



Módulo de Regressão e Séries Temporais

Mônica Barros, D.Sc.

Maio de 2007

monica@mbarros.com

1



Quem sou eu?

□ Mônica Barros

- Doutora em Séries Temporais – PUC-Rio
- Mestre em Estatística – University of Texas at Austin, EUA
- Bacharel em Matemática – University of Washington, Seattle, EUA
- Professora da PUC-Rio (Depto. De Eng. Elétrica)
- E-mails: monica@ele.puc-rio.br, monica@mbarros.com
- Home page: <http://www.mbarros.com>



monica@mbarros.com

2

Programa do Curso



□ Uma visão gerencial da Estatística e Séries Temporais

- Conceitos
- Parâmetros
- Modelo Linear
- Modelos AR(1), AR(2), AR(p) e PAR(p)
- Modelos de Séries Temporais
- Estatísticas de “erros”

monica@mbarros.com

3

Programa do Curso



□ Regressão Linear Simples

- Modelo
- Estimação (Mínimos Quadrados Ordinários e Máxima Verossimilhança)
- Análise dos resíduos
- Estimação da variância dos resíduos
- Inferências para os parâmetros
- Variáveis dummy
- Exercícios

monica@mbarros.com

4

Programa do Curso



□ Regressão Linear Múltipla

- Modelo
- Estimação (Mínimos Quadrados Ordinários e Máxima Verossimilhança)
- Análise dos resíduos
- Estimação da variância dos resíduos
- Inferências para os parâmetros
- Colinearidade
- Exercícios

Programa do Curso



□ Introdução às séries temporais

- Objetivos
- Escopo
- Métodos de estimação
- Limitações
- “Overview” dos principais métodos – quando e onde usá-los

Programa do Curso



□ Métodos de Amortecimento Exponencial

- Modelos de médias móveis
- Modelo de Brown
- Modelo de Holt-Winters
- Estimação dos parâmetros
- Estatísticas de ajustes e análise dos resíduos
- Exercícios

Programa do Curso



□ Modelagem ARIMA de Box & Jenkins sazonal e não sazonal

- Função de autocorrelação e autocorrelação parcial
- Modelo
- Identificação de (p, q, d, P, Q, D)
- Estimação
- Estatísticas de ajuste e análise dos resíduos
- Exercícios

Programa do Curso



- ❑ **Modelos de Regressão Dinâmica**
 - ❑ Função de correlação cruzada
 - ❑ Modelo
 - ❑ Identificação/estimação
 - ❑ Estatísticas de ajustes e análise dos resíduos
 - ❑ Exercícios

Nota – Instalação das Ferramentas de Análise do Excel



- ❑ Muitas das técnicas descritas aqui requerem a prévia instalação do suplemento ("add-in") "Ferramentas de Análise" do Excel. O procedimento de instalação é descrito a seguir:
- ❑ No menu **Ferramentas**, selecione "**Suplementos**" e na caixa de diálogo que será aberta marque a opção "**Ferramentas de análise**". Se esta opção não estiver presente, clique "procurar" para encontrar o arquivo correspondente (em geral chamado **Analys32.xll**) ou rode novamente o "set-up" do MS-Office.

Forecast Pro



- ❑ Neste curso também será necessário (nos módulos específicos de séries temporais), o uso do software Forecast Pro XE, já utilizado pelo ONS.
- ❑ Este software permite o ajuste de modelos de amortecimento exponencial, ARIMA e SARIMA e modelos de regressão dinâmica.

Série Temporal



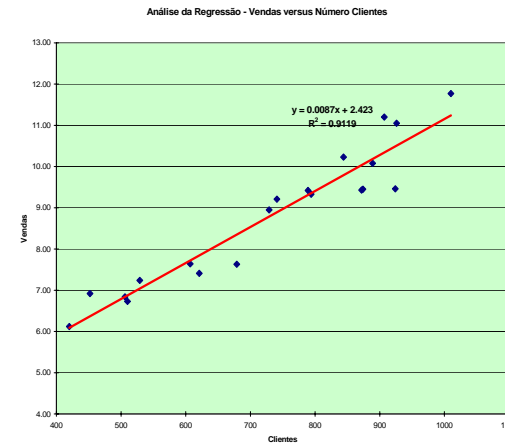
- ❑ Conjunto de observações ordenadas no tempo.
- ❑ Suponha que temos uma série temporal de cargas elétricas. Então:
- ❑ Y_1, Y_2, \dots, Y_n – onde y_t é a carga no instante t .
- ❑ Existe dependência entre as cargas de diversos instantes.
- ❑ Isto nos permite usar cargas passadas para prever a carga futura.

Modelo



- ❑ É uma representação simplificada da realidade.
- ❑ Por exemplo, você observa duas variáveis X e Y e faz um gráfico dos n pontos $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. O gráfico tem a "cara":

Modelo



Modelo Linear



- ❑ Nesse exemplo a relação entre X e Y é, como o gráfico revela, aproximadamente linear.
- ❑ **Nosso objetivo: achar a reta que relaciona Y e X. Qual reta? Como encontrá-la?**
- ❑ **O método tradicional é chamado de mínimos quadrados.**

Modelo Linear



- ❑ **Método de Mínimos Quadrados**
 - ❑ Como funciona? O objetivo é minimizar a soma dos quadrados dos "erros" (resíduos).
 - ❑ Que "erros" são esses?
 - ❑ Para cada ponto, o resíduo é o valor real menos o valor previsto pela reta.
 - ❑ Pode-se notar que o resíduo vai depender de qual reta foi ajustada.

Modelo Linear



- Por exemplo, se ajustamos a reta $y = 3 + 2x$, se o valor de x_i é 2, então y_i ajustado pela reta é igual a 7. Se agora a reta é $y = 4 + 2x$, para $x_i = 2$ temos o y_i ajustado pela reta igual a 8.
- Imagine que o valor real de y_i seja 8.5 quando x_i é 2. Então, o resíduo gerado pela reta $3 + 2x$ é $8.5 - 7 = 1.5$ e pela reta $4 + 2x$ é $8.5 - 8 = 0.5$.

Modelo Linear



- Então, no que diz respeito a esse ponto ($x_i = 2$), a reta $4 + 2x$ foi melhor que a reta $3 + 2x$ pois gerou um resíduo menor.
- Em geral, dados n pares de pontos $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, você pode ajustar uma reta $y = b_0 + b_1 \cdot x$ a todos os pares.
- **Note que os coeficientes b_0 e b_1 são os mesmos para todos os pares de pontos.**

Modelo Linear



- O resíduo do i -ésimo par é:

$$\hat{e}_i = y_i - \hat{y}_i = \text{real}_i - \text{previsto}_i$$

- O método de mínimos quadrados procura obter a reta (ou seja, os coeficientes b_0 e b_1) tais que a soma dos quadrados dos resíduos seja minimizada.

Modelo Linear



- Método de Mínimos Quadrados
- Ache b_0 e b_1 tais que a expressão:
$$\sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$
- seja minimizada.
- Esta soma acima é chamada de Soma dos Quadrados dos Resíduos (às vezes, Soma dos Quadrados dos Erros).

Modelo Linear



- ❑ Erro versus Resíduo
- ❑ Erro = variável aleatória não observável, com média zero e variância constante σ^2 .
- ❑ Resíduo = estimador ("chute") para o erro. É observável, calculado como valor real – valor previsto pelo modelo.
- ❑ Isso causa alguma confusão, pois às vezes a gente chama de "erro" o que é, na verdade, o "resíduo".

Modelo Linear



- ❑ O modelo linear é:
- ❑ $y_i = b_0 + b_1 x_i + e_i$ para $i = 1, 2, \dots, n$
- ❑ Onde e_i é o erro, uma variável com média 0 e variância constante σ^2 .
- ❑ Os parâmetros do modelo são b_0 , b_1 e σ^2 , são constantes desconhecidas que devem ser estimadas.

Modelo AR(1)



- ❑ Parâmetro = tudo que precisa ser estimado para especificar completamente o modelo.
- ❑ Modelo AR(1)
- ❑ AR(1) = autoregressivo de ordem 1
- ❑ $y_t = \phi y_{t-1} + e_t$
 - ❑ y_t = carga no mês t
 - ❑ y_{t-1} = carga no mês t-1
 - ❑ e_t = erro no mês t

Modelo AR(1)



- ❑ Parâmetros: ϕ, σ^2 (variância do erro)
- ❑ Caso Particular: Passeio Aleatório (Random Walk)
- ❑ AR(1) com $\phi = 1 \Rightarrow y_t = y_{t-1} + e_t$
- ❑ Uso: mercado de ações
- ❑ Na média, preço da ação hoje = preço da ação ontem (não ajuda muito para ganhar dinheiro!!)

Random Walk



- ❑ O gráfico a seguir mostra um exemplo de um passeio aleatório em que $y_0 = 100$ e os erros são Normais com média zero e desvio padrão 2.
- ❑ Foram geradas 500 observações.
- ❑ Você poderia confundir o gráfico com a da evolução do preço de uma ação no tempo, não é mesmo?

Random Walk



Gráfico de uma "Random Walk"



Modelo AR(1)



- ❑ Exemplo – efeito de σ^2
- ❑ $y_t = 1,05 \cdot y_{t-1} + e_t$
- ❑ Carga do mês $t = 105\%$ da carga do mês $t-1$ + erro.
- ❑ Suponha que $y_{t-1} = 100$.
 - a) $\text{VAR}(e_t) = 1$
 - b) $\text{VAR}(e_t) = 100$
- ❑ Por exemplo, se fosse possível observar o erro poderíamos encontrar algo como:

Modelo AR(1)



- a) $e_t = 1,05$ – valor pequeno pois foi “tirado” de uma distribuição de prob. com média 0 e variância 1.
 - b) $e_t = 10,32$ – valor grande porque a distribuição de prob. tem variância “grande” (100).
- ❑ Resultados:
 - a) $y_t = 105 + 1,05 = 106,05$
 - b) $y_t = 105 + 10,32 = 115,32$

Modelo AR(2)



- $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t$
- **Onde:**
 - e_t = erro no instante t com média 0 e variância σ^2
 - y_t = carga no instante t
 - Y_{t-1} = carga no instante t-1
 - Y_{t-2} = carga no instante t-2
 - Parâmetros: ϕ_1 , ϕ_2 e σ^2 .

Modelo AR(p)



- **Modelo Autoregressivo de ordem p**
- $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + e_t$
- **A carga deste mês depende das cargas dos p últimos meses.**
- **Existem p + 1 parâmetros a serem estimados: os ϕ 's e a variância do erro.**

Modelo PAR(p)



- **PAR(p) = Modelo Autoregressivo Periódico de ordem p.**
- **Contexto: Newave**
- **Modelos para previsão de vazões.**
- **Um modelo AR diferente para cada mês. Restrição do Newave - p máximo = 6 meses. (p = ordem do modelo).**

Modelo de Séries Temporais



- **Resíduo no instante t:**
- $\hat{e}_t = \text{real} - \text{previsto no instante t}$
- **Por exemplo, para dez/06, o resíduo é o valor real de dezembro de 2006 menos o valor previsto para o mesmo mês. Só vai poder ser calculado quando o valor real de 12/2006 estiver disponível!**

Modelo de Séries Temporais



- ❑ Mas posso olhar para os resíduos “in sample” (dentro da amostra) no período em que o modelo foi ajustado.
- ❑ Por exemplo, suponha que o modelo foi ajustado no período jan/03 a abr/06. Neste período todo posso fazer a comparação entre o valor real e o ajustado pelo modelo (“fitted value”), calculando o resíduo a cada instante.
- ❑ Um exemplo está na próxima figura.

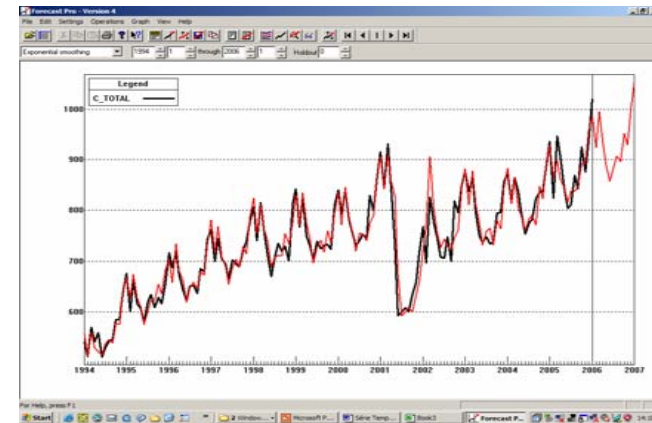
monica@mbarros.com

33

Modelo de Séries Temporais



- ❑ O gráfico a seguir mostra uma série de carga real (em preto) e ajustada por um modelo (em vermelho).

