico dos resíduos versus variáveis preditoras omitidas do modelo.
- Box-plot dos resíduos.

Verificação de Normalidade

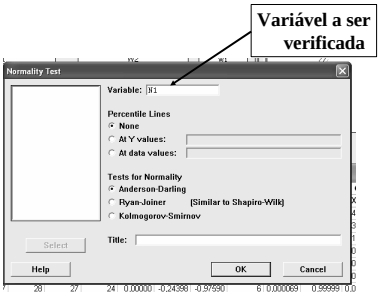
- Muitos procedimentos estatístico adotam a hipótese de normalidade dos dados
- Assim, é freqüente a necessidade de verificação de normalidade dos dados
- No Minitab, uma das maneiras de verificar a adequação do modelo (normal e outros) é através do *Probability Plot*

Verificação de Normalidade

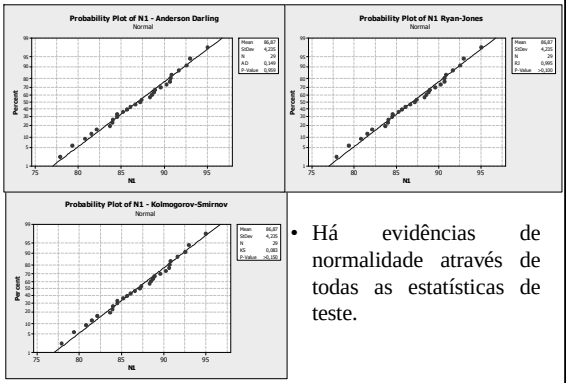
- Para verificação de normalidade, há três testes disponíveis:
  - √ Ryan-Joiner
  - √ Kolmogorov-Smirnov,
  - √ Anderson-Darling (é default do Minitab)

- Para acessá-los

Stat > Basic Statistics > Normality Teste ➔



Normalidade de Amostra *N1*

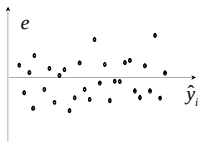


- Há evidências de normalidade através de todas as estatísticas de teste.

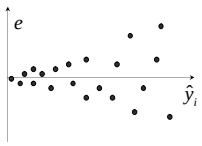
Normalidade dos Resíduos

- A normalidade é uma hipótese importante para as inferências envolvendo o modelo de regressão linear (teste de hipóteses, intervalo de confiança, etc.)
- A falta de normalidade pode ser uma indicação de heterocedasticidade ou falta de ajuste do modelo

Resíduos vs Valores Ajustados

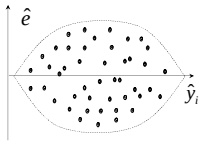


Não há defeitos óbvios no modelo

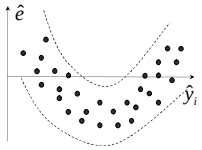


A variância é função crescente de *y*.

Resíduos vs Valores Ajustados

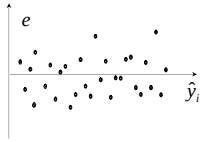


Ocorre freqüentemente quando *y* é uma proporção entre *0* e *1*.



Modelo não linear. Pode indicar necessidade de outros regressores.

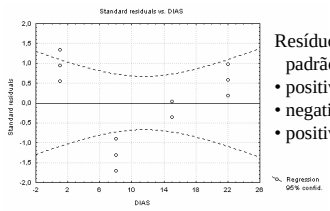
Resíduos vs. Regressor



- Exibem os mesmos padrões anteriores.
- No caso de regressão linear simples é necessários plotar apenas um dos gráficos (ou com valor ajustado ou com regressor).

Não Linearidade

- A verificação de adequação da função de regressão aos dados
- Uma verificação empírica pode ser efetuada através do gráfico dos resíduos vs. valores ajustados (ou vs. explicativas).



Resíduos vs. preditoras com padrão sistemático:
• positivos para valores baixo;
• negativos para valores médios
• positivos para valores altos.

- Nos casos em que houver um comportamento sistemático, devem-se incluir termos adicionais ou alternativos.

Heterocedasticidade

- Em geral, sua presença é verificada pelo gráfico dos resíduos vs variáveis explicativas (ou valores ajustados).
- Heterocedasticidade tende a produzir um gráfico com forma de funil.

Heterocedasticidade

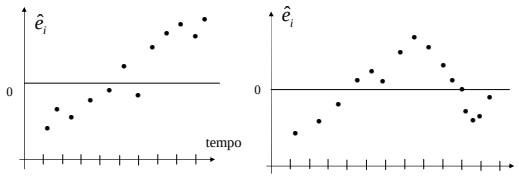
- Situações em que pode ocorrer:
  - √ Lucro vs. Tamanho da empresa:
    empresas maiores tendem a ter maior dispersão nos seus lucros.
  - √ Consumo de um Bem vs. Renda:
    pessoas ricas podem escolher melhor a proporção da renda consumida em determinado bem.

Autocorrelação Serial

- Os resíduos estão correlacionados e o valor de um resíduo passa a influenciar os resultados dos valores ajustados da resposta.
- A idéia da autocorrelação serial é que os resíduos contém mais informação sobre a variável dependente do que aquilo que foi "filtrado" pelas variáveis explicativas.
  - √ Em termos técnicos, o resíduo ainda pode ser sistematizado.

