FACULDADE DE SAÚDE PÚBLICA - USP

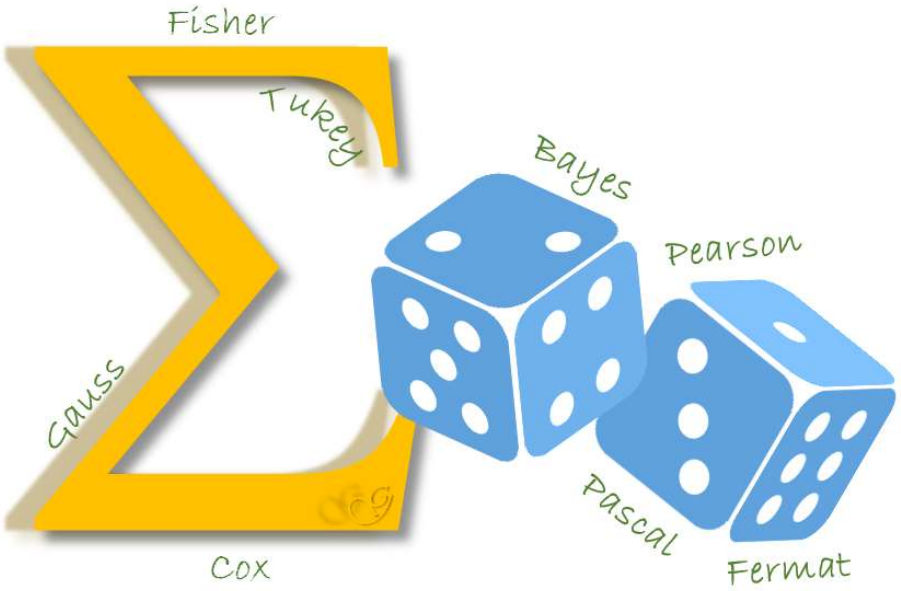
DEPARTAMENTO DE EPIDEMIOLOGIA

MODELOS DE REGRESSÃO APLICADOS EM EPIDEMIOLOGIA

Profa. Dra. MARIA DO ROSARIO DIAS DE OLIVEIRA LATORRE

Profa. Dra. GLEICE MARGARETE DE SOUZA CONCEIÇÃO

2021



MODELOS DE REGRESSÃO APLICADOS EM EPIDEMIOLOGIA

Profª. Drª. Maria do Rosário D O Latorre

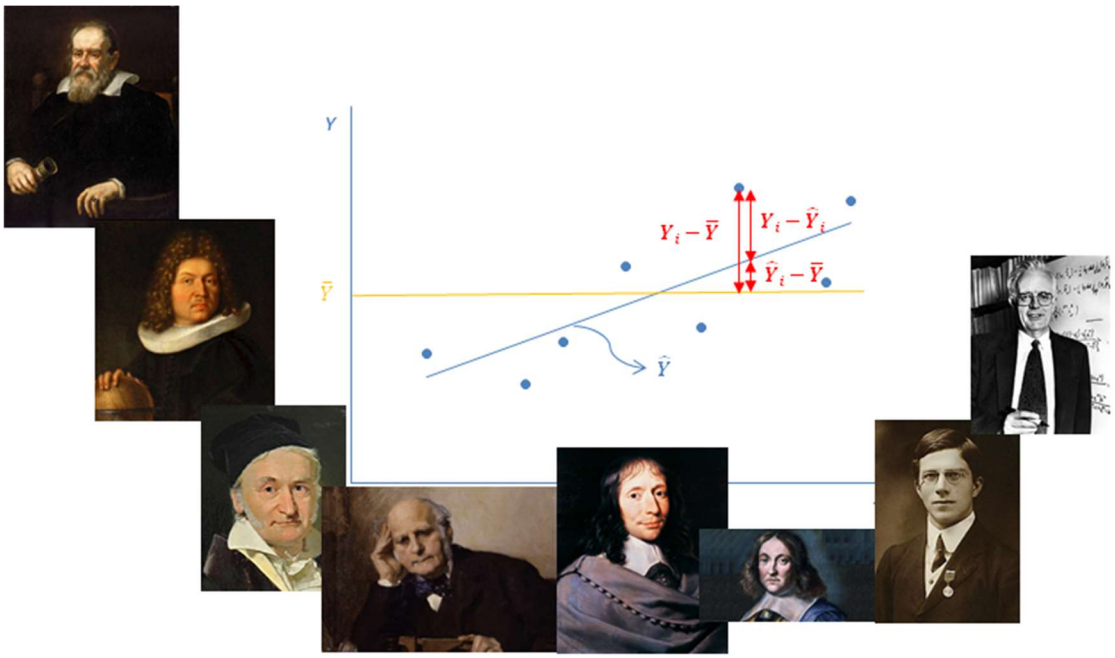
Profª. Drª. Gleice M S Conceição

Faculdade de Saúde Pública

Universidade de São Paulo

2021

MODELOS DE REGRESSÃO APLICADOS EM EPIDEMIOLOGIA



Profª. Drª. Maria do Rosário D O Latorre

Profª. Drª. Gleice M S Conceição

Faculdade de Saúde Pública

Universidade de São Paulo

2021

Programa

1. Introdução à análise de regressão
2. Noções de covariância e correlação
3. Modelo de regressão linear simples e múltipla:
    - Estimação dos parâmetros
    - Tabela de análise de variância (ANOVA)
    - Distribuições de probabilidades: Normal, t-Student, F-Snedecor e χ^2
    - Interpretação dos coeficientes
    - Análise dos resíduos
    - Teste F-parcial
    - Correlação parcial e múltipla
    - Variáveis indicadoras
    - Confusão e interação
    - Escolha do melhor modelo
4. Modelo de regressão polinomial
5. Análise de tendência em séries históricas usando modelos de regressão
6. Modelo de regressão logística simples e múltipla:
    - O modelo logístico
    - Estimação dos parâmetros
    - Interpretação dos coeficientes
    - Medidas de ajuste do modelo
    - Confusão e interação
    - Escolha do melhor modelo
    - Análise de resíduos

Bibliografia recomendada

1. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. **Applied regression analysis and other multivariable methods.** 3rd edition. Brooks/Cole Pub Co, Boston, 1997.
2. Curns AT, Mizam A. **Student solutions manual for Kleimbaum, Kupper, Muller and Nizam´s Applied regression analysis and other multivariable methods.** Brooks/Cole Pub Co, Boston, 1998.
3. Kutner MH, Christopher J. Nachtsheim CJ, Neter J, Li W. **Applied Linear Statistical Models.** 5ª ed. McGraw-Hill/Irwin, Boston, 2004.
4. Draper NR, Smith H. **Applied Regression Analysis.** John Wiley and Sons, 3rd edition. New York, 1998.
5. Kleinbaum DG, Klein M. **Logistic regression. A self-learning text**. 2nd edition. Springer-Verlag, New York, 2002.
6. Hosmer DW, Lemeshow S. **Applied logistic regression**. John Wiley and Sons, 2nd edition. New York, 2000.
7. Pereira MG. **Epidemiologia Teoria e Prática**. Rio de Janeiro: Editora Guanabara Koogan, 1999.
8. Laporta GZ, Latorre, MRDO. **Epidemiologia Aplicada via ambiente R.** Publicação independente, 2019.
9. Crawley MJ. **The R Book.** John Wiley & Sons Inc. Hoboken, 2012.
10. Wickham H, Grolemund G. **R for Data Science.** O'Reilly Media. Sebastopol, CA, 2017.

    Existe uma versão traduzida para o português:

    Wickham H, Grolemund G. **R para data science: Importe, arrume, transforme, visualize e modele dados.** Alta Books. Brasil, 2019.

História natural dos alunos do curso de Regressão...

1. Introdução à análise de regressão

Na prática há diversas situações em que a análise de regressão é apropriada:

1. Quando se deseja caracterizar a relação entre uma variável dependente (Y) e uma ou mais variáveis independentes (X_i), ié, avaliar a extensão, direção e força da relação (associação).
2. Procurar uma função matemática ou equação para descrever a variável dependente (Y) como função da variáveis independentes (X_i), ié, predizer Y em função dos X_i; determinando o melhor modelo estatístico que descreva essa relação.
3. Descrever quantitativa e/ou qualitativamente a relação entre os X_i e Y, controlando o efeito de outras variáveis (C_i).
4. Verificar o efeito interativo de 2 ou mais variáveis independentes às quais se relacionam com a variável dependente.
5. Determinar quais das muitas variáveis independentes são importantes para descrever ou predizer a variável dependente. Ordenar as variáveis independentes em sua ordem de importância em relação à variável dependente.
6. Comparar múltiplos relacionamentos derivados da análise de regressão.

É importante ser cauteloso sobre os resultados obtidos em uma análise de regressão, ou, de uma maneira mais geral, em qualquer análise utilizando técnicas estatísticas que procurem quantificar uma associação entre 2 ou mais variáveis.

A análise estatística pode estar correta, porém os dados podem estar viciados e/ou incompletos (vícios no delineamento, na amostragem, nas medidas, na escolha das variáveis e outros)

O achado de uma associação estatística significativa em um particular estudo não estabelece uma relação causal.

1.1.Questões Básicas

- Qual a função matemática mais apropriada a ser utilizada? (Em outras palavras: os dados se ajustam melhor a uma reta? A uma parábola? A uma função logística?)
- Como determinar o melhor modelo que se ajuste aos dados?
- Qual a validade e a precisão da(s) estimativa(s) do(s) coeficiente(s) de regressão?
- A presença, no modelo, de determinada variável independente melhora a precisão do mesmo?
- Dado um modelo específico, o que ele significa?

1.2.Estratégias (*stepwise*):

MODELO MAIS COMPLEXO ➔ MAIS SIMPLES

(BACKWARD SELECTION)

MODELO MAIS SIMPLES ➔ MAIS COMPLEXO

(FORWARD SELECTION)

2. Análise de Regressão Simples

- ✓ Duas variáveis quantitativas
- ✓ Descrever a relação entre elas
- ✓ Eventualmente, prever o valor de uma delas para um determinado indivíduo quando só conhecemos o valor da outra

Exemplos

- ✓ Tempo de reação a um estímulo e idade
- ✓ Perda de peso e concentração de uma determinada substância
- ✓ Peso e idade
- ✓ Peso e altura
- ✓ Número de óbitos e concentração de um determinado poluente

✓ Variável resposta, dependente ou preditiva

A variável que está sendo afetada pela outra ou outras, que acreditamos depender das outras, que pode ser explicada ou prevista pelas outras.

✓ Variável explicativa, independente ou preditora

A variável que afeta a outra, que pode ajudar a explicar a variabilidade da outra e a prever a outra.

Se o estudo envolver apenas uma variável explicativa, o método será chamado de Regressão Linear Simples.

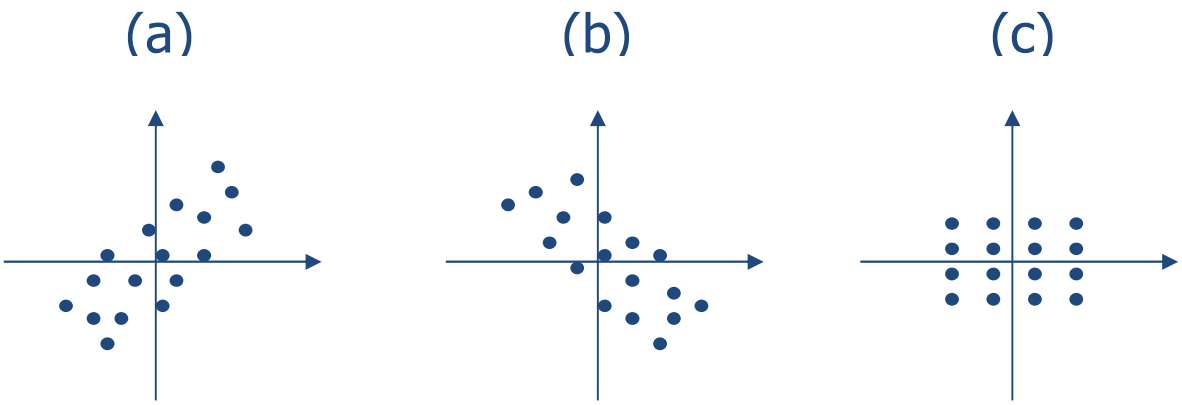
Exemplo 1

Amostra de crianças (dados hipotéticos)

Criança	Peso (libras)	Idade (anos)	Altura (pés)
1	64	8	57
2	71	10	59
3	53	6	49
4	67	11	62
5	55	8	51
6	58	7	50
7	77	10	55
8	57	9	48
9	56	10	42
10	51	6	42
11	76	12	61
12	68	9	57

2.1. Coeficiente de correlação linear de Pearson

Tipos de associação entre variáveis



Para quantificar esta associação, podemos utilizar algo do tipo:

$$X \cdot Y \text{ ou } \frac{X \cdot Y}{n}$$

O coeficiente de correlação linear de Pearson pode ser escrito como

$$r = corr(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(X_i - \bar{X})}{S_X}\frac{(Y_i - \bar{Y})}{S_Y}$$

Isto é, o coeficiente de correlação é a média dos produtos dos valores padronizados das variáveis X e Y.