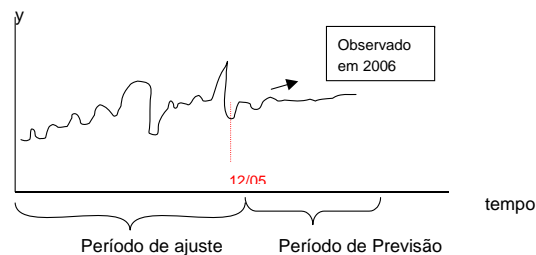


34

Modelo de Séries Temporais



- ❑ Período de ajuste (in sample)
- ❑ Período de previsão (out of sample)
- ❑ No gráfico anterior, o período de ajuste vai de janeiro de 1994 a dezembro de 2005, e o período de previsão é o ano de 2006.



monica@mbarros.com

35

Modelo de Séries Temporais



- ❑ Você ajustaria o seguinte modelo:
 - ❑ In sample = 12 meses,
 - ❑ Out of sample = 24 meses?
- ❑ Não! E o contrário? Sim, desde que a série não fosse muito mal comportada. **Em resumo – previsão de séries temporais não é futurologia! Você precisa de dados passados para fazer previsão!**

monica@mbarros.com

36

Estatísticas de “Erros”



□ Estatística de “erros” (resíduos)

Mês	Real	Previsto	Resíduo
1	y1	^y1	ê1 = y1 - ^y1
2	y2	^y2	ê2 = y2 - ^y2
...
n	yn	^yn	ên = yn - ^yn

- **MAPE = erro absoluto médio percentual**
- **“erro” percentual no instante i:**
 - $(y_i - \hat{y}_i) / y_i$
- **“erro” percentual absoluto no instante i:**
 - $|y_i - \hat{y}_i| / y_i$

Estatísticas de “Erros”



□ “Erro” absoluto percentual absoluto no instante i:

- $|y_i - \hat{y}_i| / y_i = |(\text{real} - \text{previsto}) / \text{real}|$

□ MAPE = média destes erros absolutos percentuais

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| = \frac{1}{n} \sum_{i=1}^n \left| \frac{\text{real}_i - \text{previsto}_i}{\text{real}_i} \right|$$

- **Vantagem: fácil de entender - a escala é %**
- **Desvantagem: Se o valor real é pequeno, qualquer discrepância na previsão faz o MAPE “explodir”.**

Estatísticas de “Erros”



□ MAD = mean absolute deviation = desvio absoluto médio

$$MAD = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| = \frac{1}{n} \sum_{i=1}^n |\text{real}_i - \text{previsto}_i|$$

- **O MAD está nas mesmas unidades que a sua série.**
- **Por que o módulo?**
- **Para evitar soma = 0 quando um erro é grande e positivo e outro é grande e negativo!**

Estatísticas de “Erros”



□ EQM = erro quadrático médio

$$EQM = \frac{1}{n} \sum_{i=1}^n \hat{e}_i^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n (\text{real}_i - \text{previsto}_i)^2$$

□ RMSE = root mean squared error = raiz do EQM

$$RMSE = \sqrt{EQM} = \sqrt{\frac{1}{n} \sum_{i=1}^n \hat{e}_i^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\text{real}_i - \text{previsto}_i)^2}$$

- **Vantagem sobre EQM => mesma escala que os dados.**

Estatísticas de “Erros”



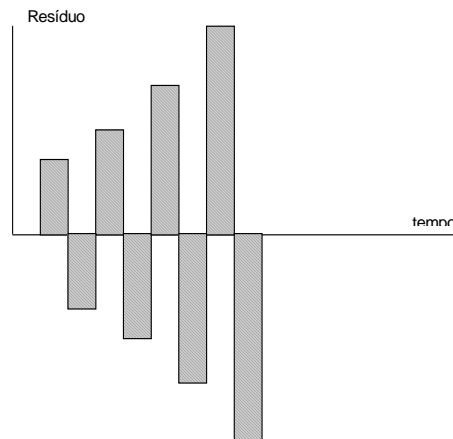
- ❑ “Erro puro”: $\hat{e}_i = y_i - \hat{y}_i$
- ❑ **Vantagem: mesma escala que os dados.**
- ❑ **Desvantagem: não dá para sair somando! Se você fizer isso corre o risco de achar que seu modelo é bom quando não é, pois erros de sinais opostos poderão se cancelar.**

Estatísticas de “Erros”



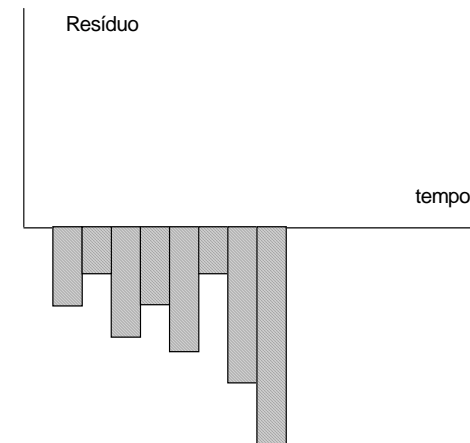
- ❑ **O que fazer com tudo isso?**
- ❑ **O que é desejável? Se o modelo é bom, resíduo não tem padrão, não tem estrutura.**
- ❑ **Faça gráficos dos resíduos “puros” ou MAPE, ou MAD ou EQM, ou RMSE.**
- ❑ **Que tipo de gráficos?**
 - ❑ Por exemplo, gráfico dos “erros” ao longo do tempo. Serve para responder como o seu modelo se comportou no “in sample” e verificar a existência de possíveis padrões.

Estatísticas de “Erros”



- **Isto é tudo que eu não quero.**
- **Por que?**
- **Resíduo com um padrão claro (positivo, negativo, positivo, negativo, ...).**

Estatísticas de “Erros”

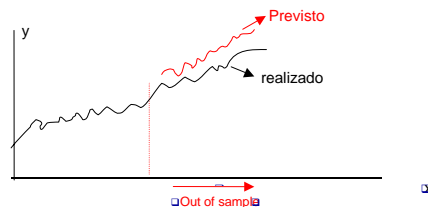


Também não quero um padrão desses – todos os resíduos são negativos, indicando que a minha previsão esteve sempre ACIMA do real

Estatísticas de “Erros”



- Gráfico dos “erros” versus o horizonte de previsão.
 - Serve para responder se as previsões se deterioram rápido ou não.



monica@mbarros.com

45

Estatísticas de “Erros”



- Gráfico da Autocorrelação dos erros
- Correlação entre os erros em diversos instantes
- ACF = autocorrelation function
- Gráfico da ACF versus lag (defasagem) é também chamado de correlograma.

O que você não quer?
Que exista correlação entre resíduos em instantes diferentes.

monica@mbarros.com

46

Estatísticas de “Erros”



- Por que você não quer que o gráfico da ACF dos resíduos mostre padrões?
- **Porque se existe dependência temporal na série, toda esta dependência teria que ser, idealmente, capturada pelo modelo, não deveria ter sobrado nada nos resíduos!**

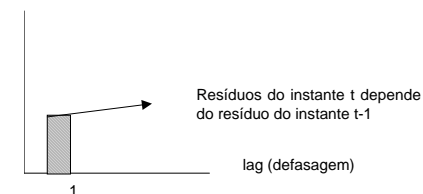
monica@mbarros.com

47

Estatísticas de “Erros”



- Casos comuns...



- **Atenção: A autocorrelação é sempre um número entre -1 e +1**
 - No gráfico anterior:
 - Resíduo grande e positivo no instante t leva a um resíduo grande e positivo no instante t+1

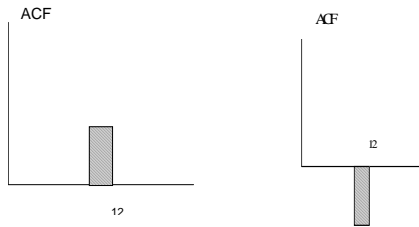
monica@mbarros.com

48

Estatísticas de “Erros”



□ Outros Padrões:



- Resíduos nos instantes t e $t-12$ têm dependência.

Estatísticas de “Erros”



- Em resumo...
- Se o modelo é bom o ruído é branco, o resíduo não tem estrutura.
- Se houver padrões no resíduo fique esperto!

Regressão Linear Simples



Regressão Linear



- Modelos de regressão linear relacionam uma variável dependente ou variável de resposta, Y , a uma ou mais variáveis explicativas (também chamadas covariáveis ou variáveis independentes), X .
- O modelo é **linear nos parâmetros** que relacionam Y aos X 's.
- A **estimação** destes modelos é geralmente feita por **mínimos quadrados ordinários**, e os estimadores obtidos por este algoritmo são ótimos sob certas condições, dadas pelo teorema de Gauss e Markov.

Objetivos dos Modelos de Regressão Linear

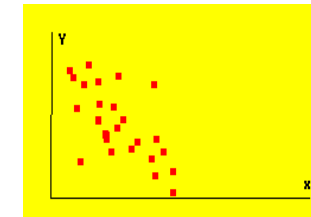
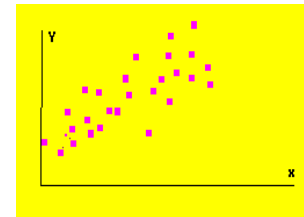


- Estudar a relação entre variáveis, para se testar causalidade (linear) entre as variáveis (ECONOMETRIA).
- Possibilitar análises de cenários ("What if analysis")
- Permitir eventualmente a previsão da variável dependente. (SÉRIES TEMPORAIS, REGRESSÃO DINÂMICA)

Regressão Linear Simples



- Dependência Linear entre X e Y
 - O diagrama de dispersão tem o seguinte aspecto:



Regressão Linear Simples



- Só uma variável explicativa
 - Estão disponíveis n pares de observações (x_i, y_i)
 - ε é um erro aleatório com média zero e variância constante σ^2 . Em muitas situações supomos que o erro é Normal, o que nos permite obter intervalos de confiança e realizar testes de hipóteses.
- Parâmetros desconhecidos: β_0, β_1 e σ^2
- A equação que a relação entre y e x é chamada de **modelo de regressão**:

$$Y = \beta_0 + \beta_1 \cdot x + \varepsilon$$

Regressão Linear Simples



- No caso de **regressão linear simples**, temos:
$$y = \beta_0 + \beta_1 x + \varepsilon$$
 - ε é uma variável aleatória chamada de **erro da regressão** e representa a variação de y que não pode ser explicada por sua relação linear com x – **por definição**, seu **valor esperado é zero**
 - Portanto, o **valor esperado ou valor médio de y está relacionado com x através de** $E(y) = \beta_0 + \beta_1 x$

Regressão Linear Simples



- **Hipóteses sobre os erros da regressão**
 - Considere o modelo $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ para $i = 1, 2, \dots, n$ (ou seja, temos n pares de observações (x_i, y_i)).
 - Os erros ε_i satisfazem as seguintes hipóteses:
 - Média zero $\Rightarrow E(\varepsilon_i) = 0$ para todo i
 - Variância constante $\Rightarrow \text{VAR}(\varepsilon_i) = \sigma^2$ para todo i (hipótese de homocedasticidade)
 - Os ε_i são Normais, o que nos permite desenvolver intervalos de confiança e testes de hipóteses.
 - Os ε_i são independentes (ou, pelo menos, descorrelatados).

Regressão Linear Simples



- **O gráfico da relação entre x e y é uma reta (reta de regressão)**
 - β_0 é chamado de intercepto (ou coef. Linear)
 - β_1 é chamado de inclinação (ou coef. Angular da reta)
- **Se os parâmetros β_0 e β_1 fossem conhecidos, poderíamos utilizar o modelo de regressão linear simples para determinar o valor esperado de y para um dado valor de x .**

Regressão Linear Simples



- **Na prática, esses parâmetros não são conhecidos e precisam ser estimados.**
- **Utilizamos observações emparelhadas de x e y para determinar os estimadores de β_0 e β_1 - b_0 e b_1 respectivamente, obtendo a seguinte equação de regressão.**

$$\hat{y}_i = b_0 + b_1 x_i \quad \text{para } i = 1, 2, \dots, n$$

- **Para determinar b_0 e b_1 utilizamos o método dos mínimos quadrados.**

Regressão Linear Simples



- **Nesse método, os valores de b_0 e b_1 são tais que a soma dos quadrados das diferenças entre os valores de y observados (y_i) e seus respectivos valores estimados pela equação de regressão (\hat{y}_i) é mínima:**

$$\min \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- **Ou seja, o método de mínimos quadrados minimiza a soma do quadrado dos resíduos.**

Regressão Linear Simples



- Ou seja, os estimadores b_0 e b_1 são obtidos de tal forma que:

$$SSE = \sum_{i=1}^n \hat{\epsilon}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

- Seja minimizada.
- E como fazer isso?
- **Note que a soma do quadrado dos resíduos (SSE) é função de ambos b_0 e b_1 .** Para garantir que esta função está sendo minimizada, precisamos encontrar seus pontos críticos (aqueles em que as primeiras derivadas em relação a b_0 e b_1 são zero) e verificar o sinal das segundas derivadas, que devem ser positivos.