- Fontes de autocorrelação serial:
  - √ Omissão de variável relevante;
  - √ Má especificação da forma funcional
- Em geral, autocorrelação é encontrada em séries de tempo.

Resíduos vs. Seqüência



- Necessário para dados obtidos em seqüência (tempo, etc)
- Se os resíduos são independentes, distribuem-se aleatoriamente em torno de zero
- Pode indicar omissão de variável

Análise dos Resíduos

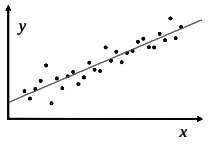


Gráfico dos dados:
(x_i, y_i)

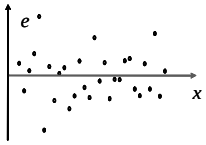


Gráfico dos resíduos:
(x_i, e_i)

- As suposições do modelo parecem satisfeitas?

Padronização dos Resíduos

- Útil para detectar valores extremos (discrepantes)
- Resíduos padronizados:

$$\hat{d}_i = \frac{\hat{e}_i}{\sqrt{\hat{s}^2}}$$

Resíduos potencialmente outliers: $\left|\hat{d}_i\right| > 3$

Resíduo Studentizado

- Melhora a escala do resíduo, utilizando o desvio-padrão exato.

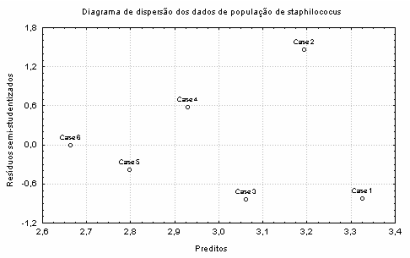
$$Var(\hat{e}_i) = s^2\left[1 - \left\{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right\}\right]$$

- Pontos próximos ao valor médio de *x* terão pior ajuste de mínimos quadrados (maior variância)
- Permitem a detecção de violações em pontos remotos (mais prováveis)

Outliers

- São valores extremos (atípicos), ou seja, são observações que não são bem ajustadas pelo modelo.
- Podem ser identificados a partir de um gráfico dos resíduos vs a variável preditora (ou valores ajustados).
- O uso de resíduos padronizados (ou studentizados) são úteis para identificar resíduos que estão muito afastados de zero, (em escala de desvios-padrão)

- Outliers podem influenciar fortemente o ajuste de mínimos quadrados;
- Não devem ser desprezados, admitindo-se seu descarte se representam um erro de registro, erro de medida, falha de equipamento ou algum problema similar.



- Gráfico de resíduos padronizados, sem outliers.

Aplicação

Um Modelo Econômico

- Objetivo: Estudar a relação entre renda familiar e despesas com alimentação.
- Experimento:

  Amostra aleatória de residências, com renda familiar semanal maior que $480
- Característica de interesse: Despesa semanal da residência com alimentação

  "Quanto foi gasto com alimentação na semana passada?"

Modelo Proposto

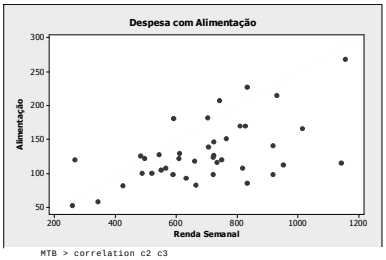
$$Y_i = ß_0 + ß_1 x_i + e_i$$

- Variáveis:

  Resposta (Y_i) : Despesa com alimentação

  Explicativa(x_i): Renda Familiar Semanal

  Erro (e_i): Todos os fatores que afetam *Y*, exceto renda
- Dados: *alimentacao*

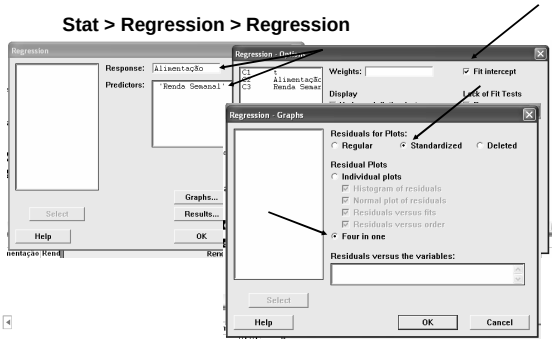



```
MTB > correlation c2 c3

Correlations: Alimentação; Renda Semanal

Pearson correlation of Alimentação and Renda Semanal = 0,546
P-Value = 0,000
```

- Há indícios de associação linear entre renda e despesas com alimentação

Ajuste de Mínimos Quadrados

Stat > Regression > Regression



Saída Minitab

Regression Analysis: Alimentação versus Renda Semanal

The regression equation is
Alimentação = 40,8 + 0,128 Renda Semanal

Predictor	Coef	SE Coef	T	P
Constant	40,77	22,14	1,84	0,073
Renda Semanal	0,12829	0,03054	4,20	0,000

S = 37,8054 R-Sq = 31,7% R-Sq(adj) = 29,9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	25221	25221	17,65	0,000
Residual Error	38	54311	1429		
Total	39	79533			