Na verdade, a definição formal para o coeficiente de correlação linear de Pearson é:

$$r = corr(X,Y) = \frac{cov(X,Y)}{S_X S_Y}$$

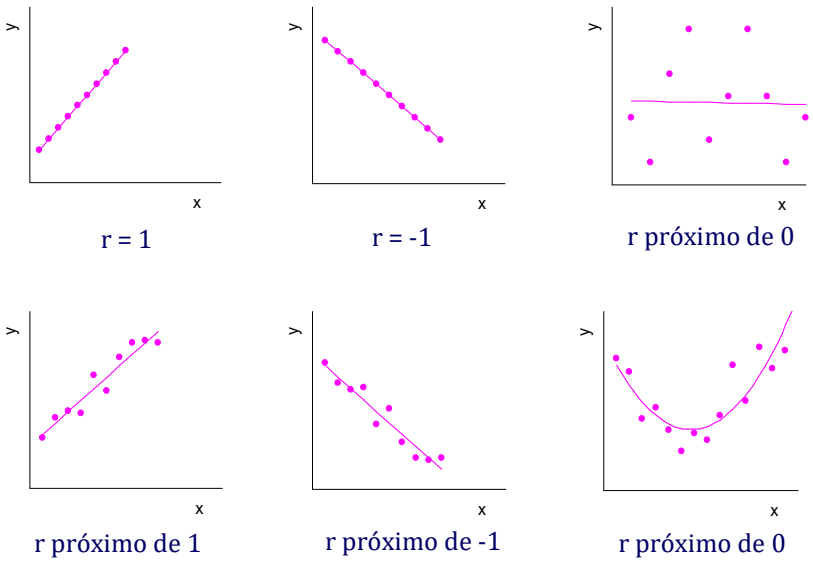
$$\text{onde } cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{n-1}$$

O coeficiente de correlação de Pearson:

- ✓ Assume valores entre -1 e 1 ($-1 \leq corr(X,Y) \leq 1$)
- ✓ Valores próximos de 1 ou –1 indicam uma associação forte
- ✓ Valores próximos de zero quando não existe associação

O coeficiente de correlação mede:

- ✓ Presença de associação linear
- ✓ Força de uma associação linear



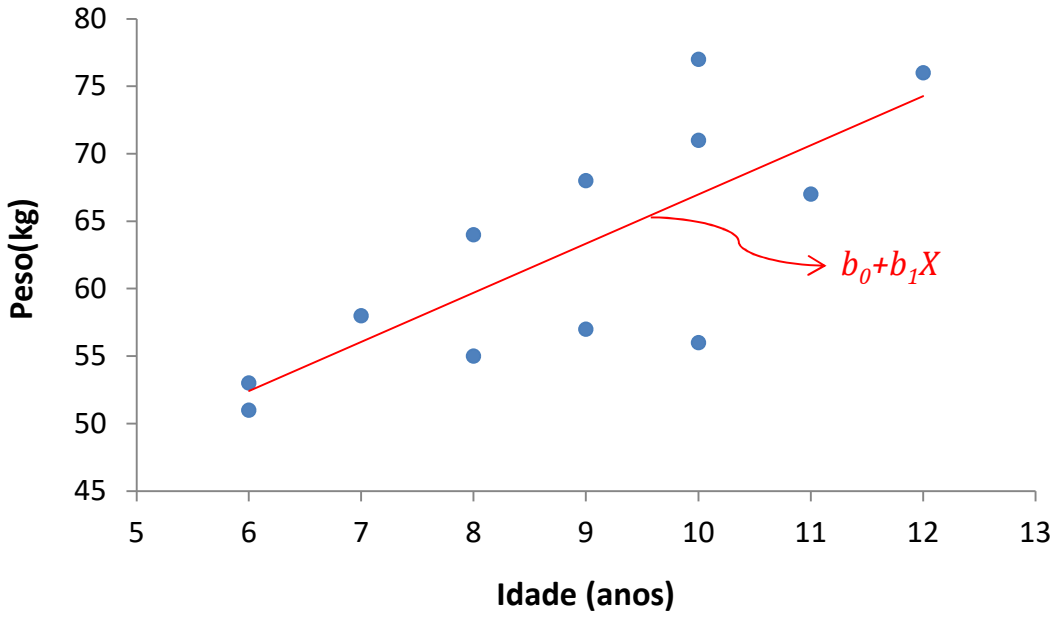
Alguns autores sugerem avaliar a presença de associação linear a partir do coeficiente de correlação do seguinte modo:

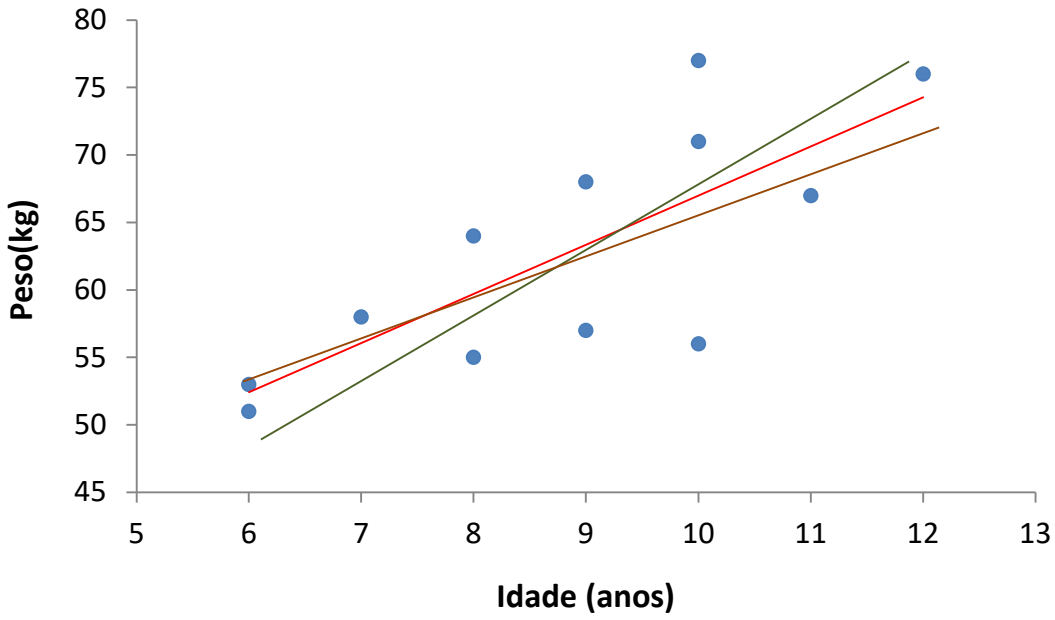
- de 0,10 a 0,39 - fraca
- de 0,40 a 0,69 - moderada
- de 0,70 até 1 - forte

Mas não há, de fato, uma norma rígida sobre isto.

Deve-se levar em conta o contexto, o tamanho da amostra e sempre avaliar a associação observando conjuntamente o coeficiente de correlação e o diagrama de dispersão.

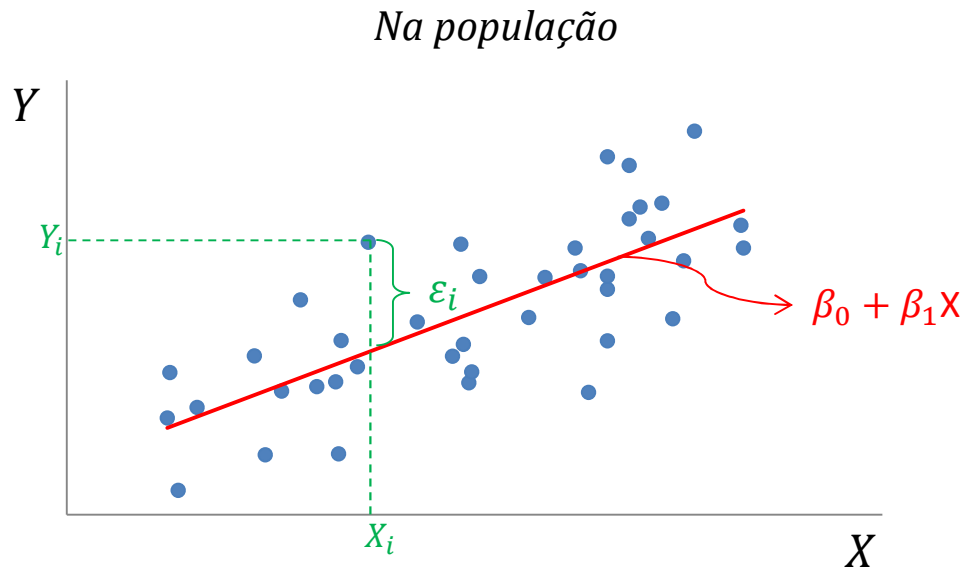
2.2. Ajustando uma reta aos dados





Como escolher a melhor reta?

2.3. O modelo de regressão linear simples



O modelo de regressão linear simples pode ser escrito como:

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

onde

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

β_0 *e* β_1 são parâmetros;

X_i é uma constante conhecida, o valor da variável preditora (ou explicativa) na i-ésima observação;

ε_i é um erro aleatório não observável;

$\varepsilon_i \sim N(0, \sigma^2)$

(ε_i , ε_i) independentes para todo *i, j*.

i = 1, ..., n

Este modelo é

- ✓ simples,
- ✓ linear nos parâmetros e
- ✓ linear na variável preditora.

✓ Obtendo a esperança e a variância de Y:

$$E(Y_i) = \mu_i = E(\beta_0 + \beta_1 X_i + \varepsilon_i) = E(\beta_0) + E(\beta_1 X_i) + E(\varepsilon_i) = \beta_0 + \beta_1 X_i$$

$$Var(Y_i) = Var(\beta_0 + \beta_1 X_i + \varepsilon_i) = Var(\beta_0) + Var(\beta_1 X_i) + Var(\varepsilon_i) = \sigma^2$$

✓ Como Y_i é a soma de uma constante $(\beta_0 + \beta_1 X_i)$ e uma variável aleatória (ε_i) com distribuição Normal, onde $(\varepsilon_i, \varepsilon_j)$ são independentes para todo *i, j*, então, Y_i~Normal com (Y_i, Y_j) independentes para todo *i, j*, .