Regressão Linear Simples



- Então, para encontrar os estimadores b_0 e b_1 basta fazer:

$$\frac{\partial SSE}{\partial b_0} = 0 \quad \text{e} \quad \frac{\partial SSE}{\partial b_1} = 0$$

- E verificar o sinal das 2as. derivadas nos pontos críticos.
- Os valores de b_0 e b_1 são dados por:

$$b_1 = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} \quad b_0 = \bar{y} - b_1 \bar{x}$$

Regressão Linear Simples



- Da equação anterior à direita notamos que a reta ajustada por mínimos quadrados passa pelo ponto (\bar{x}, \bar{y}) , isto é, pelo ponto médio dentre todos os n pontos (x_i, y_i) usados para estimar a reta de regressão.

Regressão Linear Simples



- Sejam:

$$SXX = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = \sum x_i^2 - n \bar{x}^2$$

$$SYY = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = \sum y_i^2 - n \bar{y}^2 = SST, \text{ a soma de quadrados total}$$

$$SXY = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = \sum x_i y_i - n \bar{x} \bar{y}$$

- Então podemos escrever:

$$b_1 = \frac{SXY}{SXX} \quad \text{e} \quad b_0 = \bar{y} - b_1 \bar{x}$$

Note que ambos b_0 e b_1 são variáveis aleatórias, pois são funções dos y 's, que são v.a.

Regressão Linear Simples



□ **Propriedades dos Estimadores MQO (mínimos quadrados ordinários)**

- **b_0 e b_1 são não tendenciosos, isto é:**

$$E(b_0) = \beta_0 \quad \text{e} \quad E(b_1) = \beta_1$$

- **As variâncias de b_0 e b_1 são:**

$$VAR(b_1) = \frac{\sigma^2}{SXX} \quad \text{e} \quad VAR(b_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX} \right)$$

onde σ^2 , como já vimos, é a variância do erro.

Regressão Linear Simples



□ **Propriedades dos Estimadores MQO**

- **A covariância entre b_0 e b_1 é:**

$$COV(b_0, b_1) = \frac{-\sigma^2(\bar{x})}{SXX}$$

- **Dos resultados anteriores nota-se que as variâncias e covariâncias dos estimadores MQO dependem de σ^2 (a variância do erro), que é uma quantidade desconhecida. Como estimá-la?**

$$s^2 = \frac{SSE}{n-2}$$

Regressão Linear Simples



- **O modelo de regressão desenvolvido aproxima a relação linear entre as variáveis x e y**

- **Uma pergunta importante: quão bem o modelo de regressão representa essa relação linear?**

- **O coeficiente de determinação (R^2) da regressão nos dá uma medida do ajuste do modelo de regressão aos dados utilizados e será definido daqui a pouco.**

Regressão Linear Simples



- **O i -ésimo resíduo é a diferença entre o valor observado y_i e o valor estimado pela equação de regressão, e representa o erro obtido ao estimarmos y_i .**

- **Ou seja: $\hat{e}_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$**

- **O método de mínimos quadrados encontra os coeficientes b_0 e b_1 que minimizam a soma dos quadrados desses resíduos.**

Regressão Linear Simples



- Agora suponha que queremos estimar um modelo constante, ou seja, $y_i = c + \text{erro}$
- Sem o conhecimento de nenhuma variável explicativa, ou seja, sem o modelo de regressão, a nossa melhor estimativa seria o valor médio das observações de y , \bar{y}
- Para a i -ésima observação de y , y_i , a diferença $y_i - \bar{y}$ nos fornece uma medida do erro cometido ao estimarmos y_i a partir de \bar{y}

Regressão Linear Simples



- Definimos como soma total dos quadrados, **SST** (do inglês "total sum of squares"), a soma do quadrado dessas diferenças:

$$SST = SYY = \sum (y_i - \bar{y})^2$$

- **A soma dos quadrados devidos à regressão, SSReg** (do inglês "sum of squares due to regression") é dada por:

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

Regressão Linear Simples



- Ou seja, a SST (Soma dos Quadrados Total) é a soma dos quadrado dos resíduos quando o modelo estimado é um modelo constante, isto é, quando ignoramos a relação entre x e y .
- Se a relação entre x e y for importante, o que se espera?
- Que a soma do quadrado dos resíduos seja muito menor que SST.

Regressão Linear Simples



- Agora suponha que definimos um modelo de regressão.
- O poder de explicação da relação linear entre x e y a partir do modelo de regressão, para cada y_i observado, em relação ao caso anterior (conhecimento apenas dos y_i e utilização de \bar{y} como estimativa) pode ser medido pela diferença entre os valores previstos pela regressão e \bar{y} .

Regressão Linear Simples



- Pode-se provar que SSE, SSR e SST estão relacionados da seguinte forma:

$$SST = SS\text{ Reg} + SSE$$

- O valor de SST é o mesmo independente do grau de ajuste do modelo de regressão, pois só depende dos y 's observados e da sua média.
- Devemos esperar que, quanto melhor o ajuste do modelo de regressão aos dados utilizados, menor a razão SSE/SST e maior a razão SSReg/SST.

Regressão Linear Simples



- No limite, com um ajuste perfeito (a reta de regressão passa exatamente sobre todos os pontos da amostra), teremos SSE = 0 e SST = SSReg

- O coeficiente de determinação (R^2) é uma medida da qualidade do ajuste de uma regressão, e é definido como:

$$R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} = \frac{SS\text{ Reg}}{SST} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

Regressão Linear Simples



- Quanto mais próximo de 1, melhor o ajuste do modelo de regressão aos dados da amostra.
- O coeficiente de determinação também está relacionado ao coeficiente de correlação entre x e y :

$$R^2 = \frac{SS\text{ Reg}}{SST} = \frac{(SXY)^2}{(SXX)(SYY)} = (r_{xy})^2$$

- Note que o R^2 está definido entre 0 e 1, e quanto mais próximo de 1, melhor o ajuste da regressão.

Teste de significância dos parâmetros da regressão



- No modelo de regressão linear simples, a relação de dependência entre y e x é expressa pelo parâmetro β_1 que, por sua vez, é estimado a partir de b_1 .
- E como em qualquer estimação, cometemos um erro ao utilizarmos b_1 (o estimador) ao invés de β_1 (o parâmetro real, mas desconhecido).

Teste de significância dos parâmetros da regressão



- Quando os erros são iid Normais, os estimadores dos parâmetros e valores ajustados são Normais.
- Isso acontece pois eles são apenas combinações lineares dos y_i 's (e portanto, combinações lineares dos e_i 's).
- Assim, testes e intervalos de confiança podem ser construídos baseados na distribuição t de Student.

Teste de significância dos parâmetros da regressão



- O teste de significância de β_1 é feito a partir do intervalo de confiança de β_1 para um nível de confiança especificado
- Se o intervalo de confiança de β_1 contiver o valor 0, isso indica que, para o nível de significância escolhido, nós não podemos descartar a hipótese de β_1 (o parâmetro desconhecido) ser igual a zero.

Teste de significância dos parâmetros da regressão



- Do contrário, se o intervalo de confiança não incluir zero, podemos concluir que, para o nível de significância escolhido, β_1 é diferente de zero.
- Note que testar a hipótese $\beta_1 = 0$ é equivalente a testar as seguintes hipóteses:
 - H_0 : o modelo é constante $\Rightarrow Y_i = \beta_0 + \varepsilon_i$
 - H_1 : o modelo é linear $\Rightarrow Y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$

Teste de significância dos parâmetros da regressão



- Teste da hipótese $\beta_1 = 0$ (versus a alternativa $\beta_1 \neq 0$)

$$t = \frac{b_1}{dp(b_1)} = \frac{b_1}{\sqrt{\frac{s^2}{SXX}}} = \frac{b_1}{s_{b_1}}$$

- Esta estatística tem distribuição t com n-2 graus de liberdade.
- O IC (1- α)% para β_1 é dado por:

$$b_1 \pm t_{\frac{\alpha}{2}} s_{b_1}$$

O valor de $t_{\alpha/2}$ é baseado em uma distribuição t de Student com n - 2 graus de liberdade

Utilizando o Excel



- ❑ O Excel possui uma ferramenta de análise de regressão
- ❑ No menu Ferramentas selecione Análise de Dados
 - ❑ Caso esta opção não esteja disponível, selecione Add-Ins (Suplementos) e marque a caixa Ferramentas de Análise
- ❑ Em **Análise de Dados**, selecione **Regressão**
- ❑ Forneça as informações necessárias
 - ❑ Intervalo dos valores de x
 - ❑ Intervalo dos valores de y
 - ❑ Nível de significância
 - ❑ Intervalo de saída (local na planilha onde o resultado será colocado)

monica@mbarros.com

81

Estudo de Caso – Regressão Linear Simples



- ❑ A planilha BigMac2003.xls contém preços em diversas cidades do mundo em 2003 coletados pelo banco suíço UBS.
- ❑ As variáveis da planilha são:
 - ❑ City – nome da cidade
 - ❑ BigMac – minutos de trabalho necessários para comprar um Big Mac
 - ❑ Bread - minutos de trabalho necessários para comprar 1 kg de pão
 - ❑ Rice - minutos de trabalho necessários para comprar 1 kg de arroz
 - ❑ Foodindex = índice do preço da comida (Zurique = 100)
 - ❑ Bus = custo em US\$ de uma passagem de ônibus num trecho de 10 km

monica@mbarros.com

82

Estudo de Caso – Regressão Linear Simples



- ❑ Apt = custo do aluguel de um apartamento de 3 cômodos em US\$
- ❑ TeachGI = Renda Bruta de um professor primário em milhares de US\$
- ❑ TeachNI = Renda Líquida de um professor primário em milhares de US\$
- ❑ Taxrate = alíquota de imposto paga por um professor primário
- ❑ Teachhours = número de horas trabalhadas por semana para um professor primário

- ❑ **É claro que existem inúmeras relações importantes entre as diversas variáveis nesta base de dados – a idéia geral é ver como o custo de vida (e a renda) se comporta através das grandes cidades no mundo.**

monica@mbarros.com

83

Estudo de Caso – Regressão Linear Simples

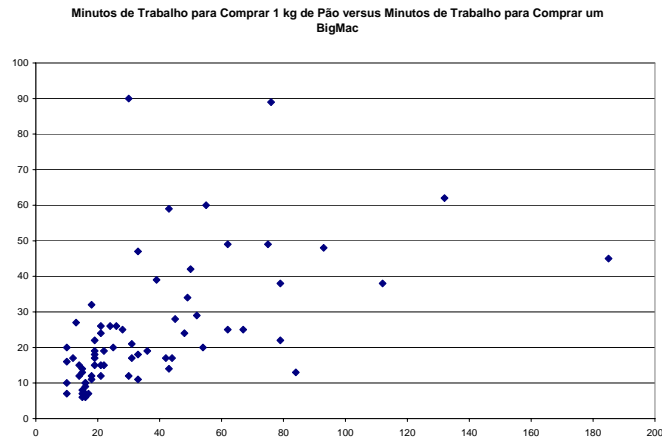


- ❑ Mas, vamos começar com algo simples.
- ❑ **É de se esperar que o tempo de trabalho necessário para comprar dois tipos de alimento seja positivamente correlacionado.**
- ❑ Então, que tal examinarmos o gráfico de "Bread" em "BigMac"?

monica@mbarros.com

84

Estudo de Caso – Regressão Linear Simples



monica@mbarros.com

85

Estudo de Caso – Regressão Linear Simples

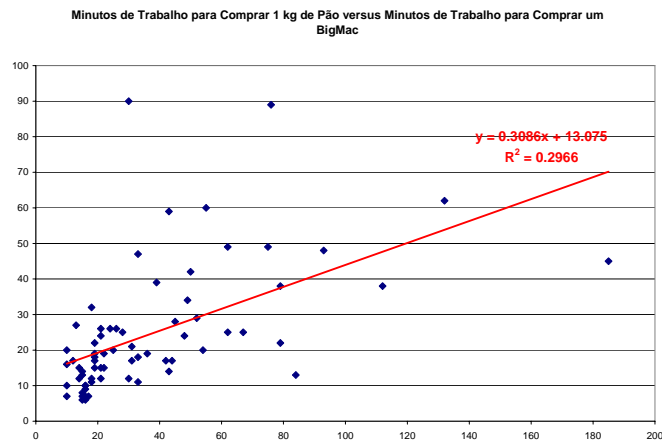


- Podemos produzir a equação de regressão sem nenhuma demora no próprio Excel, apenas clicando na opção Gráfico > Adicionar Linha de Tendência.
- Existem vários tipos de linha de tendência, escolha a "linear", e as opções de mostrar a equação e o R^2 no gráfico. O resultado está na próxima figura.