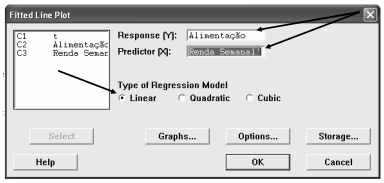
Unusual Observations

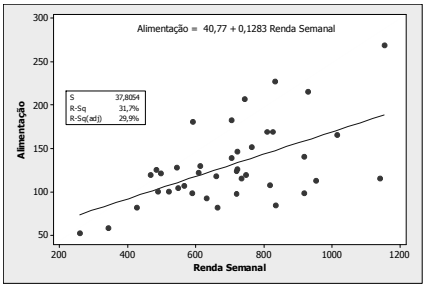
Obs	Renda Semanal	Alimentação	Fit	SE Fit	Residual	St Resid
1	258	52,25	73,90	14,70	-21,65	-0,62 X
32	833	227,11	147,67	7,27	79,44	2,14R
39	1141	115,43	187,18	14,80	-71,75	-2,06RX
40	1155	269,03	188,89	15,17	80,14	2,31RX

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.

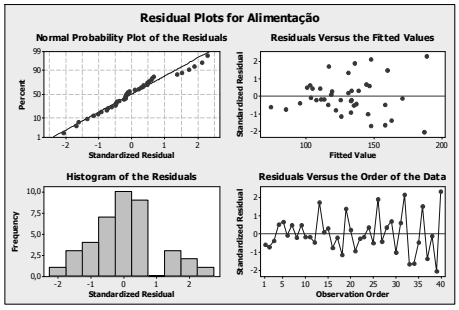
Reta de Regressão

Stat > Regression > Fitted Line Plot





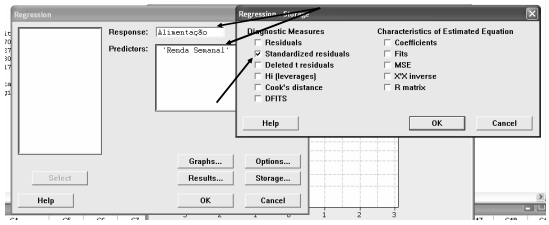
Análise dos Resíduos



Teste de Normalidade

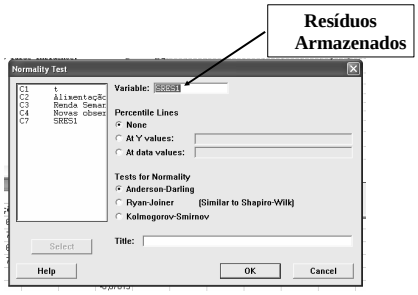
- Obtenção dos resíduos:

Stat > Regression > Regression

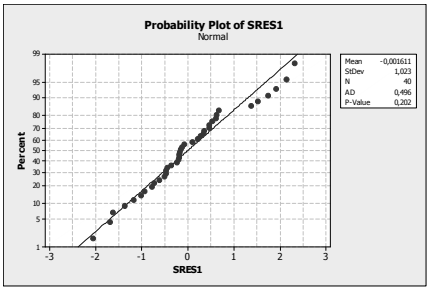


- Teste de Normalidade

Stat > Basic Statistics > Normality Teste ➔



Teste de Normalidade dos Resíduos



Não há evidências para se rejeitar a normalidade dos resíduos

Conclusões

- Não há evidências para se rejeitar a hipótese de normalidade dos resíduos;
- Os resíduos aparentam estar correlacionados
- O gráfico de resíduos vs. valores ajustados apresentam padrão indicador de heterocedasticidade ou omissão de variável relevante.

Conclusões

- Neste caso, provavelmente a renda não é suficiente para explicar o consumo (R^2 é baixo)
- Sugere-se o uso de mais variáveis explicativas para verificar se é possível o ajuste de modelo mais adequado.

Conseqüências das Violações das Hipóteses

- Heteroscedasticidade:
  - √ Sua presença tende a não viesar as estimativas dos parâmetros.
  - √ As variâncias estimadas não serão as corretas.
  - √ Logo, as inferências sobre os parâmetros estarão má especificadas.

Conseqüências das Violações das Hipóteses

- Autocorrelação serial:
  - √ As variâncias estimadas não serão as corretas,
  - √ Há possibilidade de viés nas estimativas se o problema for devido à omissão de variáveis relevantes no modelo.

Tratamento de *Outliers*

Outliers

- Pontos situados fora do padrão global das demais observações;
- Esses pontos discrepantes podem trazer problemas sérios no ajuste do modelo e na estimação dos parâmetros.

Outliers e Observações Influentes

- Uma observação é influente se sua remoção modifica substancialmente o resultado
- Os ouliers na direção de *y* têm grandes resíduos de regressão
- Os outliers na direção de *x* em geral são inlfuente para a reta de regressão, sem apresentar necessariamente grande resíduos.