✓ Assim, o modelo equivale a

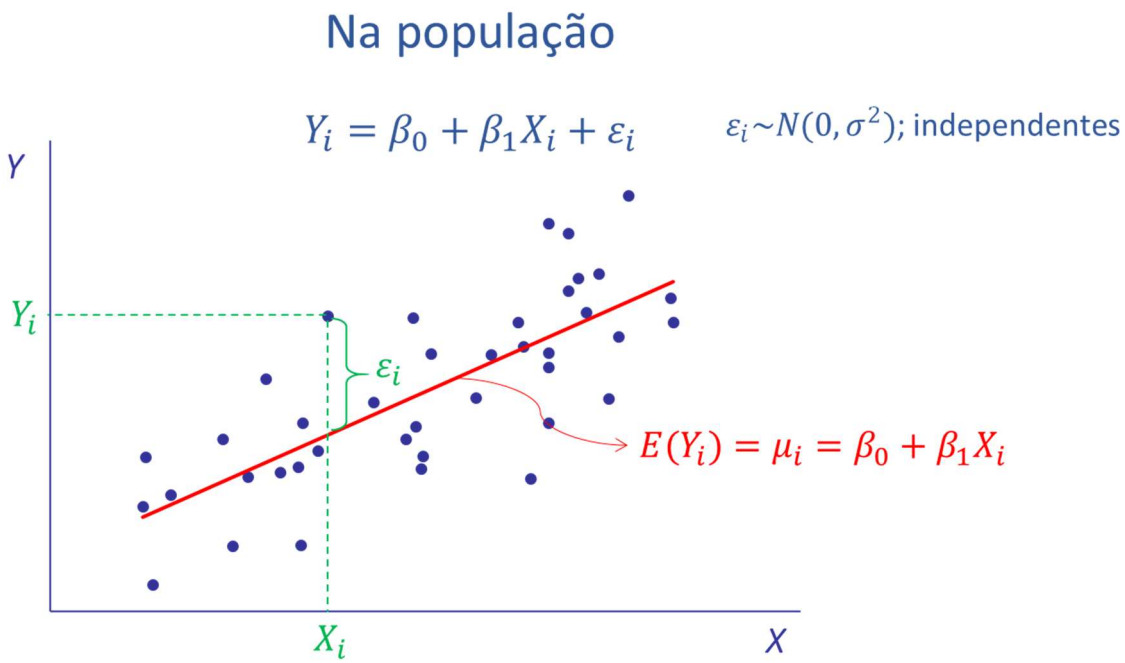
$$E(Y_i) = \mu_i = \beta_0 + \beta_1 X_i \qquad (2)$$

$$onde\ Y_i \sim N(\mu_i;\ \sigma^2), e\ são\ independentes$$

Como X_i é constante, também é comum utilizar a notação

$$E(Y_i|X_i) = \mu_i = \beta_0 + \beta_1 X_i \qquad (3)$$

$$onde\ Y_i|X_i \sim N(\mu_i;\ \sigma^2), e\ são\ independentes$$



2.4.Entendendo as suposições do modelo

1. Distribuição Normal para a variável resposta

Para um valor fixo de X, Y é uma v.a. com distribuição normal.

Então, há uma distribuição Normal para Y em cada nível de X.

$$Y_i/X_i \sim N(\mu_i; \sigma^2)$$

2. Os valores de Y são independentes uns dos outros.

(Y_i, Y_j) são independentes para todo i, j.

3. Linearidade

O valor esperado ou médio de Y_i/X_i , que é μ_i,

é uma função de linha reta sobre os X_i.

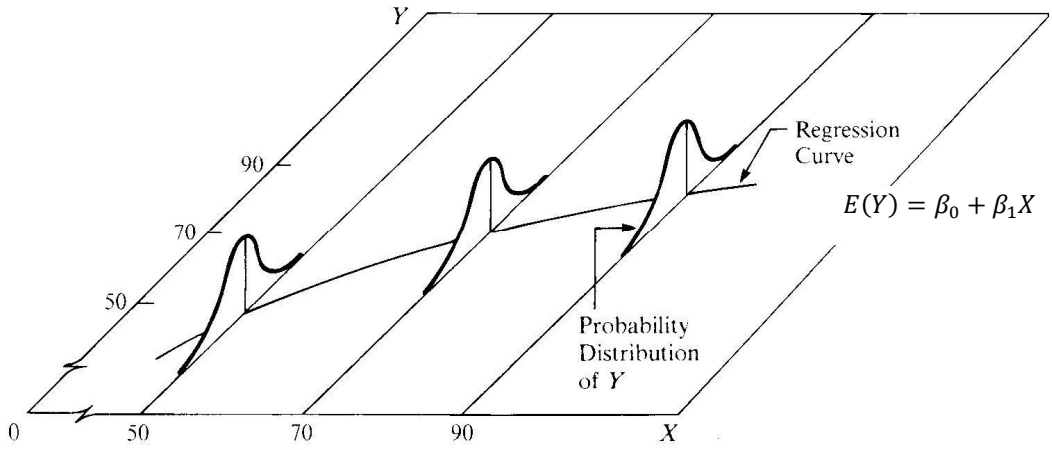
4. Homocedasticidade

A variância de Y_i/X_i é a mesma, qualquer que seja X_i.

$Var(Y_i/X_i) = \sigma^2$, constante.

Representação gráfica do modelo de regressão linear simples

Observe as suposições de normalidade, linearidade e homocedasticidade:



Uma ilustração da presença de heterocedasticidade:

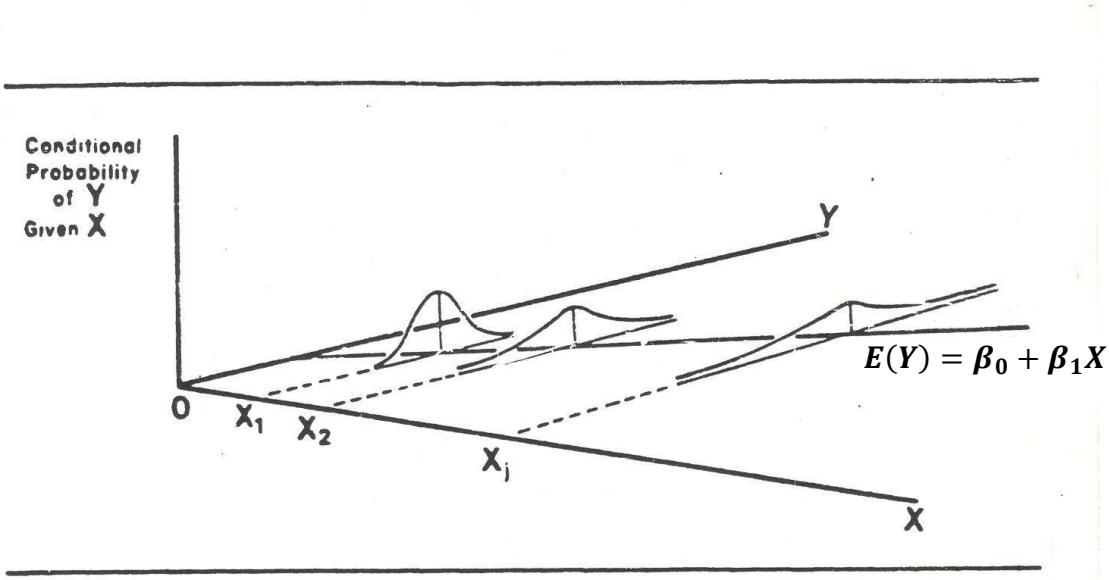


Figure 6.1: An Illustration of a Heteroscedastic Error Term Distribution for a Bivariate Regression Model: COV[VAR(ϵ), X] > 0

2.5. Interpretação dos parâmetros de modelo

$$E(Y) = \mu = \beta_0 + \beta_1 X \qquad (2)$$

Para interpretar os parâmetros do modelo, variamos X e vemos o que acontece com μ.

✓ Fazendo X=0:

$X = 0 \Rightarrow \mu_0 = \beta_0 + \beta_1 * 0 = \beta_0$

$\Rightarrow \beta_0$ é o valor de μ quando $X = 0$

✓ Aumentando X de uma unidade:

$\mu_{(X+1)} = \beta_0 + \beta_1(X + 1) = \beta_0 + \beta_1 X + \beta_1 = \mu + \beta_1$

$\Rightarrow \beta_1$ é o aumento em μ (isto é, o aumento esperado ou médio em Y)

quando aumentamos X de uma unidade.

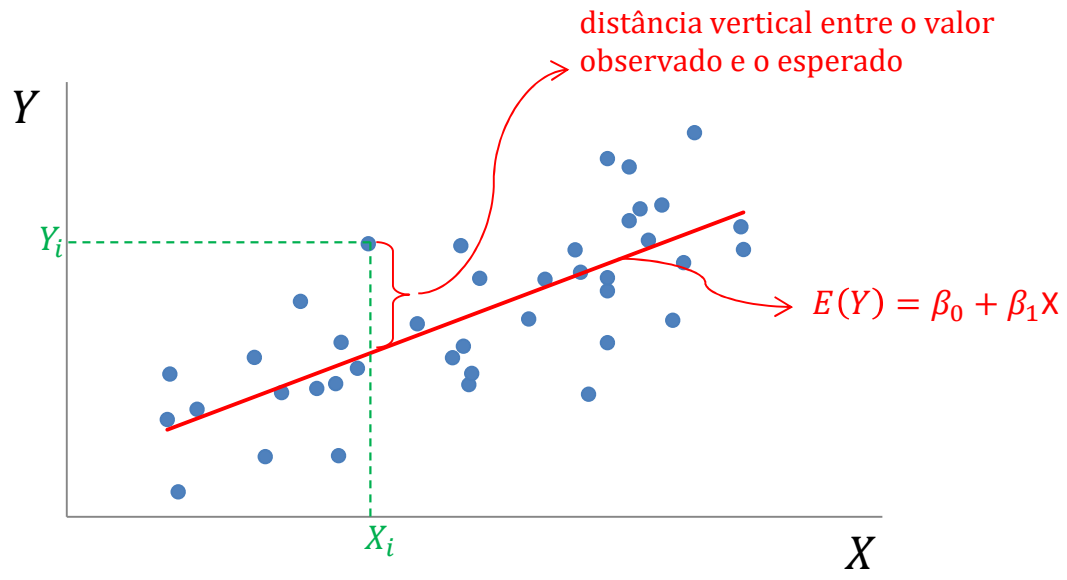
2.6.Estimando os parâmetros do modelo

- ✓ Como escolher a melhor reta?
- ✓ Como estimar os valores de β_0 e β_1?

✓ Métodos de estimação

- Método de Mínimos Quadrados
- Método de Máxima Verossimilhança

Método de Mínimos Quadrados



Para cada observação, consideramos a distância entre o valor de Y que foi observado e o valor de Y esperado (ou médio, aquele que é previsto pela reta)

$$Y_i - E(Y_i) =$$
$$Y_i - (\beta_0 + \beta_1 X_i)$$

E procuramos valores de β_0 e β_1 que minimizam essa distância:

$$Q = \sum [Y_i - (\beta_0 + \beta_1 X_i)]^2 = \sum [Y_i - \beta_0 - \beta_1 X_i]^2$$

É possível obter o ponto de mínimo da função Q utilizando derivadas:

$\frac{\partial Q}{\partial \beta_0} = 0 \ \ e \ \ \frac{\partial Q}{\partial \beta_1} = 0$ (não se preocupem com isto, vamos diretos aos resultados!)

Os estimadores de β_0 e β_1 serão os valores b_0 e b_1 para os quais Q é a menor possível na amostra que está sendo considerada:

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

Outra forma de cálculo:

$$b_1 = \frac{\sum_{i=1}^{n} X_i y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} {X_i}^2 - n\bar{X}^2}$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

A reta estimada será

$$\hat{Y}_i = b_0 + b_1 X_i \qquad (4)$$

Note que, substituindo-se o valor de b_0 na equação acima, temos:

$$\hat{Y}_i = b_0 + b_1 X_i$$

$$\hat{Y}_i = \bar{Y} - b_1\bar{X} + b_1 X_i$$

$$\hat{Y}_i = \bar{Y} + b_1(X_i - \bar{X})$$

Isso significa que quando $X_i \to \bar{X} \Rightarrow \hat{Y}_i \to \bar{Y}$

Interpretação dos coeficientes

✓ b_0

$\hat{Y} = b_0 + b_1 X$

Se $X = 0 \Rightarrow \hat{Y} = b_0$

Então, $\mathbf{b_0}$ é o valor esperado (ou médio) de Y quando X=0, é o intercepto da reta ajustada.