monica@mbarros.com

86

Estudo de Caso – Regressão Linear Simples



monica@mbarros.com

87

Estudo de Caso – Regressão Linear Simples



- O nosso modelo de regressão não foi lá "grande coisa"...
- O R^2 foi apenas 29%, mas disso a gente já desconfiava pelo gráfico, que não era "muito" linear.
- Mas, o "output" foi bastante limitado, apenas a equação da reta e o R^2 .
- Podemos ter uma análise bem mais completa através da ferramenta de Análise de dados do Excel.
- O resultado está a seguir.

monica@mbarros.com

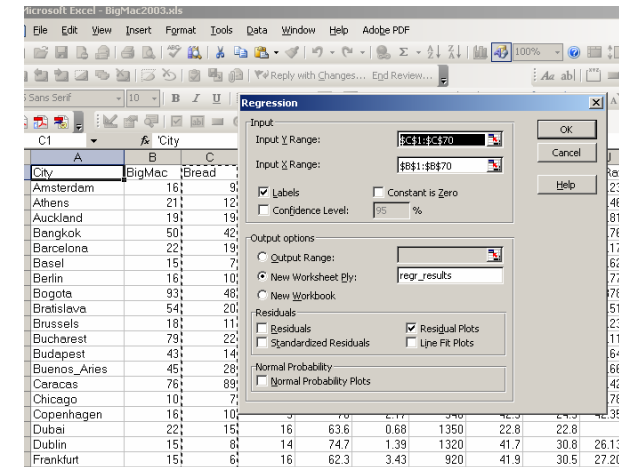
88

Estudo de Caso – Regressão Linear Simples



- ❑ Como fazer?
- ❑ Ferramentas > Análise de Dados > >Regressão
- ❑ A caixa de diálogo usada neste exemplo (para o Excel em inglês está no próximo slide)
- ❑ Note que:
 - ❑ “Labels” (Rótulos) indica que a primeira linha contém o nome da variável
 - ❑ Os resultados foram mandados para uma nova pasta (chamada regr_results neste caso)

Estudo de Caso – Regressão Linear Simples



Estudo de Caso – Regressão Linear Simples



SUMMARY OUTPUT (REGRESSÃO DE BREAD EM BIGMAC)

Regression Statistics	
Multiple R	0.54458571 << é o coef. de correlação
R Square	0.296573595
Adjusted R Square	0.286074694
Standard Error	15.04612363
Observations	69

ANOVA					
	df	SS	MS	F	Significance F
Regression	1	6394.960561	6394.961	28.24806	1.3158E-06
Residual	67	15167.85103	226.3858		
Total	68	21562.81159			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	13.075	2.822	4.632	0.000	7.441	18.709
BigMac	0.309	0.058	5.315	0.000	0.193	0.425

Estudo de Caso – Regressão Linear Simples



- ❑ Interpretação
- ❑ **Multiple R**
 - ❑ É o coeficiente de correlação – faça a raiz de R^2
- ❑ **Adjusted R-squared = R^2 ajustado**
 - ❑ É o R^2 penalizado pelo número de parâmetros do modelo
- ❑ **Std. Error (erro padrão) – é a estimativa do desvio padrão dos erros, isto é, s.**

Estudo de Caso – Regressão Linear Simples



- **Interpretação**
- **Std. Error (erro padrão)**
 - Calcule $s^2 = SSE/(n-2)$ e tire a raiz
 - Neste caso, $SSE = 15167.85$, $n = 69$
- **Observations (igual a n)**
- **Tabela ANOVA**
 - **Contém as somas dos quadrados na coluna SS**
 - **Regression = SSReg**
 - **Residual = SSE = SYY**
 - **Total = SST**

Estudo de Caso – Regressão Linear Simples



- **Tabela ANOVA**
 - **A coluna df** contém os graus de liberdade associados a cada uma destas somas de quadrados
 - **A coluna MS** (mean square) contém as somas de quadrados (SS) divididas pelos respectivos graus de liberdade (df)
 - **$F = MS(\text{Regression})/MS(\text{Residual})$**
 - **F** tem uma distribuição F com os graus de liberdade n_1 no numerador e n_2 no denominador. Neste caso, $n_1 = 1$ e $n_2 = 67$.

Estudo de Caso – Regressão Linear Simples



- **Tabela ANOVA**
- **Em geral, no caso da regressão linear simples:**

$$F = \frac{SS \text{ Reg} / 1}{SSE / (n - 2)} = \frac{MS \text{ Reg}}{MSE}$$

- **É uma variável F com distribuição F(1, n-2) graus de liberdade.**
- **Se a estatística F é grande, a regressão é significativa (ou seja, neste caso, $\beta_1 \neq 0$). O teste F é equivalente ao teste t.**

Estudo de Caso – Regressão Linear Simples



	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	13.075	2.822	4.632	0.000	7.441	18.709
BigMac	0.309	0.058	5.315	0.000	0.193	0.425

b_0 e b_1

$dp(b_0)$ e $dp(b_1)$

b_0 e b_1 divididos pelos seus d.p.

IC para b_0 e b_1 baseados na distribuição t

Se o p-value é pequeno (digamos, abaixo de 5%, os parâmetros são significantes)

Estudo de Caso – Regressão Linear Simples



- ❑ O “slide” anterior nos diz que a reta ajustada é:
- ❑ $BREAD = 13.075 + 0.309 * BIGMAC$
- ❑ Você pode perceber que esta era a reta que aparecia no gráfico da linha de tendência.
- ❑ Também, podemos notar que:
 - ❑ O modelo não é “bom”, mas os parâmetros são significantes – ou seja, o modelo linear é certamente melhor que o modelo constante neste caso.

Estudo de Caso – Regressão Linear



- ❑ A planilha `stat_case2.xls` contém as vendas de carros, TV a cores e videocassetes no mercado brasileiro entre Janeiro de 1995 e Dezembro de 1997.
- ❑ Calcule a matriz de correlação entre as variáveis.
- ❑ Faça o gráfico de vendas de carros versus vendas de TV.

Estudo de Caso – Regressão Linear



- ❑ Faça o gráfico de vendas de TV versus vendas de videocassetes.
- ❑ Faça o gráfico de vendas de carros versus vendas de videocassetes.
- ❑ Ajuste um modelo de regressão linear simples onde $y =$ vendas de TV e $x =$ vendas de videocassetes.

Estudo de Caso – Regressão Linear



- ❑ Matriz de correlação

	<i>carros</i>	<i>TV</i>	<i>Video</i>
<i>carros</i>	1		
<i>TV</i>	0.6524	1	
<i>Video</i>	0.6850	0.9495	1

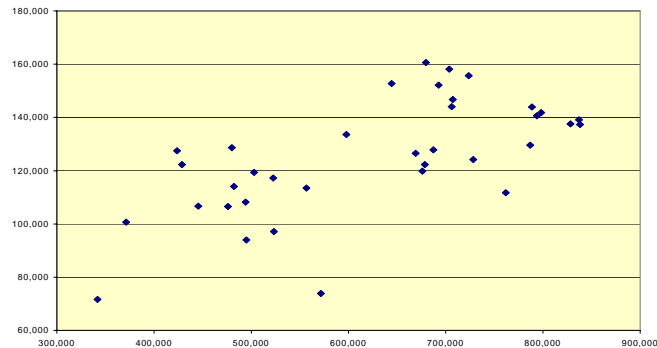
- ❑ O coeficiente de correlação de uma variável com ela mesma é sempre 1. Por que? Dica: qual a definição do coeficiente de correlação?
- ❑ Note também a alta correlação entre vendas de TV e Video.

Estudo de Caso – Regressão Linear



Gráfico carros X TV

Vendas de Carros (eixo Y) versus Venda de TVs (eixo X)



monica@mbarros.com

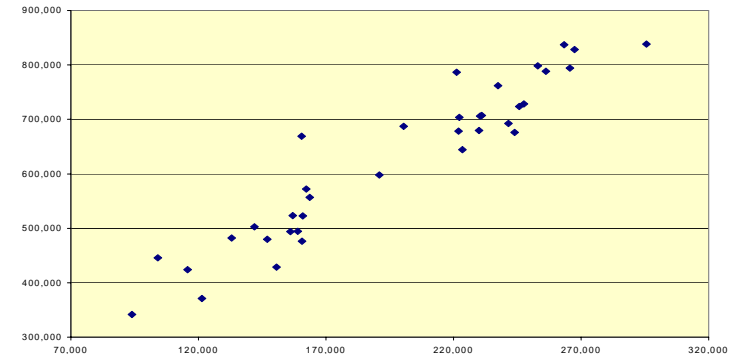
101

Estudo de Caso – Regressão Linear



Gráfico TV X Video

Vendas de TVs (eixo Y) versus Vendas de Videocassetes (eixo X)



monica@mbarros.com

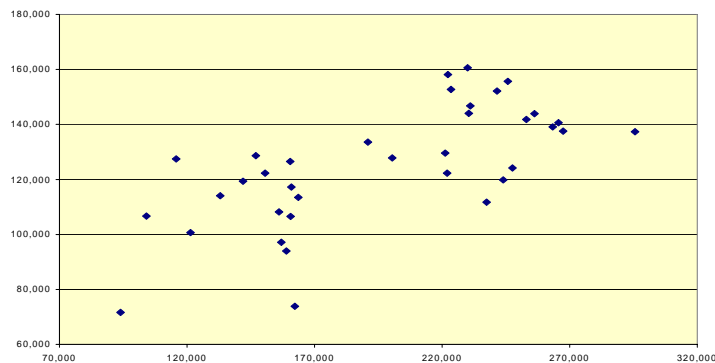
102

Estudo de Caso – Regressão Linear



Gráfico Carros X Video

Vendas de Carros (eixo Y) versus Vendas de Videocassetes (eixo X)



monica@mbarros.com

103

Estudo de Caso – Regressão Linear



Regressão Linear de TV em Video

Estatísticas da Regressão

Multiple R	0.9495	<<< é o coef. de correlação
R Square	90.1%	<<< é o coef. de determinação
Adjusted R Square	89.9%	
Standard Error	45,634	
Observations	36	<<< número de pares de observações

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
bo	126,709.81	29,167.78	4.34	0.01%	67,433.78	185,985.84
b1	2.53	0.14	17.64	0.00%	2.24	2.82

- Note que agora o ajuste da regressão é **ÓTIMO**, como já esperado, pois o gráfico entre as duas variáveis mostrava uma relação altamente linear.

monica@mbarros.com

104

Estudo de Caso – Regressão Linear



- ❑ Regressão Linear de TV em Video
- ❑ Dicas:
 - ❑ Quando os coeficientes b_0 e b_1 são significantes?
 - ❑ Em geral, se suas estatísticas t forem maiores que 2, em módulo, podemos afirmar que os coeficientes são diferentes de zero.
 - ❑ Também podemos olhar para os intervalos de confiança para b_0 e b_1 – se estes intervalos NÃO INCLUEM ZERO podemos dizer que os coeficientes são significantes, e é exatamente este o caso que acabamos de mostrar na página anterior!
- ❑ Na próxima página está o gráfico dos valores reais e ajustados (previstos) pela reta de regressão.