Exemplo 2 – Telefones *vs* ICMS

- Dados sobre número de telefones e arrecadação do ICMS em 13 sub-regiões administrativas do estado de São Paulo.
- Observações padronizadas em relação ao número de habitantes de cada sub-região

$$telefone = \frac{\#\,telefones}{\#\,habitantes} x 1.000 \quad arrecadacao = \frac{Total\ de\ ICMS\ (em\ \$1.000)}{\#\,habitantes}$$

- Planilha: *telefones*
- Fonte: Bussab et al. (1988)

Dados

Sub-região	*Qte. Telefones*	*Arrecadação ICMS*	*Região*
Dracena	42	1,95	Interior
Adamantina	44	2,39	Interior
Avaré	48	2,5	Interior
Catanduva	53	3,22	Interior
Araçatuba	56	3,63	Interior
Lins	58	3,54	Interior
Assis	58	3,65	Interior
Franca	65	4,49	Interior
S. Carlos	68	5,78	Interior
Bauru	70	5,4	Interior
S. Sebastião	77	1,14	Litoral
S. J. Campos	86	13,94	Vale do Paraíba
S. Paulo	138	12,66	Capital

Caso 1 – Apenas Interior

Regression Analysis: arrecadacao versus telefones

The regression equation is
arrecadacao = - 3,48 + 0,127 telefones

Predictor	Coef	SE Coef	T	P
Constant	-3,4768	0,6007	-5,79	0,000
telefones	0,12690	0,01055	12,03	0,000

Regressão significante

S = 0,306025 R-Sq = 94,8% R-Sq(adj) = 94,1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	13,553	13,553	144,72	0,000
Residual Error	8	0,749	0,094		
Total	9	14,302			

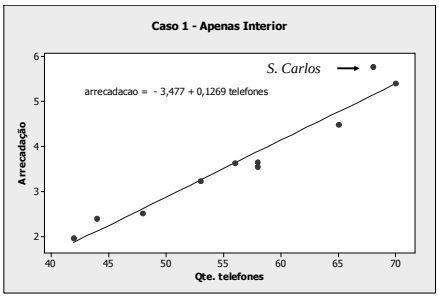
Unusual Observations

Obs	telefones	arrecadacao	Fit	SE Fit	Residual	St Resid
9	68,0	5,7800	5,1524	0,1577	0,6276	2,39R

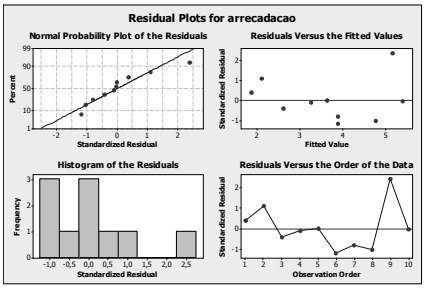
R denotes an observation with a large standardized residual.

S. Carlos →

Caso 1 – Reta Ajustada



Caso 1 – Análise Gráfica de Resíduos



Caso 2 – Interior e Litoral

Regression Analysis: arrecadacao versus telefones

The regression equation is
arrecadacao = 0,64 + 0,0480 telefones

Predictor	Coef	SE Coef	T	P
Constant	0,640	2,324	0,28	0,789
telefones	0,04797	0,03935	1,22	0,254

Regressão não significante

S = 1,38284 R-Sq = 14,2% R-Sq(adj) = 4,6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	2,842	2,842	1,49	0,254
Residual Error	9	17,210	1,912		
Total	10	20,052			

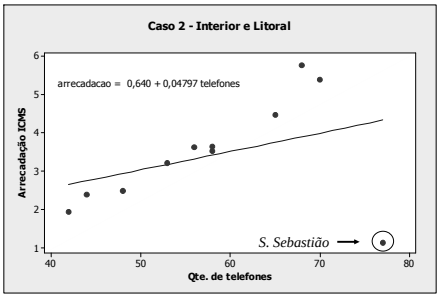
Unusual Observations

Obs	telefones	arrecadacao	Fit	SE Fit	Residual	St Resid
11	77,0	1,140	4,334	0,853	-3,194	-2,93R

R denotes an observation with a large standardized residual.

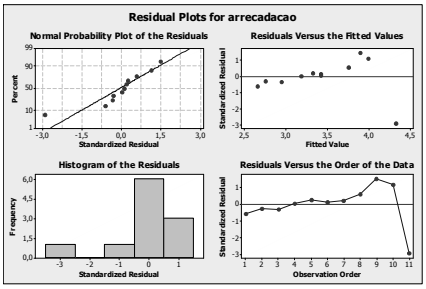
S. Sebastião

Caso 2 – Reta Ajustada



- Alta taxa de telefones instalados e baixa arrecadação

Caso 2 – Análise Gráfica de Resíduos



Caso 3 – Interior e Interior Industrializado

Regression Analysis: arrecadacao versus telefones

The regression equation is
arrecadacao = - 9,18 + 0,234 telefones

Predictor	Coef	SE Coef	T	P
Constant	-9,180	2,229	-4,12	0,003
telefones	0,23375	0,03705	6,31	0,000

Regressão significante

S = 1,50456 R-Sq = 81,6% R-Sq(adj) = 79,5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	90,094	90,094	39,80	0,000
Residual Error	9	20,373	2,264		
Total	10	110,467			

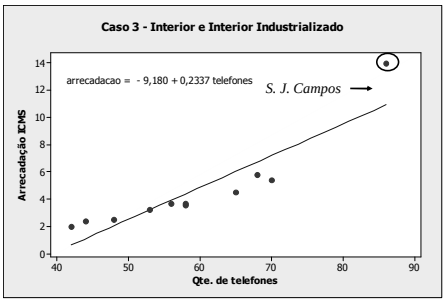
Unusual Observations

Obs	telefones	arrecadacao	Fit	SE Fit	Residual	St Resid
11	86,0	13,940	10,922	1,102	3,018	2,94R

R denotes an observation with a large standardized residual.