✓ b_1

Se aumentarmos X em uma unidade

$\hat{Y}_{novo} = b_0 + b_1(X + 1) = b_0 + b_1 X + b_1$

$\hat{Y}_{novo} = \hat{y} + b_1$

Então, $\mathbf{b_1}$ é o aumento esperado (ou médio) em y quando aumentamos X de 1 unidade, é o "efeito" de X em Y.

2.7.Resíduos

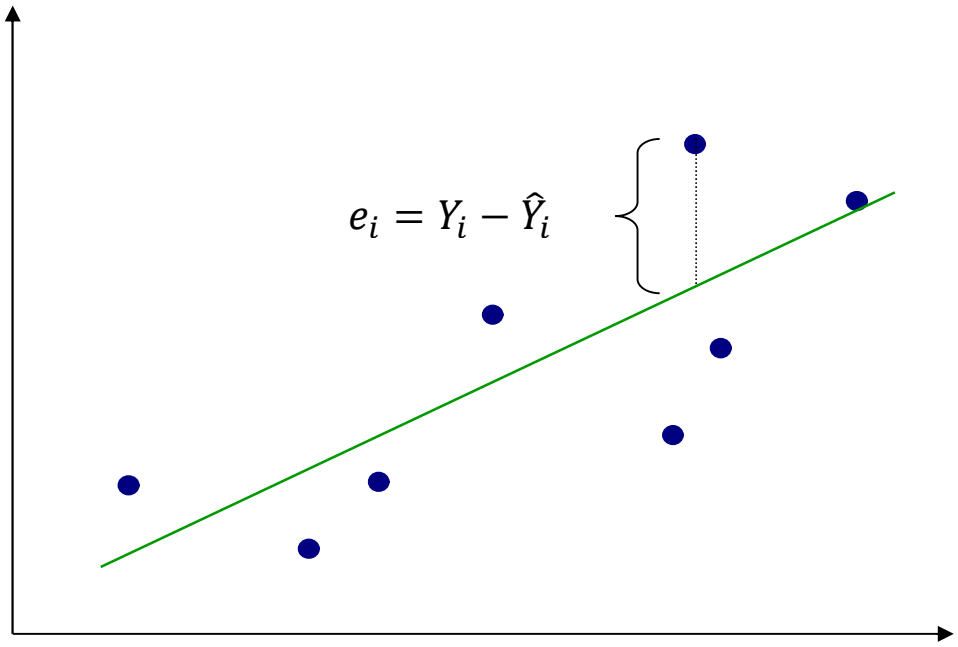
O modelo ajustado é:

$$\hat{Y}_i = b_0 + b_1 X_i \qquad (4)$$

Os resíduos do modelo são dados por:

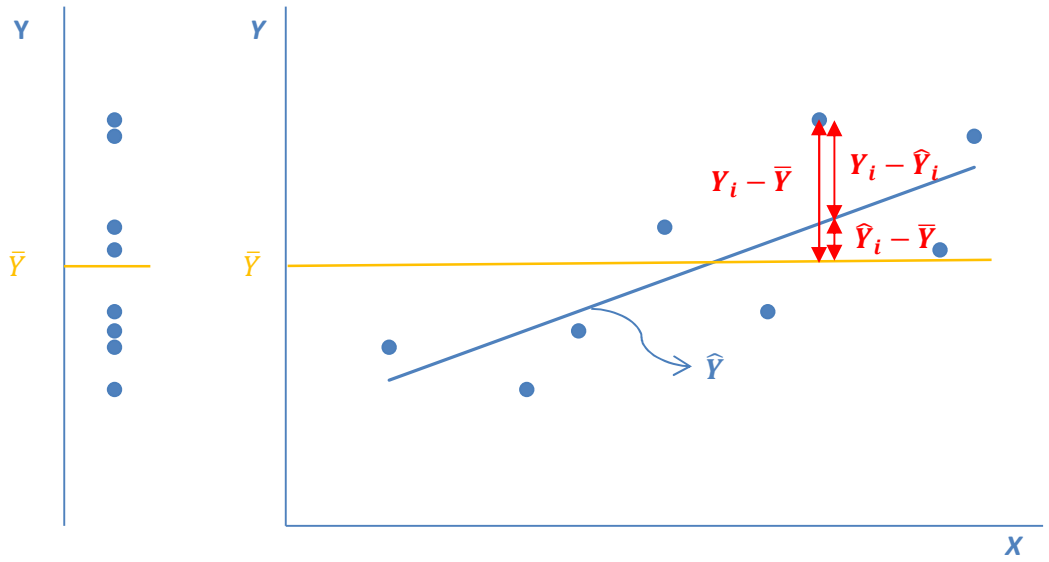
$$e_i = Y_i - \hat{Y}_i \qquad i = 1, \ldots, n \qquad (5)$$

O resíduo é a distância entre o valor de Y observado e o valor de Y ajustado pelo modelo.



3. Inferências para o modelo de regressão linear simples

3.1.Fontes de variabilidade



$$Y_i - \bar{Y} = \left(Y_i - \hat{Y}_i\right) + \left(\hat{Y}_i - \bar{Y}\right) \qquad (6)$$

3.2.Partição da Soma de Quadrados

$$Y_i - \bar{Y} = \left(\hat{Y}_i - \bar{Y}\right) + (Y_i - \hat{Y}_i) \qquad (6)$$

Elevando-se ao quadrado os 2 lados da igualdade acima e fazendo-se a soma de todas as n equações (i=1,2, ...,n), obtem-se:

$$\underset{\downarrow}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \underset{\downarrow}{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2} + \underset{\downarrow}{\sum_{i=1}^{n}(Y_i - \hat{Y})^2} \qquad (7)$$

$$SQT \quad = \quad SQM \quad + \quad SQR$$

- A Soma de Quadrados Total (SQT) é dada por

$$SQT = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

e fornece uma medida da variabilidade total das observações em relação à média geral.

O número de graus de liberdade (g.l.) associado à SQT é (n-1).

- A Soma de Quadrados do Modelo de Regressão (SQM) é dada por

$$SQM = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

e fornece uma medida da variabilidade entre a média geral e a média estimada pela reta de regressão. Quanto mais distante a reta de regressão estiver da média geral, maior será a contribuição do modelo para explicar a variabilidade de Y e maior será SQM.

O número de graus de liberdade (g.l.) associado à SQM é 1 (o número de parâmetros menos 1).

- A Soma de Quadrados dos Resíduos (SQR) é dada por

$$SQR = \sum_{i=1}^{n} (Y_i - \hat{Y})^2$$

e fornece uma medida da variabilidade das observações em torno da reta de reta de regressão. Quanto mais próximas as observações estiverem da reta, menor será SQR.

O número de graus de liberdade (g.l.) associado à SQR é (n-2).

A equação (7) é chamada a EQUAÇÃO FUNDAMENTAL DA REGRESSÃO e postula que:

$$SQT = SQM + SQR$$

Isto é, a soma dos quadrados sobre a média (SQT) = soma de quadrados devida à regressão (SQM) + soma de quadrados sobre a regressão (SQR).

Isso significa que a variação total dos $Y's$ sobre sua média pode ser explicada, em parte, pela linha de regressão e, em parte, pelos resíduos. Se todos os Y's caíssem sempre na linha de regressão a SQR seria zero!!

Portanto, quanto mais a SQM for próxima da SQT, mais a reta de regressão explica a variabilidade total de Y.

Quadrados médios

Dividindo SQM e SQR pelos correspondentes graus de liberdade obtemos, respectivamente, o quadrado médio da regressão (QMM) e o quadrado médio dos resíduos (QMR), isto é:

$$QMM = \frac{SQReg}{1} \qquad QMR = \frac{SQR}{n-2}$$

3.3.Teste de hipóteses para o modelo

$H_0: \beta_1 = 0$ (Não existe associção entre X e Y)

$H_1: \beta_1 \neq 0$ (Existe associção entre X e Y)

Estatística do teste

Pode-se demonstrar que

$$E(QMM) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$E(QMR) = \sigma^2$$

- ✓ Se H_0 for verdadeira, β_1 será igual a zero (ou seja, não existe associação entre X e Y), $E(QMM)$ e $E(QMR)$ serão iguais.
- ✓ Neste caso, espera-se que o quociente $\frac{QMM}{QMR}$ seja próximo de 1.
- ✓ Um valor observado para $\frac{QMM}{QMR}$ próximo de 1 é uma indicação de que Ho é verdadeira.
- ✓ Um valor grande para esse quociente é uma indicação de que Ho é falsa.
- ✓ O que é um valor grande?
- ✓ A estatística $\frac{QMM}{QMR} \sim F_{(1,n-2)}$, basta usar a tabela da F.
- ✓ Note que QMR é um estimador não viesado da variância σ^2.

3.4.Quadro de Análise de Variância

Fonte de variação	*g.l.*	*SQ*	*QM*	*E(QM)*	F_0	*p-valor*
Regressão	*1*	*SQM*	*QMM*	$\sigma^2 + \beta_1^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2$	$\frac{QMM}{QMR} \sim F_{(1,n-2)}$	
Resíduo	*n-2*	*SQR*	*QMR*	σ^2		
Total	*n-1*	*SQT*				

Para facilitar os cálculos:

$$SQM = b_1 \left[\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n} \right]$$

$$SQT = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n}$$

SQR: por subtração

3.5.Coeficiente de determinação ou explicação do modelo (R^2)

$$R^2 = \frac{SQM}{SQT}$$

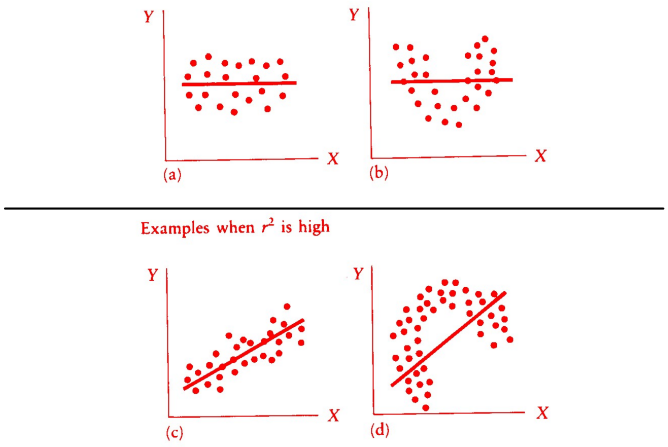
- O R^2 Mede a proporção da variabilidade total que é explicada pelo modelo adotado.
- $0 \leq R^2 \leq 1$

- Quanto mais próximo de 1 estiver o R^2, mais X contribui para explicar a variabilidade de Y e para prever Y.
- Se $R^2 = 1$ e aceitarmos que $\beta_1 \neq 0$, todos os pontos cairiam em cima da reta e o ajuste seria perfeito (altamente improvável!!!).
- Por outro lado, se R^2 está próximo de 0 e aceitarmos que $\beta_1 = 0$, X não contribui para explicar a variabilidade de Y, ou para prever Y.
- No modelo de Regressão Linear Simples, o R^2 é igual ao coeficiente de correlação (r) ao quadrado, isto é, $R^2 = r^2$.
- Sua interpretação do exige cautela. Da mesma forma que o coeficiente de correlação, deve ser observado em conjunto com outras ferramentas, como o diagrama de dispersão, por exemplo.