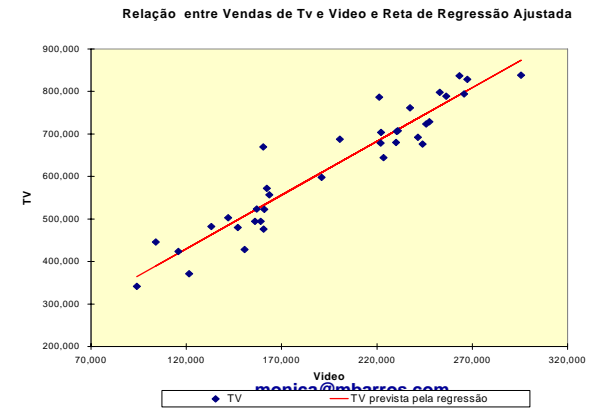
monica@mbarros.com

105

Estudo de Caso – Regressão Linear



- ❑ Valores Reais e Previstos pela Regressão



106

Estudo de Caso – Regressão Linear

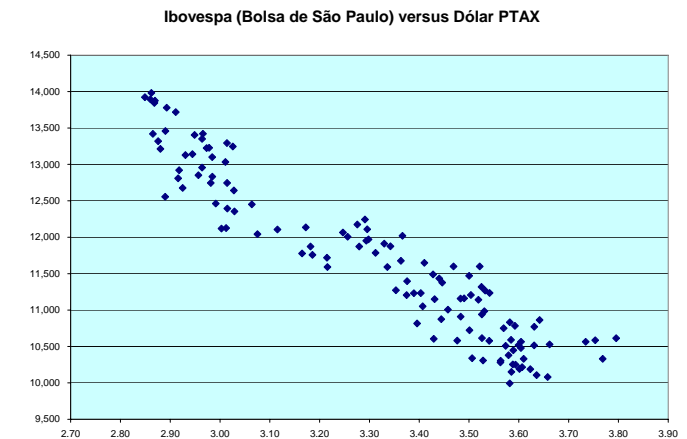


- ❑ Dólar e Índice Bovespa
- ❑ O próximo gráfico exibe a relação entre a cotação diária do dólar PTAX e o IBOVESPA no período entre 10/12/2002 e 12/06/2003.
- ❑ Ajuste um modelo de regressão linear simples para o IBOVESPA usando a PTAX como variável explicativa.

monica@mbarros.com

107

Estudo de Caso – Regressão Linear



monica@mbarros.com

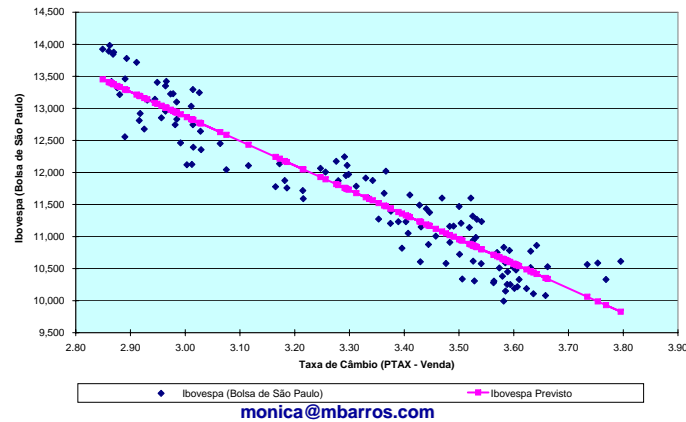
108

Estudo de Caso – Regressão Linear



Valores Reais e Previstos

IBOVESPA e Taxa de Câmbio - Valores Observados e Ajustados pela Regressão



monica@mbarros.com

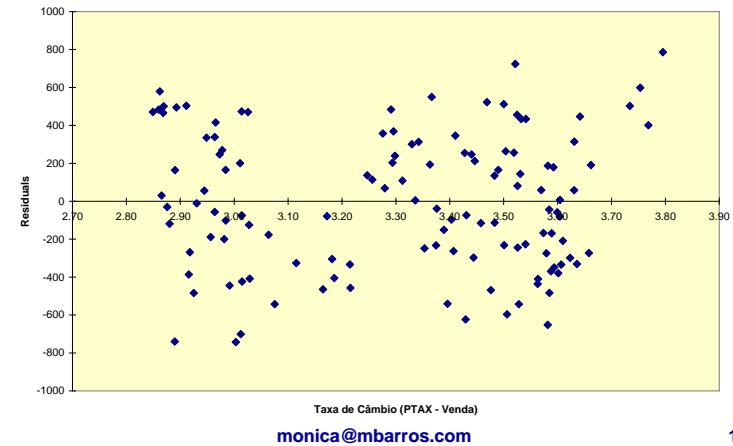
109

Estudo de Caso – Regressão Linear



Resíduos

Resíduos da Regressão



monica@mbarros.com

110

Estudo de Caso – Regressão Linear



Exemplo de Output do Excel

Regression Statistics	
Multiple R	0.946
R Square	0.895
Adjusted R Square	0.895
Standard Error	361.437
Observations	124

ANOVA					
	df	SS	MS	F	Significance F
Regression	1	136,397,746	136,397,746	1044.1	1.1981E-61
Residual	122	15,937,651	130,636		
Total	123	152,335,397			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	24,366.28	393.58	61.91	5.1E-94	23587	25145
Taxa de Câmbio (PTAX - Venda)	(3,830.74)	118.55	-32.31	1.2E-61	-4065	-3596

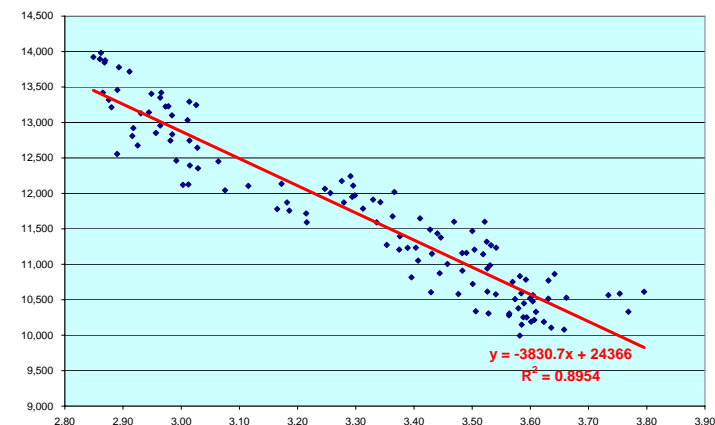
monica@mbarros.com

111

Estudo de Caso – Regressão Linear



Ibovespa versus Dólar PTAX



monica@mbarros.com

112

Estudo de Caso – Regressão Linear – Para Casa



- ❑ A planilha EE_anual.xls contém alguns dados sobre consumo anual residencial de energia elétrica no Brasil, PIB real e população (real e economicamente ativa).
- ❑ O objetivo aqui é que você explore estes dados e encontre modelos de regressão simples e (depois) múltipla que possam servir para fornecer razoáveis explicações para o crescimento do consumo de EE residencial.
- ❑ Alguns gráficos obtidos da planilha estão a seguir, e podem servir como “dicas”.

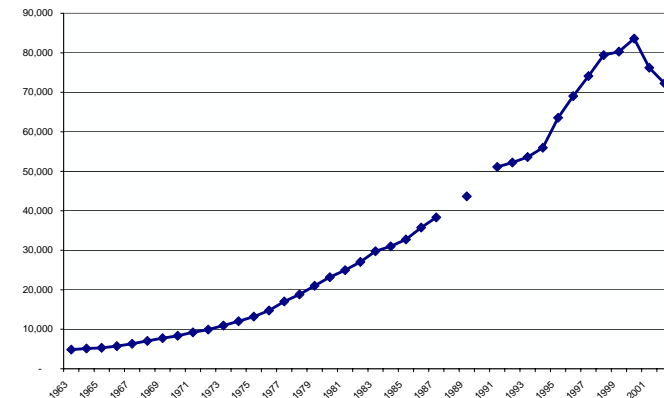
monica@mbarros.com

113

Estudo de Caso – Regressão Linear – Para Casa



Energia elétrica - consumo residencial GWh



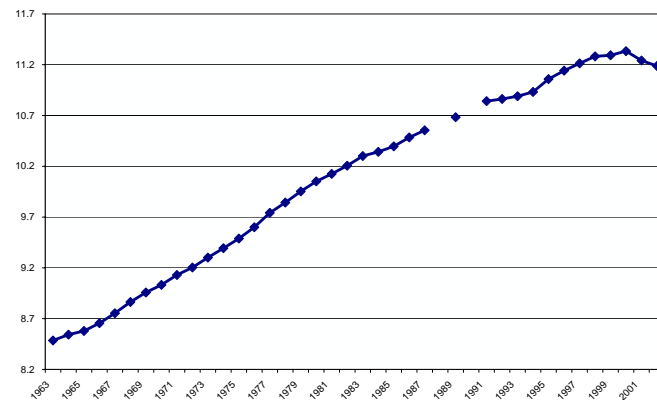
monica@mbarros.com

114

Estudo de Caso – Regressão Linear – Para Casa



Energia elétrica - log (consumo residencial GWh)



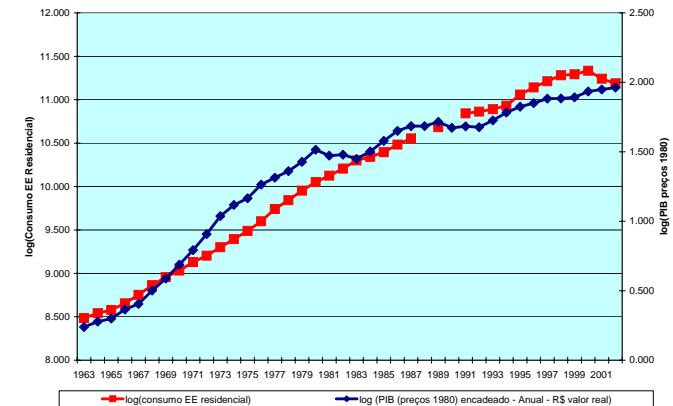
monica@mbarros.com

115

Estudo de Caso – Regressão Linear – Para Casa



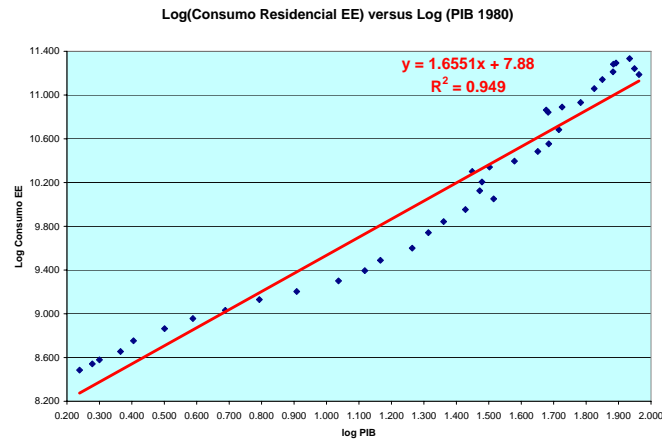
log(Consumo EE Residencial) e log(PIB preços 1980)



monica@mbarros.com

116

Estudo de Caso – Regressão Linear – Para Casa



monica@mbarros.com

117

Estudo de Caso – Regressão Linear – Para Casa



- ❑ A planilha fuel2001.xls contém dados de consumo de combustível para 50 estados americanos em 2001.
- ❑ A descrição das variáveis está na planilha.
- ❑ Construa a variável FUELPC, o consumo de combustível per capital em cada estado, através da expressão:

$$FUELPC = 1000 * \frac{FUEL}{POP}$$

monica@mbarros.com

118

Estudo de Caso – Regressão Linear – Para Casa



- ❑ Construa a variável DLIC, a proporção de pessoas acima de 16 anos no Estado que possuem carteira de motorista.

$$DLIC = \frac{NLIC}{POP}$$

- ❑ Faça o gráfico de FUELPC versus TAX.
- ❑ Faça o gráfico de FUELPC versus DLIC.
- ❑ Ajuste no Excel a regressão de FUELPC em TAX. Comente os resultados.
- ❑ Ajuste no Excel a regressão de FUELPC em DLIC. Comente os resultados.

monica@mbarros.com

119



Regressão Linear Múltipla

monica@mbarros.com

120

Regressão Linear Múltipla



- Na regressão linear múltipla, diversas variáveis independentes são usadas para modelar uma única variável resposta.
- São observados n “casos”, e em cada um deles, o valor da variável resposta (Y) e de cada variável independente X_1, X_2, \dots, X_p está disponível.

Regressão Linear Múltipla



- Então os dados disponíveis formam a matriz:

caso	Y	X ₁	X ₂	X ₃	...	X _p
1	y ₁	x ₁₁	x ₁₂	x ₁₃		x _{1p}
2	y ₂	x ₂₁	x ₂₂	x ₂₃		x _{2p}
...	...					
n	y _n	x _{n1}	x _{n2}	x _{n3}		x _{np}

- No modelo de regressão linear SIMPLES, $p = 1$.
- Nesta representação, x_{ij} refere-se ao i-ésimo caso e à j-ésima variável.

Regressão Linear Múltipla



- Modelo

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + e$$

- Onde:

- β 's são parâmetros desconhecidos (a serem estimados)
- e 's são erros, variáveis aleatórias independentes, com média zero e variância constante
- X_1, X_2, \dots, X_p são as **variáveis explicativas, preditores ou covariáveis**
- Y é a **variável dependente, ou variável resposta**

Regressão Linear Múltipla



- O modelo de regressão linear múltipla é completamente especificado a partir das seguintes hipóteses:

1) LINEARIDADE

- A relação entre a variável resposta e as variáveis explicativas é linear, como mostra a equação do slide anterior.