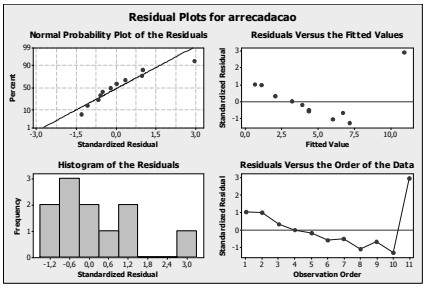
S. J. Campos

Caso 3 – Reta Ajustada



Alta arrecadação; telefones instalados fora do padrão geral

Caso 3 – Análise Gráfica de Resíduos



Caso 4 – Interior e Capital

Regression Analysis: arrecadacao versus telefones

The regression equation is
arrecadacao = - 2,66 + 0,112 telefones

Predictor	Coef	SE Coef	T	P
Constant	-2,6619	0,2682	-9,92	0,000
telefones	0,112129	0,003921	28,60	0,000

Regressão significante

S = 0,326302 R-Sq = 98,9% R-Sq(adj) = 98,8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	87,062	87,062	817,69	0,000
Residual Error	9	0,958	0,106		
Total	10	88,020			

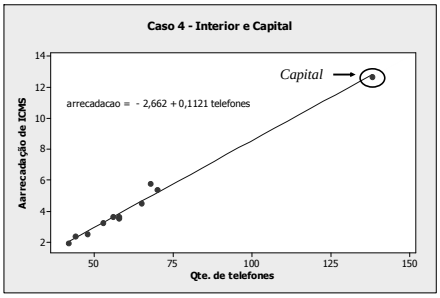
Unusual Observations

Obs	telefones	arrecadacao	Fit	SE Fit	Residual	St Resid
9	68	5,7800	4,9629	0,0999	0,8171	2,63R
11	138	12,6600	12,8120	0,3077	-0,1520	-1,40 X

Capital

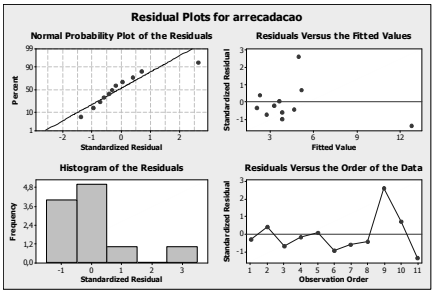
R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.

Caso 4 – Reta Ajustada

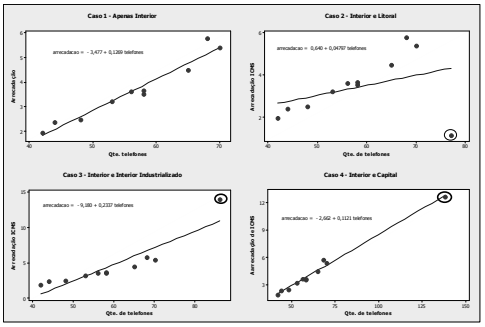


Taxa de telefones instalados muito afastada

Caso 4 – Análise Gráfica de Resíduos



Comparação dos Casos



Comparação dos Casos

	Caso 1	*Caso 2*	*Caso 3*	*Caso 4*
Inclinação	*0,12690*	*0,04797*	*0,23375*	*0,112129*
	0,01055	*0,039351*	*0,03705*	*0,003921*
Estatística T	*12,03*	*1,22*	*6,31*	*28,60*
R^2	*94,8%*	*14,2%*	*81,6%*	*98,9%*
	0,094	*1,912*	*2,264*	*0,106*
Intercepto	*-3,4768*	*0,640*	*-9,180*	*-2,6619*

Comentários

Modelo 1:

- Modelo significante;
- Análise gráfica de resíduos não fornece evidências de as suposições terem sido violadas
- O modelo parece ser adequado para explicar o aumento da arrecadação

Comentários (2)

Modelo 2:

- Observação acrescentada é litorânea, com muitas residências temporárias (alta taxa de telefones e baixa arrecadação)
- A regressão não é significante
- É razoável adotar o modelo sem esta observação

Comentários (3)

Modelo 3:

- A observação acrescentada tem arrecadação alta, mas nem tanto o nível de telefones instalados.
- É razoável adotar o modelo 1, com ressalva desta situação (esperam-se poucas situações com o mesmo comportamento)

Comentários (4)

Modelo 4:

- A observação acrescentada não afeta as estimativas do modelo 1;
- É um ponto que pode ser eliminado, pois, é praticamente caso único (valor exagerado da variável preditora *x*)

Tratamento de *outliers*

- *Outlier* é um problema sério na construção de modelos de regressão:
- Etapas para detecção e tratamento:
  - √ *Identificação de possíveis pontos discrepantes;*
  - √ *Avaliação dos efeitos sobre as estimativas e previsões*
  - √ *Análise criteriosa para eliminação ou não da observação*

Tratamento de *Outliers (2)*

- Pontos discrepantes resultados claramente de erro de mensuração devem ser corrigidos ou removidos do conjunto de dados;
- Outro tipo de *outlier*:

  Usa-se a análise de resíduos para detectar o ponto que apresenta resíduo muito grande (observações desajustadas)

Exemplo 3 – Desemprego Juvenil

- Taxas de desemprego de adolescentes e seguro desemprego, na Austrália, de 1962 a 1980
- Variáveis:
  - y_1: taxa de desemprego adolescentes homens
  - y_2: taxa de desemprego adolescentes mulheres
  - x: seguro desemprego (dólares constantes 1981)
- Fonte: Maddala
- Planilha: *desemprego_adol_Australia*