O que R^2 não mede:

- a magnitude da inclinação de uma reta de regressão
- a linearidade da relação entre Y e X
- se o modelo está bem ajustado (quem faz isto é a análise de resíduos)

quando r2 é baixo



3.6.Estimador para σ^2

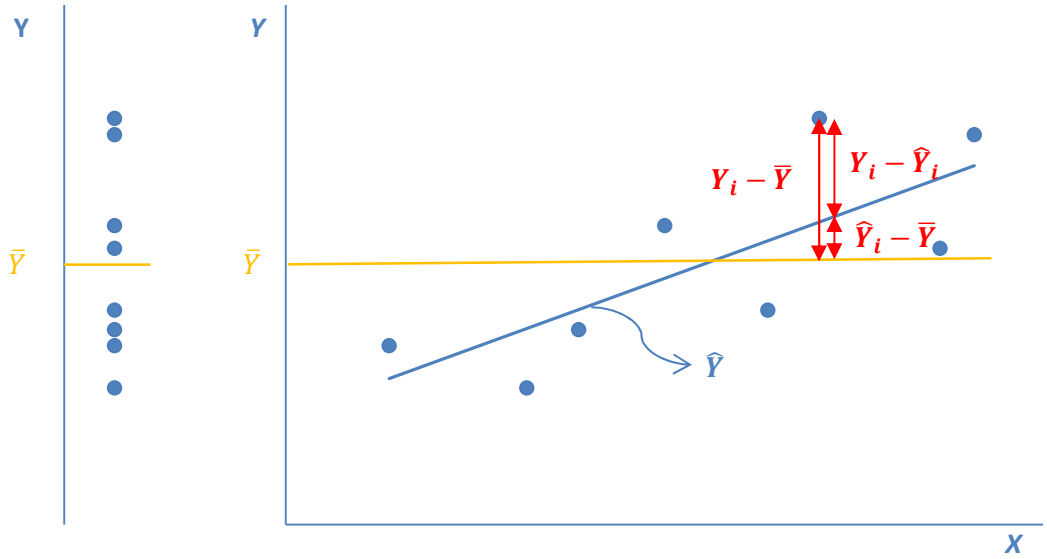
A figura abaixo mostra a representação de dois modelos distintos para Y.

$Y_i = \bar{Y} + \varepsilon_i$ ou

$E(Y_i) = \mu_i = \bar{Y}$

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ ou

$E(Y_i) = \mu_i = \beta_0 + \beta_1 X_i$



No primeiro modelo, explicamos Y apenas pela sua média.

No segundo, explicamos Y levando em conta a idade.

Qual deles explica mais o comportamento de Y?

Qual deles apresenta a menor variabilidade em torno da média?

No primeiro modelo, a variabilidade de Y em torno de sua média, é estimada por

$$S_Y^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$$

que é um estimador não viesado para $VAR(Y)$.

No segundo modelo, a variabilidade de Y em torno de sua média (que é a reta), é estimada por

$$S_Y^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{n-2} = QMR$$

De fato, o quadrado médio do resíduo (QMR) que é um estimador não viesado para $VAR(Y) = \sigma^2$, isto é:

$$S_Y^2 = QMR$$

3.7.Inferências acerca de β_1

3.7.1.Teste de hipóteses para β_1

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

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 X_i \qquad (2)$$

$H_0: \beta_1 = 0$ (Não existe associção entre X e Y)

$H_1: \beta_1 \neq 0$ (Existe associção entre X e Y)

Note que, se $\beta_1 = 0$, a reta de regressão é paralela ao eixo X e o modelo fica

$$Y_i = \beta_0 + \varepsilon_i$$

$$E(Y_i) = \mu_i = \beta_0$$

O estimador para β_1 é

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Pode-se demonstar que b_1 tem distribuição Normal com

$$E(b_1) = \beta_1$$

$$VAR(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Como não conhecemos σ^2, nós o substituímos por $S_Y^2 = QMR$, e o estimador para a $VAR(\beta_1)$ fica

$$S_{\beta_1}^2 = \frac{QMR}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

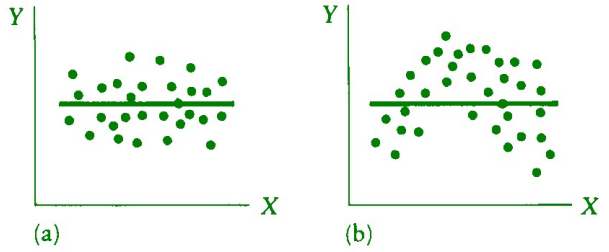
De modo que:

$$\frac{b_1 - \beta_1}{S_{\beta_1}} \sim t_{n-2}$$

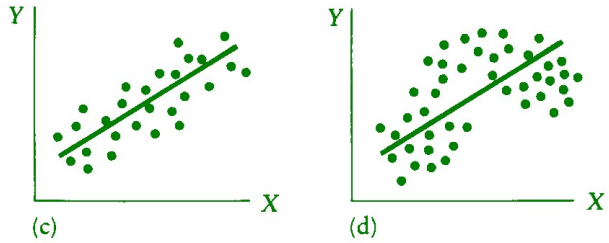
Como, sob H_0, $\beta_1 = 0$, a estatística do teste será

$$\frac{b_1}{S_{\beta_1}} \sim t_{n-2}$$

H0:B1=0 não é rejeitada



Examples when H_0: $\beta_1 = 0$ is rejected



3.7.2. Intervalo de confianca para β_1

Lembrando que o formato usual para um intervalo de confiança de um parâmetro é:

Estimador do parâmetro	$\mp$	Quantil de uma distribuição	X	Desvio padrão do estimador

O intervalo de confiança para β_1, com coeficiente de confiança $(1-\alpha)$, será:

$$IC(\beta_1; 1-\alpha) = b_1 \mp t_{(1-\alpha;\, n-2)} S_{\beta_1}$$

3.8.Inferências acerca de β_0

Existem poucas situações nas quais desejamos fazer inferências sobre β_0. Elas ocorrem, basicamente, as situações nas quais faz sentido que X assuma o valor

zero e, preferencialmente, quando a faixa de valores observados de X inclui o valor zero.

3.8.1.Teste de hipóteses para β_0

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (1)$$

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 X_i \quad (2)$$

$H_0: \beta_0 = 0$

$H_1: \beta_0 \neq 0$

O estimador para β_0 é

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Pode-se demonstar que b_0 tem distribuição Normal com

$$E(b_0) = \beta_0$$

$$VAR(b_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$

Como não conhecemos σ^2, nós o substituímos por $S_Y^2 = QMR$, e o estimador para a $VAR(\beta_0)$ fica

$$S_{\beta_0}^2 = QMR \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$

De modo que:

$$\frac{b_0 - \beta_0}{S_{\beta_0}} \sim t_{n-2}$$

Como, sob H_0, $\beta_0 = 0$, a estatística do teste será

$$\frac{b_0}{S_{\beta_0}} \sim t_{n-2}$$

3.8.2. Intervalo de confianca para β_0

O intervalo de confiança para β_0, com coeficiente de confiança $(1-\alpha)$, será:

$$IC(\beta_1; 1-\alpha) = b_0 \mp t_{(1-\alpha;\, n-2)} S_{\beta_0}$$

3.9.Estimação e Predição

Os principais objetivos da análise de regressão são a descrição e a previsão.

A reta de regressão ajustada $(\hat{Y}_i = b_0 + b_1 X_i)$ fornece a descrição da relação linear entre a variável resposta e a explicativa e quantifica, por meio dos coeficientes estimados, a velocidade com a qual Y varia a partir de X.

É possível fazer previsões sobre o valor de Y para um dado valor de X a partir da reta de regressão ajustada.

Existem duas situações de interesse para as quais desejamos fazer previsões e obter intervalos de confiança.

A primeira envolve a previsão do valor médio de Y para um dado nível de X, isto é, a previsão de $E(Y/X)$ ou μ, dada por $\hat{Y}$, que pertence à reta de regressão estimada.

A segunda envolve a previsão dos possíveis valores de Y que podem ser observados (e não a sua média) em um dado nível de X, isto envolve a distribuição de probabilidades de Y em torno da média μ.

3.9.1. Estimação

Vamos chamar de X_h o valor de X para o qual queremos estimar a resposta média μ_h, ou $E(Y_h)$. X_h pode ser um valor observado de X na amostra ou qualquer outro valor de X (observado ou não), preferencialmente dentro da faixa dos valores de X na amostra.

O estimador pontual de μ_h é, obviamente,

$$\hat{Y}_h = b_0 + b_1 X_h.$$

O estimador por intervalo de μ_hé chamado de Intervalo de Estimação e é dado por:

$$IC(\mu_h; 1-\alpha) = \hat{Y}_h \mp t_{(1-\frac{\alpha}{2};n-2)} S_{\hat{Y}_h}$$

$$S^2_{\hat{Y}_h} = QMR\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

Note que a variância de $\hat{Y}_h$ é proporcional a $X_h - \bar{X}$, que é a distância de X_h em relação à média de X, ou seja, quanto mais próximo X_h estiver de $\bar{X}$, menor será a variância do estimador $\hat{Y}_h$ e menor será a a amplitude do seu intervalo de confiança. Analogamente, quanto mais distante X_h estiver de $\bar{X}$, maior será a variância do estimador $\hat{Y}_h$ e maior a amplitude do seu intervalo de

confiança. Isto significa que as estimativas de $\hat{Y}$ na reta de regressão vão se tornando menos precisas à medida que X se distancia de $\bar{X}$.

O teste de hipóteses para μ_h é da forma

$H_0: \mu_h = \mu_0$
$H_1: \mu_h \neq \mu_0$

A estatística do teste será:

$$t = \frac{\hat{Y}_h - \mu_0}{S_{\hat{Y}_h}} \sim t_{n-2}$$

3.9.2. Predição

Agora estamos interessados na predição de um possível valor de Y que possa ser observado (e não a sua média) em um dado nível de X.

Novamente, seja X_h o valor de X para o qual desejamos prever um possível valor para Y. Este valor deve pertencer à distribuição de Y/X_h. Vamos denotar este novo valor por $Y_{h(novo)}$.

O estimador pontual de $Y_{h(novo)}$ é,

$$\hat{Y}_{h(novo)} = b_0 + b_1 X_h.$$

O estimador por intervalo de μ_h é chamado de Intervalo de Estimação e é dado por:

O intervalo que delimita, com $(1-\alpha)\%$ de confiança, onde os possíveis valores de Y pertencentes à distribuição Normal com média μ_h podem ocorrer é chamado Intervalo de Predição e é dado por:

$$IC(Y_h; 1-\alpha) = \hat{Y}_{h(novo)} \mp t_{(1-\frac{\alpha}{2};n-2)} S_{pred}$$

$$S^2_{pred} = QMR\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

Note que este intervalo é maior do que o anterior, já que pretende englobar não só a média μ_h, mas a distribuição dos valores de Y em torno da média μ_h. Novamente, quanto mais próximo X_h estiver de $\bar{X}$, menor será a variância do estimador $\hat{Y}_{h(novo)}$ e menor será a a amplitude do seu intervalo de confiança, e vice-versa.

O teste de hipóteses para $Y_{h(novo)}$ é da forma

$H_0: Y_{h(novo)} = Y_{h0}$
$H_1: Y_{h(novo)} \neq Y_{h0}$

A estatística do teste será:

$$t = \frac{\hat{Y}_{h(novo)} - Y_{h0}}{S_{pred}} \sim t_{n-2}$$

3.9.3. Bandas de confiança para a reta de regressão

Frequentemente, estamos interessados em obter bandas de confiança para toda a reta de regressão. Tais bandas permitem visualizar toda uma região

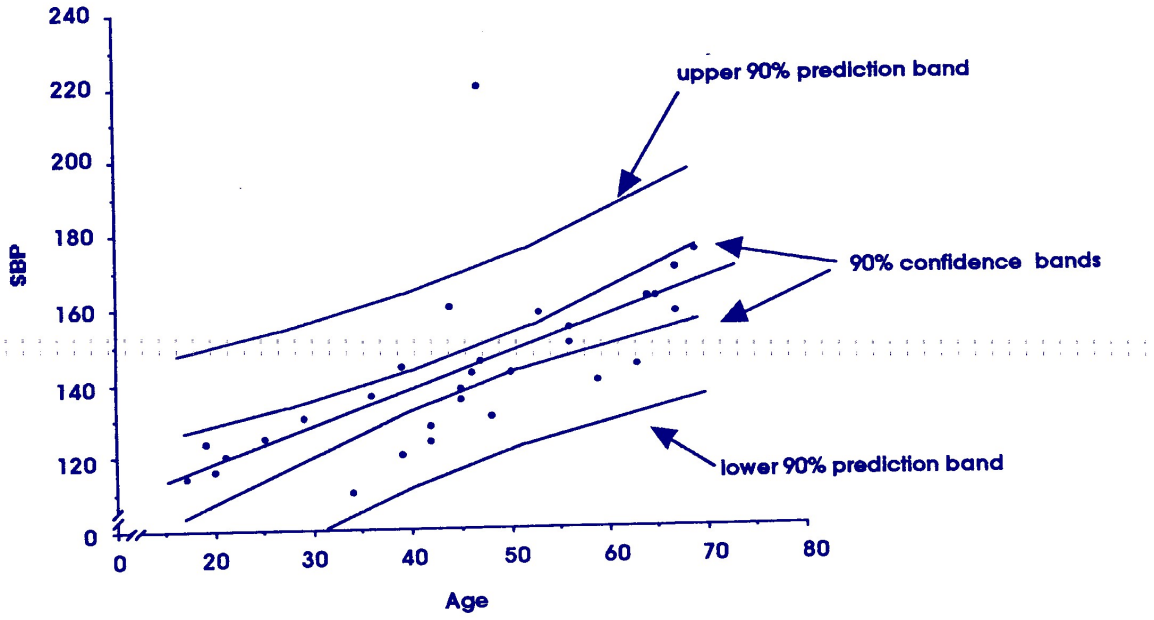
onde a verdadeira reta de regressão $(E(Y_i) = \mu_i = \beta_0 + \beta_1 X_i)$ poderia estar, isto é, a região que com probabilidade $(1 - \alpha)$ contem a verdadeira reta.