2) NÃO OCORRÊNCIA DE MULTI-COLINEARIDADE PERFEITA

- Em outras palavras, nenhuma variável explicativa é uma combinação linear das outras.
- Pode-se reescrever esta condição em termos da seguinte matriz, conhecida como matriz de design:



$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \dots & & & & \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

X é a **matriz de design**. Note que a coluna de “uns” corresponde ao termo constante da regressão

- A condição 2 pode ser reescrita como:

$$\text{Posto}(X) = p + 1 = \min(n, p+1)$$

Nota: o posto de uma matriz é o número de linhas (colunas) linearmente independentes.



3) VARIÁVEIS EXPLICATIVAS NÃO SÃO ESTOCÁSTICAS

4) DISTRIBUIÇÃO DOS ERROS

e_i são iid $N(0, \sigma^2)$ para $i = 1, 2, \dots, n$

- Algumas conseqüências desta hipótese são:
 - Todos os erros têm a mesma variância (**homocedasticidade**)
 - Todos os erros são **descorrelatados** (o que é uma conseqüência da independência).



□ Modelo em Forma Matricial

- É conveniente escrever o modelo em representação matricial como a seguir:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}, e = \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_p \end{pmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \dots & & & & \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

- Dimensões: \underline{Y} é $n \times 1$, \underline{e} é $n \times 1$, $\underline{\beta}$ é $(p+1) \times 1$
- e X é $n \times (p+1)$

Regressão Linear Múltipla



- A equação de regressão múltipla pode ser escrita em forma matricial como:

$$Y = X\beta + e$$

- Pode-se escrever as hipóteses sobre a distribuição dos erros em termos de uma distribuição Normal multivariada:

$$e \approx N_n(0, \sigma^2 I) \text{ onde } I \text{ é a matriz identidade } n \times n \text{ e } 0 \text{ é um vetor de zeros de dimensão } n \times 1$$

Regressão Linear Múltipla



- Soma do quadrado dos Resíduos
- A soma do quadrado dos resíduos (SSE) pode ser escrita em termos matriciais como:

$$SSE = \hat{e}'\hat{e} = (Y - \hat{Y})'(Y - \hat{Y}) = (Y - Xb)'(Y - Xb)$$

- Onde \underline{b} é o vetor $(p+1) \times 1$ de coeficientes estimados, ou seja, o vetor de estimadores dos β 's.
- O vetor \hat{Y} é o vetor $n \times 1$ de valores ajustados da regressão.

Regressão Linear Múltipla



- Estimadores de Mínimos Quadrados
- Como na regressão simples, o objetivo é encontrar os b_i 's (aqui: b_0, b_1, \dots, b_p) que minimizem a soma do quadrado dos resíduos. Ou seja, temos que resolver a equação:

$$\frac{\partial SSE}{\partial b} = 0 \Rightarrow \frac{\partial \{(Y - Xb)'(Y - Xb)\}}{\partial b} = 0$$

$$\frac{\partial \{Y'Y - 2b'X'Y + b'X'Xb\}}{\partial b} = 0$$

$$-2X'Y + 2X'Xb = 0$$

Equações Normais



$$X'Xb = X'Y$$

Regressão Linear Múltipla



- Estimadores de Mínimos Quadrados
- O sistema anterior (as equações normais) pode ser resolvido pré-multiplicando pela inversa de $X'X$, e resulta em:

$$b = (X'X)^{-1} X'Y$$

- Uma vez encontrados o vetor de estimadores b , pode-se obter o vetor de valores ajustados $\hat{Y} = Xb$ e o vetor de resíduos:

$$\hat{e} = Y - \hat{Y} = Y - Xb$$

Regressão Linear Múltipla



- **Estimadores de Mínimos Quadrados**
- Os resíduos e a matriz de design são decorrelatados, isto é:

$$X^t \cdot \hat{e} = 0$$

- A soma dos resíduos para todas as observações é nula. Isso pode ser provado da equação anterior fazendo-se o produto da 1a. coluna de X (que só contém 1's) com os resíduos.

$$\sum_{i=1}^n \hat{e}_i = \sum_{i=1}^n (y_i - \hat{y}_i) = 0$$

Regressão Linear Múltipla



- **Propriedades dos Estimadores**

- Suponha que $E(e) = 0$, um vetor coluna de zeros de dimensão n, e $\text{Var}(e) = \sigma^2 I$, onde I é a matriz identidade de dimensão n. Então:
- **b é não tendencioso** para β , isto é: $E(b_i) = \beta_i$ para $i = 0, 1, 2, \dots, p$
- **A matriz de variância-covariância dos estimadores é:**

$$\text{VAR}(b) = \sigma^2 (X^t X)^{-1}$$

Regressão Linear Múltipla



- **Estimador da variância dos erros**

$$s^2 = \frac{SSE}{n-p-1} = \frac{(Y - Xb)^t (Y - Xb)}{n-p-1} = \frac{Y^t Y - b^t (X^t X) b}{n-p-1} = \frac{Y^t Y - b^t X^t \hat{Y}}{n-p-1}$$

Regressão Linear Múltipla



- **Análise de variância**

- Na análise de variância, o modelo completo (incluindo todas as variáveis explicativas X_1, X_2, \dots, X_p) é comparado com o modelo constante, em que não estão presentes quaisquer variáveis explicativas, apenas a constante β_0 existe no modelo.
- Já vimos que no modelo constante, o estimador de mínimos quadrados de β_0 é a média dos Y 's e portanto a soma do quadrado dos resíduos neste modelo é $SYY = SST$.

Regressão Linear Múltipla



□ Tabela ANOVA

Tabela ANOVA			
Fonte	graus de liberdade	SS	MS
Regressão nos X_i 's	p	SSReg	SSReg/p
Resíduo	n-p-1	SSE	SSE/(n-p-1) = s^2
TOTAL	n-1	SST = SYY	

- É claro que $SSE < SYY$, e novamente (como na regressão simples) a diferença entre ambos é $Ssreg$, a soma de quadrados em Y explicada pelo modelo maior que NÃO é explicada pelo modelo menor.

Regressão Linear Múltipla



□ Tabela ANOVA

- A importância da regressão nos X 's é determinada pelo tamanho relativo de SSReg em relação a SSE.
- A estatística F a ser usada é:

$$F = \frac{MS_{Reg}}{MS_{Residual}} = \frac{SS_{Reg}/p}{SSE/(n-p-1)} = \frac{SS_{Reg}/p}{s^2} \sim F(p, n-p-1)$$

- Se a estatística F é "grande" então o modelo incluindo os X 's é significativamente melhor que o modelo constante. A distribuição desta estatística é EXATAMENTE F se os erros são iid Normais.

Regressão Linear Múltipla



□ O coeficiente de determinação (R^2)

- Como na regressão simples:

$$R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} = \frac{SS_{Reg}}{SST} = \frac{b'X'Y}{SST}$$

- Fornece a proporção da variabilidade dos Y 's explicada pela regressão nos X 's. Pode-se mostrar que o R^2 é o quadrado do coeficiente de correlação múltipla entre Y e os X 's, ou seja, é o quadrado da MÁXIMA correlação entre Y e uma função linear dos X 's.

Regressão Linear Múltipla



□ R^2 ajustado

- Proposto por Theil para comparar modelos com número diferente de variáveis:

$$R_{adj}^2 = 1 - \frac{SSE/(n-p)}{SST/(n-p-1)} = 1 - \frac{n-1}{n-p-1} (1 - R^2)$$

- Exemplo numérico
- $R^2 = 0.90$, $p = 3$, $n = 20$ obs, o R^2 ajustado será: 0.8813. Se o modelo passa agora a ter $p = 5$ variáveis, o R^2 ajustado cai para 0.8643.

Regressão Linear Múltipla



- ❑ O cálculo da estatística F geral da regressão não costuma ser muito interessante.
- ❑ O fato é que normalmente se sabe a priori que as variáveis são relacionadas e então já se espera que o valor da estatística F seja grande.
- ❑ Mais interessante é olhar para o teste de hipóteses referentes a variáveis individuais.

Regressão Linear Múltipla



- ❑ Teste t para cada parâmetro na regressão

- ❑ Sabe-se que:

$$b_i \sim N(\beta_i, \sigma^2 a_{ii}) \quad \text{onde } a_{ii} \text{ é o } i\text{-ésimo elemento da diagonal principal de } (X'X)^{-1}$$
$$\frac{(n-p-1)s^2}{\sigma^2} \sim \chi_{n-p-1}^2 \quad \text{independente de } b_i$$

- ❑ Isso nos permite formar estatísticas t:

$$t = \frac{b_i - \beta_i}{s\sqrt{a_{ii}}} \sim t_{n-p-1} \quad \text{para } i=0,1,2,\dots,p$$

Regressão Linear Múltipla



- ❑ O teste de hipótese compara o valor observado na amostra com percentis de interesse da distribuição t.
- ❑ Importante é o caso da significância de uma determinada variável, por exemplo, X_k . Esta variável será significativa no modelo se a estatística t associada ao seu coeficiente for diferente de zero.

Regressão Linear Múltipla



- ❑ Teorema de Gauss-Markov

- ❑ Os estimadores de mínimos quadrados ordinários são BLUE (“best linear unbiased estimators”), ou seja, são os MELHORES estimadores LINEARES NÃO TENDENCIOSOS.

- ❑ “Melhores” em que sentido?

- ❑ No sentido de apresentarem a MENOR VARIÂNCIA.

- ❑ Na classe dos estimadores lineares e não tendenciosos, os estimadores por mínimos quadrados são os que apresentam variância mínima.

Regressão Linear Múltipla



□ Previsão

- Os resultados da estimação das equações de regressão podem ser usados para prever a variável dependente em valores das variáveis explicativas diferentes daqueles que constavam da amostra original.
- Seja u um vetor $(p+1) \times 1$ de valores observados "novos" das variáveis explicativas X_1, X_2, \dots, X_p , tal que o 1º elemento de u é 1 (para levar em conta o termo constante da regressão).

Regressão Linear Múltipla



□ Previsão

- Então o valor ajustado no ponto u é:

$$\hat{Y} = u^t \cdot b$$

- E o erro padrão do valor ajustado é:

$$erropadrao = s \cdot \sqrt{u^t (X^t X)^{-1} u}$$

- Mas, não podemos esquecer que, a qualquer previsão está associado um erro aleatório, isto é, a equação de previsão é:

$$Y_{previsao} = u^t \cdot b + e$$

- E portanto o erro padrão da previsão torna-se:

$$e.p._{previsao} = s \cdot \sqrt{1 + u^t (X^t X)^{-1} u}$$

Regressão Linear Múltipla



□ Estimação por Máxima Verossimilhança

- Suponha que os erros do modelo são independentes, Normalmente distribuídos e com variância constante.
- Então os Y_i são independentes Normais com média:

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}$$

- E variância σ^2 .

Regressão Linear Múltipla



□ Estimação por Máxima Verossimilhança

- Em termos vetoriais, podemos escrever que o vetor Y tem distribuição Normal multivariada de dimensão n com vetor de médias $X \cdot \beta$ e matriz de variância-covariância $\sigma^2 \cdot I$.
- Isso nos permite escrever (facilmente?) a verossimilhança:

$$L(\beta_0, \beta_1, \dots, \beta_p, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \left(\exp \left\{ \frac{-1}{2\sigma^2} (Y - X\beta)^t (Y - X\beta) \right\} \right) \\ = (2\pi\sigma^2)^{-n/2} \left(\exp \left\{ \frac{-1}{2\sigma^2} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_{i1} - \dots - \beta_p X_{ip})^2 \right\} \right)$$

Regressão Linear Múltipla



□ Estimação por Máxima Verossimilhança

□ A log-verossimilhança é:

$$l(\beta_0, \beta_1, \dots, \beta_p, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_{i1} - \dots - \beta_p X_{ip})^2$$

- Da última expressão do lado direito vemos que maximizar a log-verossimilhança para os β_i 's é equivalente a minimizar a soma de quadrado dos resíduos e portanto o método de máxima verossimilhança fornece os **MESMOS** estimadores dos β_i 's que o método de mínimos quadrados.

Regressão Linear Múltipla



□ Estimação por Máxima Verossimilhança

□ E o estimador da variância?

□ É um estimador tendencioso, dado por:

$$\hat{\sigma}^2 = \frac{1}{n} (Y - Xb)' (Y - Xb) = \frac{SSE}{n} = \frac{(n-p-1)s^2}{n}$$

□ Já vimos que s^2 é um estimador não tendencioso de σ^2 , portanto:

$$E(\hat{\sigma}^2) = E\left(\frac{SSE}{n}\right) = E\left(\frac{(n-p-1)s^2}{n}\right) = \frac{(n-p-1)\sigma^2}{n}$$

Regressão Linear Múltipla



□ Variáveis Dummy

- Até agora, todas as variáveis nos nossos modelos eram quantitativas. Neste caso, a magnitude da variável é de interesse.
- Na prática, precisamos também incorporar o efeito de variáveis qualitativas nos nossos modelos de regressão. Por exemplo, raça, sexo, região do país, estação do ano são todas variáveis qualitativas.
- Aqui iremos considerar apenas variáveis **EXPLICATIVAS** qualitativas – a variável dependente será sempre quantitativa.