Homens

Regression Analysis: Homens versus Dólares ctes 1981

The regression equation is
Homens = 2,48 + 0,212 Dólares ctes 1981

Predictor	Coef	SE Coef	T	P
Constant	2,478	1,227	2,02	0,060
Dólares ctes 1981	0,21183	0,03431	6,17	0,000

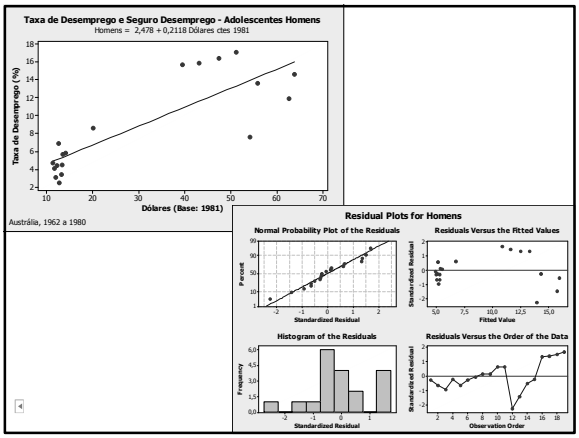
S = 2,98987 R-Sq = 69,2% R-Sq(adj) = 67,3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	340,76	340,76	38,12	0,000
Residual Error	17	151,97	8,94		
Total	18	492,72			

Obs	Dólares ctes 1981	Homens	Fit	SE Fit	Residual	St Resid	
1	13,3	4,500	5,296	0,887	-0,796	-0,28	
2	13,2	3,400	5,275	0,889	-1,875	-0,66	
3	12,7	2,500	5,169	0,900	-2,669	-0,94	
4	12,2	4,400	5,063	0,911	-0,663	-0,23	
5	11,9	3,100	4,999	0,918	-1,899	-0,67	
6	11,6	4,100	4,936	0,925	-0,836	-0,29	
7	11,3	4,700	4,872	0,931	-0,172	-0,06	
8	14,1	5,800	5,465	0,869	0,335	0,12	
9	13,4	5,700	5,317	0,884	0,383	0,13	
10	12,5	6,900	5,126	0,904	1,774	0,62	
11	20,0	8,600	6,715	0,762	1,885	0,65	
12	54,1	7,600	13,939	1,083	-6,339	-2,27R	73-74
13	62,7	11,900	15,760	1,325	-3,860	-1,44	
14	63,8	14,600	15,993	1,357	-1,393	-0,52	
15	55,8	13,600	14,299	1,129	-0,699	-0,25	
16	51,1	17,100	13,303	1,006	3,797	1,36	77-80
17	47,4	16,400	12,519	0,917	3,881	1,35	
18	43,1	15,900	11,608	0,826	4,292	1,49	
19	39,5	15,700	10,846	0,764	4,854	1,68	

R denotes an observation with a large standardized residual.



Mulheres

Regression Analysis: Mulheres versus Dólares ctes 1981

The regression equation is
Mulheres = 3,31 + 0,226 Dólares ctes 1981

Predictor	Coef	SE Coef	T	P
Constant	3,314	1,411	2,35	0,031
Dólares ctes 1981	0,22606	0,03943	5,73	0,000

S = 3,43628 R-Sq = 65,9% R-Sq(adj) = 63,9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	388,06	388,06	32,86	0,000
Residual Error	17	200,74	11,81		
Total	18	588,79			

Obs	Dólares ctes 1981	Mulheres	Fit	SE Fit	Residual	St Resid	
1	13,3	5,900	6,321	1,019	-0,421	-0,13	
2	13,2	4,500	6,298	1,021	-1,798	-0,55	
3	12,7	4,400	6,185	1,034	-1,785	-0,54	
4	12,2	5,100	6,072	1,047	-0,972	-0,30	
5	11,9	5,500	6,004	1,055	-0,504	-0,15	
6	11,6	5,700	5,937	1,063	-0,237	-0,07	
7	11,3	6,000	5,869	1,071	0,131	0,04	
8	14,1	6,300	6,502	0,999	-0,202	-0,06	
9	13,4	5,100	6,343	1,016	-1,243	-0,38	
10	12,5	6,100	6,140	1,039	-0,040	-0,01	
11	20,0	9,300	7,835	0,876	1,465	0,44	
12	54,1	7,800	15,544	1,245	-7,744	-2,42R	} 73-74
13	62,7	12,300	17,488	1,523	-5,188	-1,68	
14	63,8	16,700	17,737	1,560	-1,037	-0,34	
15	55,8	15,600	15,928	1,297	-0,328	-0,10	
16	51,1	18,800	14,866	1,156	3,934	1,22	} 77-80
17	47,4	18,800	14,029	1,054	4,771	1,46	
18	43,1	18,900	13,057	0,950	5,843	1,77	
19	39,5	17,600	12,244	0,879	5,356	1,61	

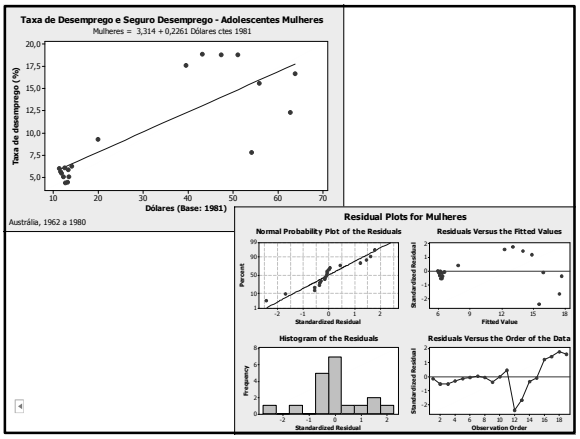
R denotes an observation with a large standardized residual.