Estas bandas são dadas por

$$\hat{Y}_h \mp W S_{\hat{Y}_h}$$

$$W^2 = 2F_{(1-\alpha;2,n-2)}$$

Note que a fórmula para as bandas de confiança é parecida com a do intervalo de estimação da reposta média μ_h para um dado valor X_h, exceto que a distribuição t foi substituida pela F. Com isto, as bandas terão amplitude maior do que o intervalo de estimação, o que faz sentido, já que devem compreender toda a reta de regressão e não apenas um único ponto, como no caso do intervalo de estimação.



4. O coeficiente de correlação e a análise de regressão

4.1. Definição

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

onde σ_{XY} é a covariância entre X e Y, definida como

$$\sigma_{XY} = E[X - E(X)][Y - E(Y)]$$

Anteriormente, aprendemos sobre r que, na verdade, é o estimador de ρ, obtido a partir da amostra.

$$\hat{\rho} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y} = \mathrm{r} = \frac{cov(X,Y)}{S_X S_Y} = \frac{1}{n-1} \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{S_X S_Y}$$

Não é difícil mostrar que

$$\rho = \frac{\sigma_X}{\sigma_Y}\beta_1 \;\Rightarrow\; \beta_1 = \rho\frac{\sigma_Y}{\sigma_X}$$

Note que o sinal de ρ é o mesmo de β_1.

O mesmo vale para o estimador de ρ:

$$\hat{\rho} = r = \frac{S_X}{S_Y} b_1 \;\Rightarrow\; b_1 = \rho\frac{S_Y}{S_X}$$

O sinal de r é o mesmo de b_1.

4.2. Teste de hipótese para ρ

$H_0: \rho = 0$

$H_1: \rho \neq 0$

Isto equivale a testar as hipóteses

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

uma vez que $\beta_1 = \frac{\sigma_Y}{\sigma_X}\rho$

A estatística do teste é:

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{(n-2)}$$

Note que, como $H_0: \rho = 0$ pode ser escrito inteiramente em termos de r e de n, pode-se realizar o teste de hipótese mesmo sem o ajuste de uma reta de regressão.

4.3. Intervalo de confiança para ρ

Uma vez que distribuição de r é complicada quando $\rho \neq 0$, o intervalo de confiança é obtido por meio de uma transformação com aproximação pela Normal

$$z' = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right)$$

e o intervalo de confiança será dado por

$$IC(z', 1-\alpha) = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right) \mp z_{(1-\frac{\alpha}{2})} \frac{1}{\sqrt{n-3}}$$

Uma vez obtido o intervalo para z', é necessário transformar de volta para obter o intervalo para ρ.

5. Exercícios

Exercício 1.

No exemplo 1, vamos estudar a relação entre peso e idade. Para tanto:

a) Identifique a variável resposta (ou dependente) e a explicativa (ou indenpendente).

b) Construa o diagrama de dispersão, com a variável dependente no eixo Y e a independente no eixo X. Interprete-o.

c) Calcule a média, a variância e o desvio padrão de ambas as variáveis.

d) Calcule o coeficiente de correlação linear de Pearson entre idade e peso. Interprete-o.

e) Obtenha a reta de regressão do peso em função da idade.

f) Interprete os coeficientes da reta de regressão ajustada.

g) Desenhe a reta de regressão ajustada no diagrama de dispersão.

h) Obtenha as estimativas para o peso esperado (ou médio) em crianças com 8 e 11 anos.

i) Construa o quadro de análise de variância para os dados do exemplo 1.

j) Especifique e teste as hipóteses correspondentes.

k) Obtenha o coeficiente de determinação do modelo. Interprete-o.

l) Teste as hipóteses referentes a β_1.

m) Construa um intervalo de confiança para β_1.

n) Construa um intervalo de estimação para a média do peso na população de crianças de 8 e 11 anos.

o) Construa um intervalo de predição para o peso na população de crianças de 8 e 11 anos.

p) Verifique se a expressões abaixo são verdadeiras

$$\hat{\rho} = r = \frac{S_X}{S_Y} b_1 \quad e \quad b_1 = r \frac{S_Y}{S_X}$$

6. ANÁLISE DOS RESÍDUOS (ε_i=e_i):

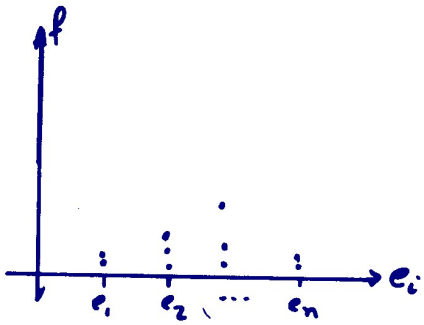
$$e_i = Y_i - \hat{Y}_i \quad , \quad i = 1,2, \ldots\ldots n$$

Suposições:

a) os e_i são independentes, ié, COV(e_i,e_k) =0, para i≠k.

b) $e_i \sim N(0, S_e)$, onde S_e^2 = constante

6.1. Análise Global:



o gráfico deve ter a aparência de uma curva normal

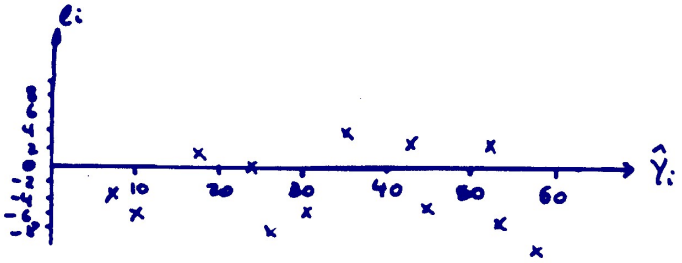
se $e_i \sim N(0; S_e) \Rightarrow \dfrac{e_i - \overline{X}_e}{S_e} \sim N(0;1)$

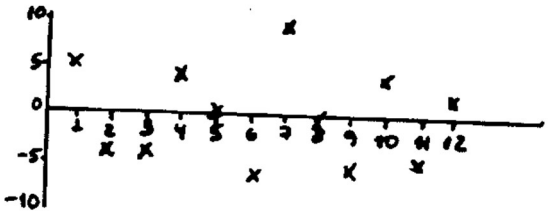
onde $S_e^2 = \dfrac{\sum\left(e_i - \bar{X}_e\right)^2}{n - p} = \dfrac{\sum e_i^2}{n - p}$; p = no. de variáveis indep.

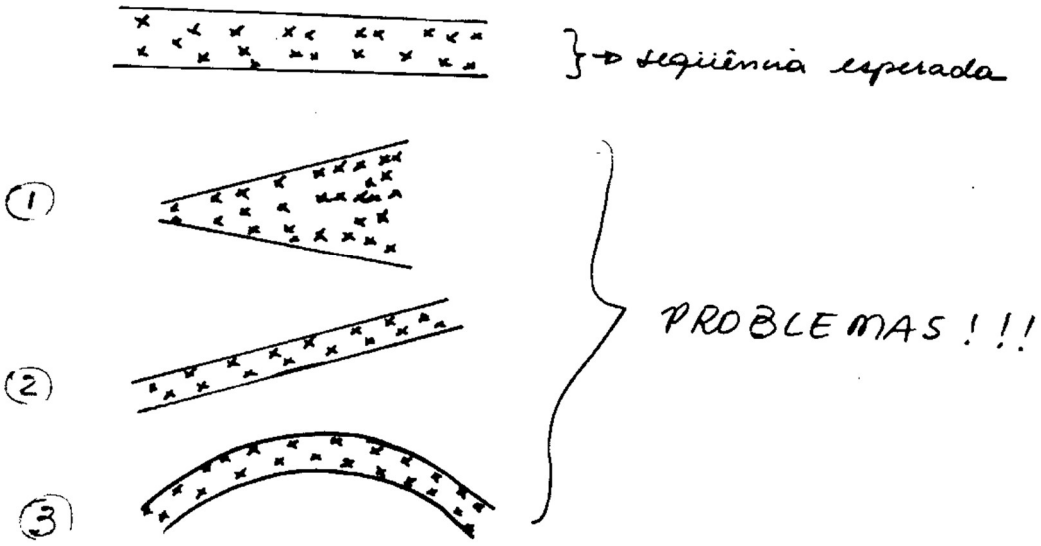
$\therefore IC_{95\%}(e_i) = [-1.96; +1.96]$

teste estatístico: aderência dos e_i à curva Normal.

6.2. Gráfico $e_i \times \hat{Y}_i$







❶ a variância não é constante (conforme suposto): deve-se fazer uma transformação na variável dependente Y_i, antes da análise de regressão ou fazer a estimação por mínimos quadrados ponderados.

❷ erro na análise de regressão: o modelo está viciado.

❸ o modelo é inadequado. São necessários termos adicionais (ex:quadrático ou produtos cruzados) ou é necessário que se faça uma transformação na variável dependente Y antes da análise.

6.3. Gráfico $e_i \times X_i$: idem ao 6.2.

6.4. Seqüência no tempo

(obs: é necessário que se conheça a seqüência, no tempo, em que os resíduos ocorrem)

❶ a variância não é constante no tempo: deve-se utilizar mínimos quadrados ponderados.

❷o tempo deve ser uma variável independente a ser introduzida no modelo (termo linear).

❸idem ao ❷, mas acrescentar, também, o termo de 2o. grau

6.5. testes estatísticos:

dos sinais e outros.

7. VALORES ABERRANTES (*OUTLIERS*)

Um valor aberrante é um ponto peculiar do conjunto de dados e, por isso, deve ser examinado cuidadosamente para que se descubra a razão de sua particularidade.

Não é prudente descartá-lo sem antes se proceder à uma investigação. Ele pode ser descartado quando seu valor for devido à um erro de mensuração e/ou registro ou devido à outro fator externo ao estudo.