Regressão Linear Múltipla



□ Variáveis Dummy

- Fatores qualitativos muitas vezes aparecem na forma de variáveis binárias, por exemplo: uma pessoa é homem ou mulher, uma família tem renda acima de 20 S.M. ou não.
- Nestes exemplos, a informação relevante pode ser capturada definindo-se uma variável binária.
- Em estatística, variáveis binárias são geralmente chamadas de "dummies".
- Ao definir uma "dummy" precisamos decidir quem será o valor 0 e quem será o 1 e esta escolha é, até certo ponto, arbitrária.

Regressão Linear Múltipla



□ Variáveis Dummy

- Por que usar os valores 0 e 1 para representar uma variável qualitativa? Porque simplifica a interpretação dos resultados, mas a princípio você poderia usar quaisquer valores.

□ Considere o seguinte modelo:

$$\text{salario} = \beta_0 + \beta_1(\text{mulher}) + \beta_2(\text{educacao}) + e$$

Regressão Linear Múltipla



□ Variáveis Dummy

- Onde "salario" e "educacao" representam o salário por hora trabalhada e o número de anos de educação formal, e "mulher" é uma dummy com valor 1 se a pessoa é mulher e 0 se é homem.
- A interpretação do coeficiente β_1 é a diferença nos salários devida ao sexo da pessoa, dado o mesmo nível educacional. Ou seja, β_1 nos diz se há discriminação contra as mulheres – se $\beta_1 < 0$ então para o mesmo nível de educação, os salários das mulheres são menores que os dos homens.

Regressão Linear Múltipla



□ Variáveis Dummy

- Se $\beta_1 < 0$, os homens ganham uma quantidade fixa a mais por hora. A reta correspondente aos homens é **paralela** à das mulheres, mas está **ACIMA** desta última – para todos os níveis de educação os homens ganham mais, a diferença é $-\beta_1$.
- Ou seja, as duas retas têm interceptos diferentes e o mesmo coeficiente angular. Na reta das mulheres, o intercepto é $\beta_0 + \beta_1$, que é menor que o intercepto da reta dos homens (β_0).

Regressão Linear Múltipla



□ Variáveis Dummy

- Imagine que você quisesse incluir duas dummies, uma para homens e outra para mulheres. O que fazer? Retirar a constante, senão você terá colinearidade perfeita.
- Mas, em termos de interpretação do modelo, não é mais fácil interpretar os resultados deste último do que o modelo original que propusemos.

Regressão Linear Múltipla



□ Variáveis Dummy

- Em geral, se desejamos modelar g categorias, devemos criar $g-1$ variáveis dummy num modelo que inclui uma constante. O grupo "base" é o que está sendo omitido da especificação, e o intercepto da regressão representa o intercepto deste grupo.
- Os coeficientes das dummies são as diferenças entre o intercepto daquele grupo e do grupo "base".
- Uma especificação alternativa é incluir dummies para todos os grupos e retirar a constante do modelo, mas isso torna a interpretação das diferenças entre grupos mais complicada.

monica@mbarros.com

157

Regressão Linear Múltipla



□ Variáveis Dummy

□ Exemplo (adaptado de Wooldridge)

- Os dados a seguir se referem a uma amostra de americanos em 1976. Além das variáveis "salario", "educ", "mulher", adicionamos ao modelo a variável "tenure", que representa o número de anos no emprego atual.
- O modelo de regressão a seguir foi ajustado no software estatístico Minitab.

monica@mbarros.com

158

Regressão Linear Múltipla



□ Variáveis Dummy

The regression equation is

$$\text{salario} = -1.57 + 0.572 \text{ educ} + 0.0254 \text{ exper} + 0.141 \text{ tenure} - 1.81 \text{ mulher}$$

Predictor	Coef	SE Coef	T	P
Constant	-1.5679	0.7246	-2.16	0.031
educ	0.57150	0.04934	11.58	0.000
exper	0.02540	0.01157	2.20	0.029
tenure	0.14101	0.02116	6.66	0.000
mulher	-1.8109	0.2648	-6.84	0.000

S = 2.958

R-Sq = 36.4%

R-Sq(adj) = 35.9%

monica@mbarros.com

159

Regressão Linear Múltipla



□ Variáveis Dummy

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	2603.11	650.78	74.40	0.000
Residual Error	521	4557.31	8.75		
Total	525	7160.41			

Source	DF	Seq SS
educ	1	1179.73
exper	1	432.52
tenure	1	581.86
mulher	1	408.99

monica@mbarros.com

160

Regressão Linear Múltipla



- **Variáveis Dummy**
- **Interpretação dos resultados**
 - Todas as variáveis significantes ao nível 5%.
 - Coeficiente de "mulher" = -1.81, indicando que o salário médio por hora de uma mulher, mantidos todos os outros fatores constantes, é inferior ao de um homem em US\$ 1.81.

Regressão Linear Múltipla



- **Variáveis Dummy**
 - Como já controlamos o efeito das outras variáveis (educação, experiência e "tenure"), a variável "dummy" captura o efeito no salário que não pode ser explicado por estes fatores e que pode ser atribuído ao sexo da pessoa.
- **Interpretação do coeficiente – modelo na escala dos logs**
 - Se a variável dependente está expressa como log, o coeficiente da dummy deve ser interpretado como um percentual.
 - Em geral, se β é o coeficiente de uma variável dummy x_1 , a diferença percentual exata em y quando $x_1 = 1$ versus quando $x_1 = 0$ é: $100(\exp(\beta) - 1)$

Regressão Linear Múltipla



- **Estudo de Caso**
 - Considere a planilha ceosal1.xls
 - O objetivo é prever o salário do CEO (principal executivo) de uma empresa com base na performance recente da mesma e no setor em que ela atua.
 - A planilha contém as seguintes variáveis:
 - salario = salário do CEO em milhares de dólares de 1990
 - var%_salario = variação percentual do salário em 1989/1990
 - vendas = vendas da empresa em 1990
 - ROE = retorno da empresa – 1989/1990
 - var%ROE = variação % do ROE – 1989/1990

Regressão Linear Múltipla



- **Estudo de Caso (continuação)**
 - retorno_ação = retorno da ação 1988 a 1990
 - set_indus = dummy indicadora do setor "indústria"
 - set_finan = dummy indicadora do setor financeiro
 - set_consumo = dummy indicadora do setor de bens de consumo
 - set_transp = dummy indicadora do setor de transportes e concessionárias de serviços públicos
 - log_salario = logaritmo do salário
 - log_venda = logaritmo das vendas

Regressão Linear Múltipla



□ Estudo de Caso (continuação)

- Ajuste um modelo para $\log(\text{salario})$ em $\log(\text{vendas})$, roe e nas variáveis "dummy". Use como "grupo base" (variável dummy omitida) o setor de "indústrias".
- **O que você conclui?**
- Qual a diferença percentual em salário, mantidas todas as outras variáveis constantes, entre um CEO no setor "transporte e concessionárias" e outro no "grupo base"?

Regressão Linear Múltipla



□ Estudo de Caso (continuação)

- Qual a diferença percentual em salário (mantidas todas as outras variáveis constantes) entre um CEO no setor "financeiro" e outro no "grupo base"?
- E entre um no setor financeiro e um no setor de bens de consumo?

Regressão Linear Múltipla



□ Análise dos Resíduos

- Lembre-se das hipóteses básicas sobre os erros:
 - Independência
 - Normalidade
 - Variância constante
- A análise dos resíduos deve se preocupar em verificar até que ponto estas hipóteses estão sendo violadas e, se possível, corrigir os problemas.

Regressão Linear Múltipla



□ Análise dos Resíduos

- Lembre-se que os **erros** do modelo não são observáveis e portanto as hipóteses feitas não podem ser diretamente testadas NELES.
- O que fazer? Olhamos para os **resíduos**, que são os estimadores dos erros do modelo e verificamos se os resíduos são iid Normais com variância constante.

Regressão Linear Múltipla



- ❑ **Análise dos Resíduos**
- ❑ **Que tipo de comportamento você quer identificar?**
 - ❑ Se os resíduos são "muito" diferentes de uma Normal (através de um gráfico de probabilidade Normal)
 - ❑ Se existem padrões óbvios nos resíduos (por exemplo, se y cresce, o resíduo cresce)
 - ❑ Se existe autocorrelação nos resíduos.
 - ❑ Na prática – alguns destes gráficos, como a autocorrelação, não estão disponíveis no Excel, apenas softwares estatísticos permitem que você faça uma análise mais sofisticada.

monica@mbarros.com

169

Regressão Linear Múltipla



- ❑ **Análise dos Resíduos**
 - ❑ A seguir apresentamos os resultados e alguns gráficos de resíduos do modelo ajustado para o salário dos CEOs.
 - ❑ O modelo foi ajustado no Minitab.
- The regression equation is
 $lsalary = 4.59 + 0.257 lsales + 0.0112 roe + 0.158 finance + 0.181 consprod - 0.283 utility$

Predictor	Coef	SE Coef	T	P
Constant	4.5881	0.2950	15.55	0.000
lsales	0.25719	0.03203	8.03	0.000
roe	0.011152	0.004300	2.59	0.010
finance	0.15796	0.08900	1.77	0.077
consprod	0.18089	0.08477	2.13	0.034
utility	-0.28300	0.09923	-2.85	0.005

S = 0.4598 R-Sq = 35.7% R-Sq(adj) = 34.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	23.8110	4.7622	22.53	0.000
Residual Error	203	42.9112	0.2114		
Total	208	66.7222			

monica@mbarros.com

170

Regressão Linear Múltipla



	Valor ajustado		Resíduo		Resíduo padronizado	
Obs	lsales	lsalary	Fit	SE Fit	Residual	St Resid
15	7.4	7.6064	7.1103	0.1748	0.4961	1.17 X
28	9.9	8.2543	7.2778	0.0761	0.9765	2.15R
62	6.3	5.5452	6.4528	0.0883	-0.9076	-2.01R
82	8.1	6.0890	7.0154	0.0684	-0.9263	-2.04R
87	7.9	8.3292	6.9367	0.0695	1.3925	3.06R
108	9.1	8.8009	7.2000	0.0730	1.6009	3.53R
127	8.7	6.0568	7.1514	0.0741	-1.0946	-2.41R
129	5.7	5.8861	6.2583	0.1426	-0.3722	-0.85 X
133	5.2	5.4072	6.3599	0.1146	-0.9527	-2.14R
164	8.7	9.3266	7.2639	0.0611	2.0628	4.53R
166	8.3	8.2014	6.9840	0.0871	1.2173	2.70R
174	7.7	9.6039	6.9599	0.0644	2.6439	5.81R

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

- ❑ **O MINITAB fornece uma lista de observações com resíduo padronizado grande ou que têm grande influência na regressão.**

monica@mbarros.com

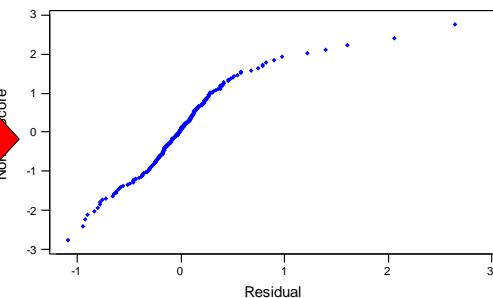
171

Regressão Linear Múltipla



- ❑ **Análise dos Resíduos**
- ❑ **Resíduos x Distribuição Normal**

Normal Probability Plot of the Residuals
 (response is lsalary)

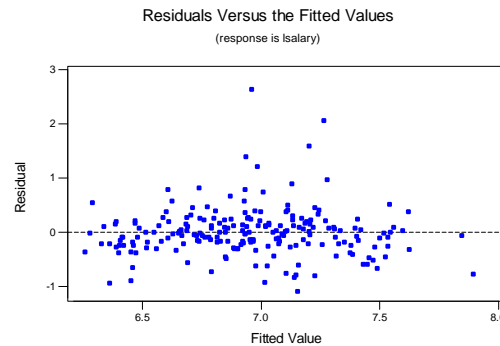


Uma reta indica que os resíduos são Normais – quanto mais "curvo" este gráfico, menos Normais os resíduos