Resumo

$$\hat{y}_1 = \underset{(1,227)}{2,478} + \underset{(0,0343)}{0,212x} \qquad R^2 = 69,2\%$$

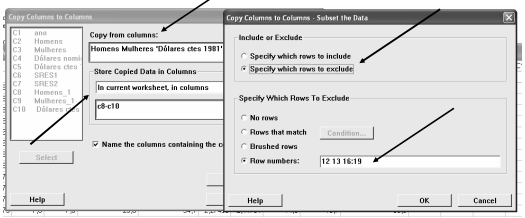
$$\hat{y}_2 = \underset{(1,410)}{3,314} + \underset{(0,0394)}{0,226x} \qquad R^2 = 65,9\%$$

- As estimativas dos coeficientes de x não mudaram muito e estão mais precisas
- Os valores de R^2 estão mais altos



Exclusão dos Valores Outliers – Minitab

Data > Copy > Columns to Columns ➔



Exclusão dos \

Regression Analysis: Homens_1 versus Dólares ctes 1981_1

The regression equation is
Homens_1 = 2,16 + 0,203 Dólares ctes 1981_1

Predictor	Coef	SE Coef	T	P
Constant	2,1557	0,6103	3,53	0,005
Dólares ctes 1981_1	0,20269	0,02296	8,83	0,000

S = 1,40590 R-Sq = 87,6% R-Sq(adj) = 86,5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	154,04	154,04	77,93	0,000
Residual Error	11	21,74	1,98		
Total	12	175,78			

Regression Analysis: Mulheres_1 versus Dólares ctes 1981_1

The regression equation is
Mulheres_1 = 2,80 + 0,225 Dólares ctes 1981_1

Predictor	Coef	SE Coef	T	P
Constant	2,7985	0,3936	7,11	0,000
Dólares ctes 1981_1	0,22505	0,01481	15,20	0,000

S = 0,906736 R-Sq = 95,5% R-Sq(adj) = 95,0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	189,90	189,90	230,97	0,000
Residual Error	11	9,04	0,82		
Total	12	198,94			

Resumo – Exclusão Valores

$$\hat{y}_1 = \underset{(0,610)}{2,156} + \underset{(0,0230)}{0,203x} \qquad R^2 = 87,6\%$$

$$\hat{y}_2 = \underset{(0,394)}{2,799} + \underset{(0,0148)}{0,225x} \qquad R^2 = 95,5\%$$

- Equações sugerem que aumento no seguro desemprego leva a aumento na taxa de desemprego
- Resíduos grandes para 1973, 1974 e 1977 a 1980.

Comentários

- Neste exemplo não seria correto a exclusão das observações com altos resíduos
- Os resíduos podem ser devidos ao lapso de tempo entre o requerimento do benefício e sua concessão, que deveria ser incorporado no modelo
- O exemplo ilustra o fato de que nem todos os *outliers* devem ser excluídos

Transformações Estabilizadoras

Homocedasticidade

- É hipótese fundamental na derivação das propriedades amostrais dos estimadores
- Exemplo de heterocedasticidade:

  Espera-se que o ganho assalariado segundo o nível de instrução não seja homogêneo :

  √ Baixo nível de instrução ➔ Salários baixos e próximos uns dos outros

- A ausência de homogeneidade apresenta quase sempre um padrão identificável;
- A variabilidade é, em geral, função da variável preditora ou de uma outra variável independente.
- Os EMQO, em presença de heterocedasticidade continuam não-viciados, mas não possuem mais a propriedade de variância mínima;
- É importante corrigir a heterocedasticidade

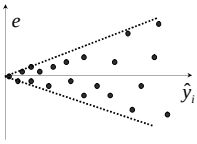
Correção da Heterocedasticidade

- Mínimos quadrados ponderados;
- Transformação da variável respota *y* ou da preditora *x*

  Transformações estabilizadoras da variância

Transformação de Estabilização Variância (1)

Gráfico do Resíduo:



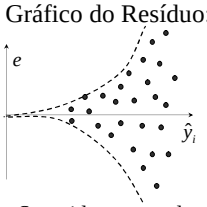
Transformação:

$\sqrt{y}$

- Recomendada quando $Var(e_i)$ é proporcional a x_i.

Transformação de Estabilização Variância (2)

Gráfico do Resíduo:



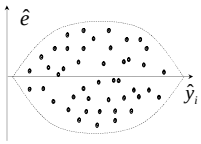
Transformação:

$\log y$

- Sugerido quando a variância cresce mais acentuadamente, isto é, proporcional a x_i^2.

Transformação de Estabilização Variância (2)

Gráfico do Resíduo:



Transformação:

$arcsen\sqrt{y}$

- Sugerido quando a variável resposta é do tipo proporção, isto é, $0 = y = 1$.