ANÁLISE DE REGRESSÃO LINEAR MÚLTIPLA (MULTIVARIADA ????)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon \quad ; \quad \text{k : numero de variaveis}$$

$$\mathrm{Y} = \mathrm{f}(\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_\mathrm{k}) \quad \text{, utilizando amostra de tamanho} \quad n$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_k X_k$$

Y : v.a. dependente

X_j : v.a. independentes (regressores)

β_j : coeficientes de regressão (a serem estimados)

(cada β_j representa a mudança em $\bar{Y}_{Y/X_1,...,X_k}$ para uma unidade de cada X_j, quando todas as outras variáveis independentes permanecem constantes)

ex:

$$\mathrm{E}(Y / X_1 = X_2 = ... = X_k = 0) = \beta_0$$

$$\mathrm{E}(Y / X_1 = 1, X_2 = ... = X_k = 0) = \beta_0 + \beta_1$$

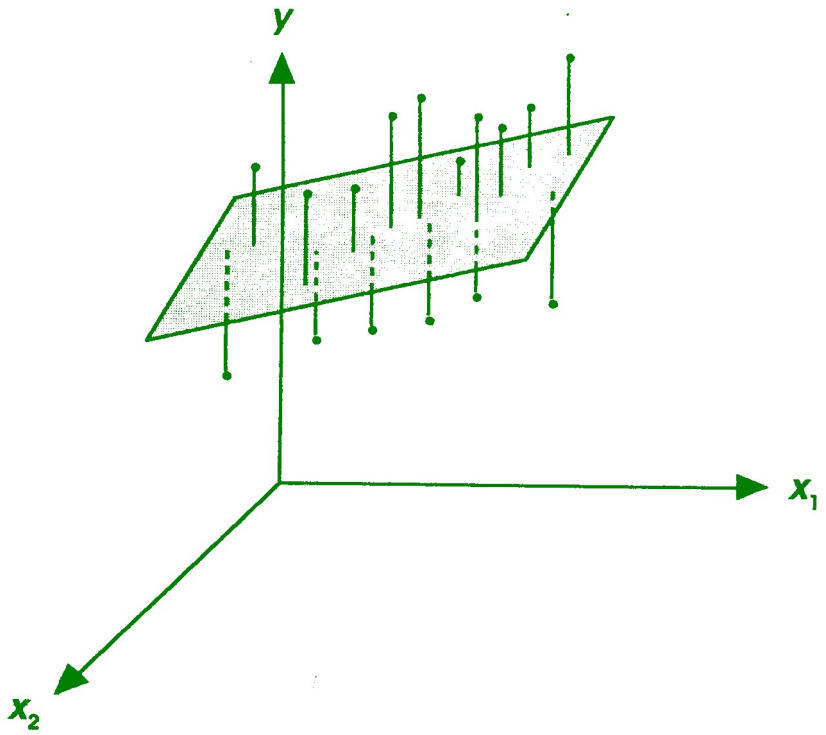
$$\mathrm{E}(Y / X_2 = 1, X_1 = X_3 = ... = X_k = 0) = \beta_0 + \beta_2$$

$$\mathrm{E}(Y / X_1 = X_2 = 1, X_3 = X_4 = ... = X_k = 0) = \beta_0 + \beta_1 + \beta_2$$

ESTIMATIVA POR MÍNIMOS QUADRADOS:

$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \rightarrow$ achar os β_j que minimizam esta expressão

ε : erro = resíduo (desvio do verdadeiro valor de Y em relação ao valor estimado pelo modelo, ié, $\left(Y_i - \hat{Y}_i\right)$



SUPOSIÇÕES BÁSICAS

São as mesmas do modelo simples, porém com extensão para múltiplas variáveis.

1. Distribuição Normal

Para um conjunto de valores fixos das v.a. X_j (que, idealmente, devem ser contínuas), Y é uma v.a. com distribuição normal, com média e variância finitas (aqui se trabalha em um espaço k-dimensional).

$$Y_i \sim N\left(\bar{Y}_{Y/X_1,X_2,\ldots,X_k};S\right)$$

2. Os valores de Y são independentes uns dos outros.

3. Linearidade

O valor médio de Y ($\bar{Y}_{Y/X_1,X_2,\ldots,X_k}$) é uma função de linear sobre os X_j.

4. Homocedasticidade

A variância de Y é a constante, qualquer que seja o conjunto dos X_j.

5. Não existe correlação entre os erros, ié, para quaisquer 2 amostras tem-se que : $COV(\varepsilon_i,\varepsilon_l)=0\,,\quad \forall\ i\neq l\,.$

6.Cada variável independente não está correlacionada com o termo de erro, ié, para cada X_j, $COV\left(X_j, \varepsilon_{i,j}\right) = 0$

7.Não há colinearidade perfeita entre as variáveis independentes, ié, nenhuma variável independente está relacionada linearmente, de maneira perfeita, com uma ou mais variáveis independentes.

EQUAÇÃO GERAL DA REGRESSÃO

$$\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + 0 \qquad ⑤$$

↓ ↓ ↓

SQTSQR SQM

SQTotal=SQ devida ao resíduo + SQ devida à regressão

ANOVA (modelo geral)

FONTE	SQ	GL	MQ	F_{TOTAL}
regressão	$\sum\left(\hat{Y}_i - \bar{Y}\right)^2$	k	SQM/k	$F_o(k, n-k-1) =$
resíduo	$\sum\left(Y_i - \hat{Y}_i\right)^2$	n-k-1	$SQR/n-k-1$	MQM/MQR
TOTAL	$\sum\left(Y_i - \bar{Y}\right)^2$	n-1		

$$r^2 = SQM/SQT \;\; ; F_c \sim F_{k,n-k-1}$$

MATRIZ DE CORRELAÇÃO

É uma matriz (k+1) x (k+1), sendo k o número de variáveis independentes que serão testadas no modelo múltiplo. Nesta matriz aparecem os coeficientes de correlação (r) entre todas as variáveis de estudo, sendo que na primeira linha deverão estar os coeficientes de correlação entre a variável dependente e as variáveis independentes. Esta é uma matriz com a diagonal unitária

	Y	X_1	X_2	X_3	.....	X_k	
Y	1	$r_{Y,X1}$	$r_{Y,X2}$	$r_{Y,X3}$		$r_{Y,Xk}$	ordem de entrada das variáveis independentes
X_1		1	$r_{X1,X2}$	$r_{X1,X3}$		$r_{X1,Xk}$	
X_2			1	$r_{X2,X3}$		$r_{X2,Xk}$	colinearidade
...							
...							
X_k						1	

ANOVA (adição de variáveis)

FONTE		SQ	GL	MQ	$F_{parcial}$
regressão	X_1	*	1	$SQM_{X_1}/1$	$F_o(1,n-1-1)=MQM_{X_1}/MQR$
	X_2/X_1	*	1	$SQM_{X_2}/1$	$F_o(1,n-2-1)=MQM_{X_2}/MQR$
	....	*	.....	.....	.....
	$X_k/X_1,X_2,...X_{k-1}$		1	$SQM_{X_k}/1$	$F_o(1,n-k-1)=MQM_{X_k}/MQR$
resíduo		$\sum\left(Y_i-\hat{Y}_i\right)^2$	n-k-1	$SQR/n-k-1$	
TOTAL		$\sum\left(Y_i-\bar{Y}\right)^2$	n-1		

* fórmulas nas páginas seguintes.

TESTES DE HIPÓTESES

1. Teste de significância do modelo geral

$$\begin{cases} \mathrm{H}_0: \beta_1 = \beta_2 = \ldots\ldots = \beta_k = 0 \\ \mathrm{H}_a: \text{ existe pelo menos um dos } \quad \beta_j \neq 0 \end{cases}$$

$$F_o = {MQM}/{MQR} \qquad \text{, onde } F_c \sim F_{k,n-k-1}$$

$$F_0 = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}}$$

2. teste do intercepto

$$\begin{cases} \mathrm{H}_0: \beta_0 = 0 \\ \mathrm{H}_a: \beta_0 \neq 0 \end{cases}$$

$$F_0 = \frac{\frac{SQR(\text{modelo sem } \beta_0) - SQR(\text{modelo com } \beta_0)}{1}}{\frac{SQR(\text{modelo com } \beta_0)}{\text{n - k - 1}}} \qquad , F_c \sim F_{1,n-k-1}$$

$$F_0 = \frac{\frac{n\bar{Y}^2}{1}}{\frac{\sum\left(Y_i - \bar{Y}\right)^2}{n-1}} \qquad , \qquad F_c \sim F_{1,n-1}$$

3. Teste do F parcial

$$\begin{cases} \mathrm{H}_0 : \beta^* = 0 \ , \text{no modelo } \ \mathrm{Y} = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \beta^* X^* \\ \mathrm{H}_a : \beta^* \neq 0 \\ \mathrm{H}_a : \mathrm{X}^* \text{ melhora significativamente a predicao de Y,} \\ \text{dado que } \mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_p \text{ já estao no modelo} \end{cases}$$

$$SQM\left(X^* / X_1, X_2, \ldots, X_p\right) = SQM\left(X_1, X_2, \ldots, X_p, X^*\right) - SQM\left(X_1, X_2, \ldots, X_p\right)$$

$$\therefore \ F_{p_o}\left(X^* / X_1, X_2, \ldots, X_p\right) = \frac{SQM\left(X^* / X_1, X_2, \ldots, X_p\right) \Big/ 1}{MQR\left(X_1, X_2, \ldots, X_p, X^*\right)}$$

$$F_{p_c}\left(X^* / X_1, X_2, \ldots, X_p\right) \sim F_{1, n-(p+1)-1}$$

4. Teste múltiplo do F parcial

$$\begin{cases} H_0 : \beta_1^* = \beta_2^* = \ldots = \beta_k^* = 0 \text{ no modelo} \\ \quad Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \underbrace{\beta_1^* X_1^* + \beta_2^* X_2^* + \ldots + \beta_k^* X_k^*}_{\text{bloco de variáveis}} \\ \\ H_a : \text{pelo menos um } \beta_j^* \neq 0 \\ H_a : \text{o bloco inteiro dos } X_j^* \text{ melhora significativamente a} \\ \text{predição de Y, dado que } X_1, X_2, \ldots, X_p \text{ já estão no modelo} \end{cases}$$

$$SQM\left(X_1^*, X_2^*, \ldots, X_k^* / X_1, X_2, \ldots, X_p\right) =$$
$$= SQM\left(X_1, X_2, \ldots, X_p, X_1^*, X_2^*, \ldots, X_k^*\right) - SQM\left(X_1, X_2, \ldots, X_p\right)$$

$$\therefore \; F_{mp_o}\left(X_1^*, X_2^*, \ldots, X_k^* / X_1, X_2, \ldots, X_p\right) = \frac{SQM\left(X_1^*, X_2^*, \ldots, X_k^* / X_1, X_2, \ldots, X_p\right) / k}{MQR\left(X_1, X_2, \ldots, X_p, X_1^*, X_2^*, \ldots, X_k^*\right)}$$

$$F_{mp_c}\left(X_1^*, X_2^*, \ldots, X_k^* / X_1, X_2, \ldots, X_p\right) \sim F_{k, n-(p+k)-1}$$

OBS:
1. como reconhecer variável de confusão?
2. como testar interação entre 2 variáveis independentes?

CORRELAÇÃO MÚLTIPLA

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

DEF.

$$r_{Y/X_1,X_2,\ldots X_k} = r_{Y,\hat{Y}} = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)\left(\hat{Y}_i - \overline{\hat{Y}}\right)}{\sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{\hat{Y}}\right)^2}}$$

$$r_{Y,\hat{Y}} = \frac{\sum_{i=1}^{n} Y_i \hat{Y}_i - n\overline{Y}^2}{\sqrt{\left(\sum_{i=1}^{n} Y_i - n\overline{Y}\right).\left(\sum_{i=1}^{n} \hat{Y}_i - n\overline{\hat{Y}}\right)}}$$

DEF: coeficiente de determinação múltipla (r^2)

$$r^2{}_{Y/X_1,X_2,\ldots X_k} = R^2{}_{Y,\hat{Y}} = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 - \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = \frac{SQM}{SQT}$$

Coef. de determinação múltipla ajustado ($r^2_{aj.}$)

$$r^2_{aj} = r^2 - \frac{k}{n-k-1}\left(1-r^2\right) = \frac{(n-1)r^2 - k}{n-k-1}$$

r^2_{aj} ➔ leva em conta a chance de contribuição de cada variável incluída, subtraindo-se o valor que seria esperado se nenhuma variável independente fosse associada à variável dependente.