Regressão Linear Múltipla



- ❑ **Análise dos Resíduos**
- ❑ **Resíduos x Valores Ajustados**

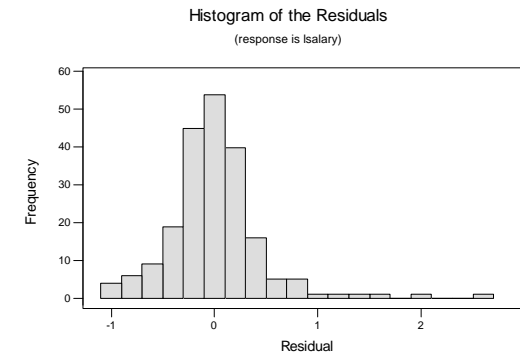


173

Regressão Linear Múltipla



- ❑ **Análise dos Resíduos**
- ❑ **Histograma dos Resíduos**



174

Regressão Linear Múltipla



- ❑ **Inadequação das Hipóteses do Modelo: multicolinearidade**
 - ❑ Uma das hipóteses do modelo de regressão múltipla é que não ocorre multicolinearidade perfeita, ou seja, o posto da matriz X é $(p+1)$, igual ao número de β 's a serem estimados.
 - ❑ O que acontece se esta hipótese for violada? **A matriz (X^tX) não poderá ser invertida e portanto os estimadores de mínimos quadrados não poderão ser calculados.**

monica@mbarros.com

175

Regressão Linear Múltipla



- ❑ **Multicolinearidade**
- ❑ **Mas, este é um caso extremo. Os problemas em que a inversão matemática desta matriz é impossível são chamados de "mal condicionados".**
- ❑ **Na prática, o que se vê é alta correlação entre uma ou mais variáveis. Qual o efeito disso?**

monica@mbarros.com

176

Regressão Linear Múltipla



- Multicolinearidade
- Exemplo 1 – multicolinearidade perfeita

$$X = \begin{pmatrix} 1 & 2 \\ 3 & 6 \end{pmatrix} \Rightarrow X'X = \begin{pmatrix} 10 & 20 \\ 20 & 40 \end{pmatrix}$$

- X^tX não pode ser invertida (seu determinante é zero)

Regressão Linear Múltipla



- Multicolinearidade
- Exemplo 2 – multicolinearidade “quase” perfeita

$$X = \begin{pmatrix} 1 & 2 \\ 3 & 5.9 \end{pmatrix} \Rightarrow X'X = \begin{pmatrix} 10 & 19.7 \\ 19.7 & 38.81 \end{pmatrix}$$
$$\Rightarrow (X'X)^{-1} = \begin{pmatrix} 3881 & -1970 \\ -1970 & 1000 \end{pmatrix}$$

- X^tX pode ser invertida (seu determinante é pequeno mas diferente de zero). **Note como os elementos da inversa de X^tX são grandes!**

Regressão Linear Múltipla



- Multicolinearidade
- Exemplo 3 – NÃO EXISTE multicolinearidade

$$X = \begin{pmatrix} 1 & 2 \\ 3 & 5 \end{pmatrix} \Rightarrow X'X = \begin{pmatrix} 10 & 17 \\ 17 & 29 \end{pmatrix}$$
$$\Rightarrow (X'X)^{-1} = \begin{pmatrix} 29 & -17 \\ -17 & 10 \end{pmatrix}$$

- X^tX pode ser invertida SEM PROBLEMAS. **Note como os elementos da inversa de X^tX são muito diferentes do exemplo anterior.**

Regressão Linear Múltipla



- Multicolinearidade
- Lembre-se que o estimador de mínimos quadrados é:

$$b = (X'X)^{-1} X'Y$$

- Dos exemplos anteriores, o efeito da multicolinearidade sobre este estimador deve ter ficado claro. Se as colunas de X forem “muito” dependentes (exemplo 2), X^tX pode ser invertida mas seu determinante é quase zero, e a inversa tem valores muito grandes.

Regressão Linear Múltipla



□ Multicolinearidade

- O impacto sobre os estimadores é também grande.
- Qualquer pequena mudança numa das variáveis pode mudar completamente o valor estimado do parâmetro.
- **Além disso...**
$$VAR(b) = \sigma^2 (X^T X)^{-1}$$
- **Se há multicolinearidade, a variância dos estimadores será grande (enorme?).**

Regressão Linear Múltipla



□ Multicolinearidade

- **Qual o impacto deste último problema?**
- Lembre-se que o desvio padrão dos estimadores (com s substituindo σ) entra no cálculo das estatísticas t que avaliam a significância dos parâmetros.
- **Então, o que pode acontecer?**
- O denominador da estatística t será muito grande, a estatística t será muito pequena, indicando que o parâmetro não é significante....

Regressão Linear Múltipla



□ Multicolinearidade

- **Logo, muitas vezes a multicolinearidade estará "escondendo" a significância dos parâmetros.**
- Como saber se isso está acontecendo de fato?
- **Olhe para o R^2 e a estatística F da regressão.** Nenhum deles envolve o cálculo da inversa de $(X^T X)$.
- Se o R^2 for alto, a estatística F for significativa e existirem várias estatísticas t insignificantes, é um bom indício de colinearidade.

Regressão Linear Múltipla



□ Multicolinearidade

- **Como e por que surge?**
- Na verdade, está muitas vezes relacionada a um tamanho insuficiente de amostra.
- Por exemplo, no caso de dados econômicos, é comum que diversas variáveis, num determinado intervalo de tempo, se "mexam" na mesma direção, gerando a multicolinearidade.

Regressão Linear Múltipla



- **Multicolinearidade**
- **Como identificar?**
 - **Dê uma olhada na matriz de correlação amostral das suas variáveis explicativas.**

 - **Se existem diversas variáveis altamente correlacionadas, muito provavelmente você vai enfrentar o problema da multicolinearidade.**

Regressão Linear Múltipla



- **Multicolinearidade**
- **Como resolver?**
 - **Retire algumas das variáveis altamente correlacionadas do seu modelo, isto é, comece a rodar seqüências de modelos menores.**

 - **Uma solução automática para isso é o procedimento de regressão "stepwise", que tem um único problema: pode gerar modelos sem muito sentido físico ou econômico.**

Regressão Linear Múltipla



- **Multicolinearidade**
- **E se a gente não fizer nada, o que acontece?**
 - **Depende...**
 - **Se o modelo for empregado para gerar previsões, a multicolinearidade nem sempre é um transtorno.**
 - **Os estimadores continuam não tendenciosos – o problema é a variância estimada!**

Regressão Linear Múltipla



- **Inadequação das Hipóteses do Modelo: heterocedasticidade**
 - **Em muitas situações práticas, a hipótese de que a variância dos erros é constante é questionável.**

 - **Por exemplo, se uma variável dependente positiva tem valores num intervalo que vai até milhares, é intuitivo que as respostas próximas de zero geralmente serão menos variáveis que as respostas em valores altos, pois estas últimas tem mais "espaço" para oscilar.**

Regressão Linear Múltipla



- **Heterocedasticidade**
- **Como identificar?**
 - Gráficos de resíduos versus valores ajustados ou variáveis explicativas.
 - Heterocedasticidade é observada através de um padrão como um “cone” aberto para a direita ou para a esquerda.
 - Um padrão parecido com uma “noz” pode ser observado se a variável resposta é um uma porcentagem – percentuais altos ou baixos são menos variáveis que aqueles próximos de 50%.

Regressão Linear Múltipla



- **Heterocedasticidade**
- **Como resolver?**
 - Transformações para a estabilização da variância \Rightarrow transformações aplicadas à variável resposta Y.
 - Algumas transformações comuns são:
 - **raiz(Y)** \Rightarrow dados de contagens
 - **Log(Y)** \Rightarrow quando $Y > 0$ e o intervalo de valores possíveis de Y é muito largo
 - **Log(Y + 1)** \Rightarrow mesma situação que acima, mas alguns dos $Y = 0$

Regressão Linear Múltipla



- **Heterocedasticidade**
- **Como resolver?**
 - **$1/Y$** \Rightarrow quando a variável resposta está “amontoadada” perto de zero, mas decresce rápido, embora existam alguns valores altos.
 - **$1/(Y + 1)$** \Rightarrow mesma situação que acima, mas alguns $Y = 0$
 - **Arc seno{ raiz(Y) }** \Rightarrow se $0 \leq Y \leq 1$ (ou seja, Y é uma proporção Binomial)

Regressão Linear Múltipla



- **Inadequação das Hipóteses do Modelo: autocorrelação dos resíduos**
 - Lembre-se que uma das hipóteses fundamentais do modelo é a independência dos erros.
 - Logo, se os resíduos são correlacionados, esta hipótese foi violada.
 - Por que? O que geralmente leva a este problema?
 - O mais comum são os problemas de especificação.

Regressão Linear Múltipla



Autocorrelação dos resíduos

- **Ao omitir alguma variável relevante, os resíduos do modelo poderão se tornar autocorrelacionados.**

- **Muitas vezes também procedimentos de dessazonalização de séries temporais geram autocorrelação nos resíduos.**

Regressão Linear Múltipla



Autocorrelação dos resíduos

Conseqüências

- **Suponha o seguinte modelo:**

$$Y_t = \beta_0 + \beta_1 x_t + e_t \quad \text{onde:}$$

$$e_t = \rho \cdot e_{t-1} + u_t \quad \text{e os } u_t \text{ são iid e } |\rho| < 1$$

Regressão Linear Múltipla



Autocorrelação dos resíduos

Conseqüências

- **Os estimadores de mínimos quadrados são ainda não tendenciosos, mas sua variância não é mínima.**

- **Agora:** $VAR(b) = (X'X)^{-1} X' \Omega X (X'X)^{-1}$

$$\text{onde } \Omega = \sigma_e^2 \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{bmatrix}$$

Regressão Linear Múltipla



Autocorrelação dos resíduos

Conseqüências

- **Os estimadores de mínimos quadrados ainda são não tendenciosos, MAS...**
- **Suas variâncias são **subestimadas**, invalidando os testes de hipóteses.**

Como detectar?

- **Caso mais simples – autocorrelação de lag 1 – Estatística de Durbin-Watson**



- Autocorrelação dos resíduos
- Como detectar?

$$DW = \frac{\sum_{t=2}^n (\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^n \hat{e}_t^2}$$

- onde n é o número de observações e \hat{e}_t é o t-ésimo resíduo do modelo
- A estatística de Durbin-Watson está relacionada com o estimador de ρ (que mede a autocorrelação dos erros).

- Pode-se provar que:
 $DW \cong 2(1 - \hat{\rho})$



- Autocorrelação dos resíduos
- Como detectar?

- Note que o numerador da estatística DW consiste na diferença ao quadrado entre resíduos sucessivos, e o denominador é apenas a soma dos quadrados dos resíduos.
- Existem n-1 termos na soma no numerador e n no denominador (por que?)



- Autocorrelação dos resíduos
- Como usar a estatística DW?

- Se $\rho = 0$ então $DW \cong 2$
- Se $0 < \rho < 1$ então $0 < DW < 2$ (caso mais comum na prática)
- Se $-1 < \rho < 0$ então $2 < DW < 4$

- Durbin e Watson derivaram a distribuição amostral de sua estatística. Dependendo do valor encontrado, o teste pode ser inconclusivo.



- Autocorrelação dos resíduos
- Como usar a estatística DW?

- O teste mais comum é o das hipóteses:
 $H_0: \rho = 0$ versus
 $H_1: \rho > 0$

- A hipótese nula será rejeitada se DW estiver "longe" de 2 (se DW for "pequeno").
- Em geral é preciso compara a estatística DW com DOIS valores críticos, d_L e d_U . A hipótese nula é rejeitada se $DW < d_L$. Se DW cai no intervalo entre d_L e d_U , o teste é inconclusivo, e se $DW > d_U$, a hipótese nula não é rejeitada.

Regressão Linear Múltipla



- **Autocorrelação dos resíduos**
- **Limitações da estatística DW**
 - Não usar se o modelo contém, como variável explicativa, a variável dependente defasada.
 - A estatística DW não mede o efeito de autocorrelações de ordem maior que 1. Também não captura efeitos MA nos erros.
 - O modelo estimado deve incluir uma constante.

Regressão Linear Múltipla



- **Referências – módulo de regressão**
 - Gujarati, D. (2003) – Basic Econometrics, McGraw-Hill, New York.
 - Pindyck, R.S. & Rubinfeld, D. (1998) - Econometric Models and Economic Forecasts, McGraw-Hill, New York.
 - Vasconcellos, M.A.S. & Alves, D. (editores) (2000) – Manual de Econometria: nível intermediário, Ed. Atlas, São Paulo
 - Weisberg, S. (1980) – Applied Linear Regression, John Wiley & Sons, New York.
 - Wooldridge, J. M. (2000) – Introductory Econometrics: A Modern Approach. South-Western College Publishing, a division of Thomson Learning.