Exemplo 4 – Empresa de Eletricidade

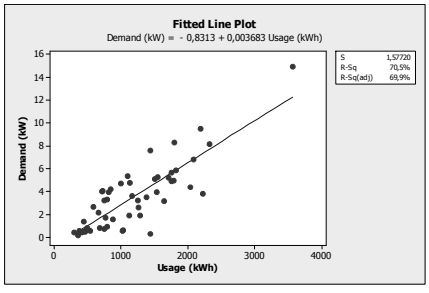
- Investigar relacionamento entre demanda diária de pico (*kW*) e consumo total de energia elétrica (*kWh*)
- Problema: enquanto o consumidor paga pelo uso da energia, o sistema de geração deve ser grande o suficiente para atender a máxima demanda
- Planilha: *energia_eletrica*

Saída – Dados Originais

Regression Analysis: Demand (kW) versus Usage (kWh)

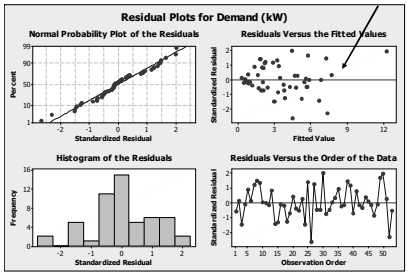
```
The regression equation is
Demand (kW) = - 0,8313 + 0,003683 Usage (kWh)

S = 1,57720   R-Sq = 70,5%   R-Sq(adj) = 69,9%

Analysis of Variance

Source       DF       SS        MS       F      P
Regression    1  302,633  302,633  121,66  0,000
Error        51  126,866    2,488
Total        52  429,499
```

Regressão Ajustada – Dados Originais



Gráficos de Resíduos – Dados Originais

Heterocedasticidade



Saída – Dados Transformados

Regression Analysis: SQR(Demand)* versus Usage (kWh)

The regression equation is
SQR(Demand)* = 0,5822 + 0,000953 Usage (kWh)

S = 0,464044 R-Sq = 64,8% R-Sq(adj) = 64,2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	20,2585	20,2585	94,08	0,000
Error	51	10,9822	0,2153		
Total	52	31,2407			

Regressão Ajustada – Dados Transformados

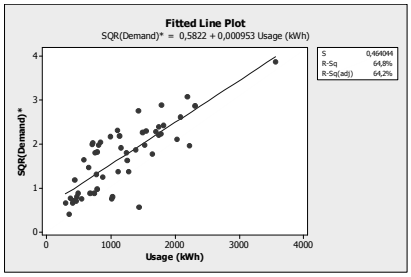
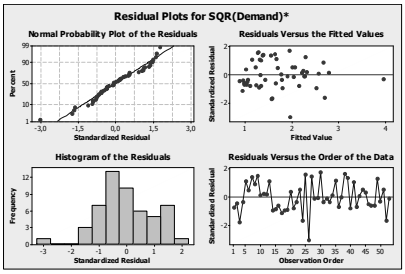
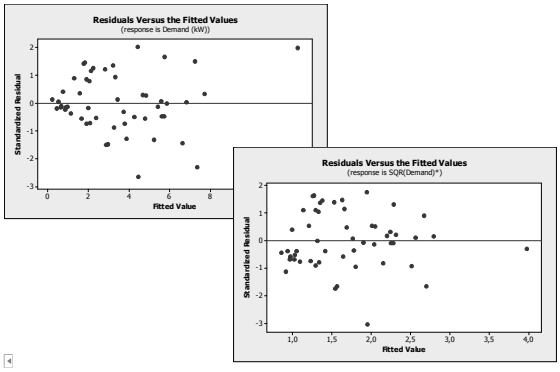


Gráfico de Resíduos – Dados Transformados



Comparação Gráfico de Resíduos



Transformações Estabilizadoras

Relação de s^2 a E(Y)	Transformação
$s^2 \propto constante$	$y' = y$
$s^2 \propto E(Y)$	$y' = \sqrt{y}$
$s^2 \propto E(Y)[1 - E(Y)]$	$y' = arcsen\left(\sqrt{y}\right)$
$s^2 \propto [E(Y)]^2$	$y' = \log y$
$s^2 \propto [E(Y)]^3$	$y' = \left(\sqrt{y}\right)^{-1}$
$s^2 \propto [E(Y)]^4$	$y' = y^{-1}$

Comentários

- A força da transformação depende da quantidade de curvatura que ela induz
  - √ Pode ir de moderada ($v x$), `a forte ($1/x$)
- Geralmente transformação moderada sobre faixa de valores estreita produz pouco efeito

$$\frac{y_{\max}}{y_{\min}} < 2 \; ou \; 3$$

- Por outro lado, transformação forte sobre faixa de valores estreito trará efeito dramático na análise
- Em geral, a transformaçãofornecerá estimativas mais precisas dos parâmetros do modelo
- Intervalos de confiança e de predição podem ser covertidos diretamente de uma métrica a outra, mas não se assegura que eles sejam os menores possíveis.

Omissão de Variáveis

Referências

Bibliografia Recomendada

- Hill, R. C., Griffiths, W. E. e Judge, G. (Saraiva) *Econometria*
- Gujarati, D. N. (Pearson) *Econometria Básica*
- Maddala, G. S. (LTC) *Introdução à econometria*