O COEFICIENTE DE CORRELAÇÃO PARCIAL

$r_{Y,Xi/Xj}$ ➔ é uma estimativa de $\rho_{Y,Xi/Xj}$

Vamos supor a situação em que tenho apenas duas variáveis independentes X_1 e X_2.

$$\rho^2_{Y,X_1/X_2} = \frac{\sigma^2_{Y/X_2} - \sigma^2_{Y/X_1,X_2}}{\sigma^2_{Y/X_2}}$$

Nesta situação particular, tem-se que o coeficiente de correlação parcial ao quadrado é:

$$r^2_{Y,X_1/X_2} = \frac{SQR(\text{do modelo so com } X_2) - SQR(\text{do modelo completo, ie, com } X_1 \text{ e } X_2)}{SQR(\text{modelo so com } X_2)}$$

$$r^2_{Y,X_1/X_2} = \frac{\text{extra } SQ \text{ devido a adicao de } X_1, \text{ dado que } X_2 \text{ ja estava no modelo}}{SQR(\text{modelo so com } X_2)}$$

$$r_{Y,X_1/X_2} = \frac{r_{Y,X_1} - r_{Y,X_2} \cdot r_{X_1,X_2}}{\sqrt{\left(1 - r^2_{Y,X_2}\right).\left(1 - r^2_{X_1,X_2}\right)}}$$

A estatística $F_{parcial}(X_p/X_1, X_2,...X_k)$ é a utilizada para testar se $r_{Y,Xp/X_1,X_2,...,X_k} = 0$.

Representação alternativa do modelo de regressão.

Todos os coeficientes de regressão podem ser escritos em função das correlações parciais.

Por exemplo, para k=3 (ié, 3 variáveis independentes), tem-se:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2 - \hat{\beta}_3 X_3$$

$$\hat{\beta}_1 = r_{Y,X_1/X_2X_3} \cdot \frac{S_{Y/X_2.X_3}}{S_{X_1/X_2.X_3}}$$

$$\hat{\beta}_2 = r_{Y,X_2/X_1X_3} \frac{S_{Y/X_1.X_3}}{S_{X_2/X_1.X_3}}$$

$$\hat{\beta}_3 = r_{Y,X_3/X_1X_2} \frac{S_{Y/X_1.X_2}}{S_{X_3/X_1.X_2}}$$

COLINEARIDADE

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2$$

pode-se demonstrar que: $\beta_j = c_j \left[\frac{1}{1 - r^2_{X_1 X_2}} \right]$ e

que $\hat{\beta}_0, \hat{\beta}_1$ e $\hat{\beta}_2$ são diretamente proporcionais a $1 / 1 - r^2_{X_1, X_2}$

FIV: fator inflacionário da variância

$$FIV = \frac{1}{1 - R_j^2}$$

quando $FIV > 10 \Rightarrow$ há colinearidade

$FIV > 10 \Rightarrow R_j^2 > 0.90 \Rightarrow r_j > 0.95$

Para se evitar a colinearidade pode-se "centralizar" a variável.

VARIÁVEIS CATEGÓRICAS EM REGRESSÃO LINEAR

Há dois métodos para se analisar variáveis categóricas em regressão linear:

MÉTODO 1

Estimar uma equação de regressão para cada categoria da variável.

MÉTODO 2

Definir uma(algumas) variável(eis) *dummy* e incorporá-la(s) no modelo. Este método é menos poderoso.

VARIÁVEIS INDICADORAS

Variáveis indicadoras (ou *dummy*) são quaisquer variáveis que têm um número finito de valores que representam diferentes categorias de uma variável qualitativa.

Exemplo:

Y= PAS

X = idade ;

Z = sexo $\begin{cases} Z = 0 \Rightarrow \text{sexo} = \text{masculino} \\ Z = 1 \Rightarrow \text{sexo} = \text{feminino} \end{cases}$

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ \qquad (1)$$

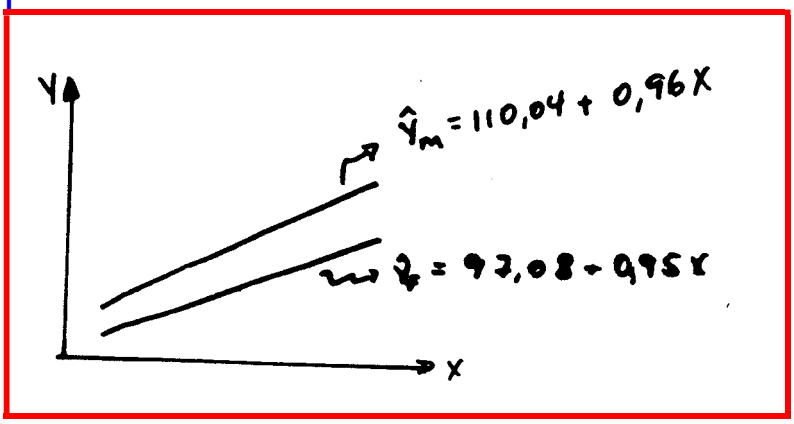
$$\text{qdo } Z = 0 \Rightarrow Y_M = \beta_0 + \beta_1 X \qquad (2)$$

$$\text{qdo } Z = 1 \Rightarrow Y_F = \beta_0 + \beta_1 X + \beta_2 + \beta_3 X \Leftrightarrow$$
$$Y_F = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X \qquad (3)$$

O modelo (1) incorpora as 2 equações de regressão separadas [(2) e (3)] em um único modelo.

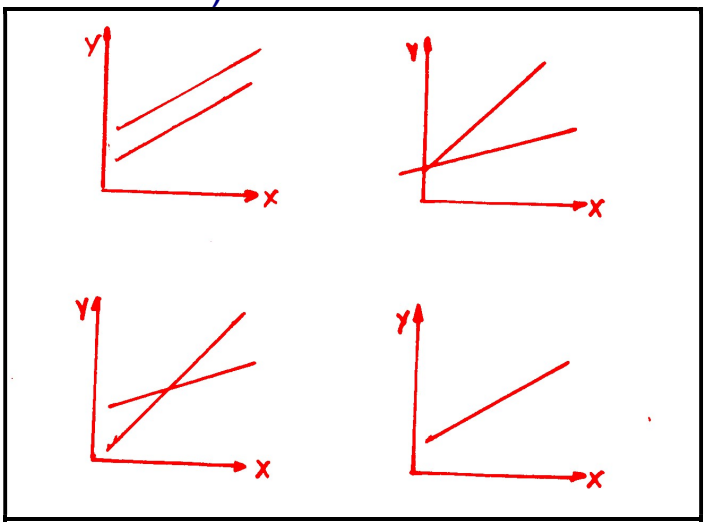
COMPARAÇÃO DE 2 RETAS DE REGRESSÃO

Questão:será que a associação entre PAS e idade é a mesma para homens e mulheres?



Perguntas:

1. As inclinações das 2 retas são iguais?(ié, existe paralelismo?)
2. Os interceptos das 2 retas são iguais?(somente no caso das 2 retas não serem paralelas)
3. As 2 retas têm interceptos e inclinações iguais?(ié, são coincidentes?)



CONTINUAÇÃO DO MÉTODO 1

1. teste de paralelismo de 2 retas

$$\hat{\beta}_1 = \frac{(n_M - 1)\, S^2_{X_M}\, \hat{\beta}_{1M} + (n_F - 1)\, S^2_{X_F}\, \hat{\beta}_{1F}}{(n_M - 1)\, S^2_{X_M} + (n_F - 1)\, S^2_{X_F}}$$

$$\begin{cases} H_0 : \beta_{1M} = \beta_{1F} \\ H_a : \beta_{1M} \neq \beta_{1F} \end{cases}$$

$$t_o = \frac{\hat{\beta}_{1M} - \hat{\beta}_{1F}}{S_{\hat{\beta}_{1M} - \hat{\beta}_{1F}}} \qquad t_c \sim t_{n_F + n_M - 4}$$

$$S_{\hat{\beta}_{1M} - \hat{\beta}_{1F}} = S^2_{P,Y/X} \left[\frac{1}{(n_M - 1) S^2_{X_M}} + \frac{1}{(n_F - 1) S^2_{X_F}} \right]$$

$$S^2_{P,Y/X} = \frac{(n_M - 2) S^2_{Y/X_M} + (n_F - 2) S^2_{Y/X_F}}{n_M + n_F - 4}$$

2.teste do intercepto

$$\hat{\beta}_0 = \frac{n_M \hat{\beta}_{0M} + n_F \hat{\beta}_{0F}}{n_M + n_F}$$

$$\begin{cases} H_0: \beta_{0M} = \beta_{0F} \\ H_a: \beta_{0M} \neq \beta_{0F} \end{cases}$$

$$t_o = \frac{\hat{\beta}_{0M} - \hat{\beta}_{0F}}{S_{\beta_{0M} - \hat{\beta}_{0F}}} \qquad t_c \sim t_{n_F + n_M - 4}$$

$$S^2_{\beta_{0M} - \hat{\beta}_{0F}} = S^2_{P,Y/X} \left[\frac{1}{n_M} + \frac{1}{n_F} + \frac{\bar{X}^2_M}{(n_M - 1) S^2_{X_M}} + \frac{\bar{X}^2_F}{(n_F - 1) S^2_{X_F}} \right]$$

3. teste de coincidência de 2 retas

Se ambas as hipóteses nulas forem aceitas: a de paralelismo e mesmo intercepto.

"PASSOS" PARA SE FAZER MODELAGEM EM REGRESSÃO

1. Selecionar as variáveis independentes, não se esquecendo das possíveis variáveis de confusão;

2. Codificar previamente as variáveis;

3. Fazer gráficos de dispersão (*scatter plot*) com todas as variáveis, 2 a 2;

4. Fazer a análise univariada das variáveis independentes, não se esquecendo de fazer a análise de resíduos.

5. Fazer a matriz de correlação para avaliar a colinearidade das variáveis independentes e definir a ordem de entrada das mesmas no modelo múltiplo.

6. Fazer a análise múltipla, avaliando a significância do modelo geral, de cada uma das variáveis e do incremento de cada uma delas, através do teste F e $F_{parcial}$. Não se esquecer de avaliar os possíveis efeitos de confusão e a colinearidade entre as variáveis;

7. Decidir pelo melhor modelo, ié, o mais "ajustado". Fazer a estimação por ponto e por intervalo de cada um dos β_j;

8. Avaliar as interações apenas para as variáveis de confusão;

9. Fazer análise dos resíduos.

ANÁLISE DE REGRESSÃO POLINOMIAL

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots\ldots + \beta_k X^k$$

ANOVA (regressão polinomial)

FONTE		SQ	GL	MQ	$F_{parcial}$
regressão	X	*	1	$SQM_X / 1$	$F_o(1, n-1-1) = MQM_X / MQR$
	X^2/X	*	1	$SQM_{X^2} / 1$	$F_o(1, n-2-1) = MQM_{X^2} / MQR$
	.... $X^k/X, X^2, \ldots X^{k-1}$	*	 1	..... $SQM_{X^k} / 1$	..... $F_o(1, n-k-1) = MQM_{X^k} / MQR$
resíduo		$\sum \left(Y_i - \hat{Y}_i \right)^2$	n-k-1	$SQR / n-k-1$	
TOTAL		$\sum \left(Y_i - \bar{Y} \right)^2$	n-1		

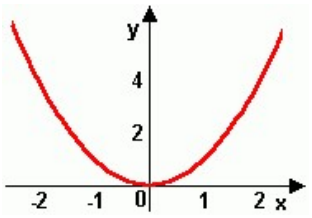
* fórmulas iguais às já citadas.

MODELO DE REGRESSÃO LINEAR

$$Y = \beta_0 + \beta_1 X$$

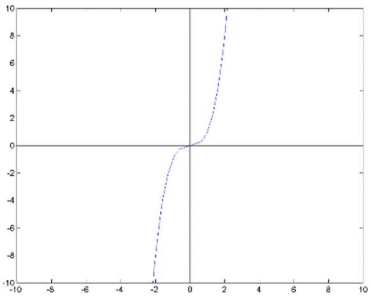
MODELO DE REGRESSÃO DE 2a ORDEM

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2$$



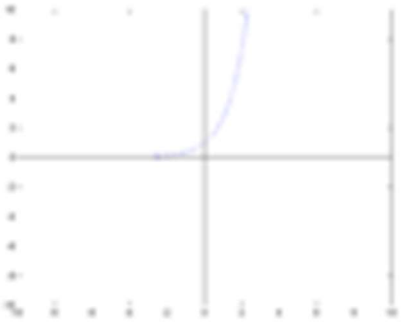
MODELO DE REGRESSÃO DE 3a ORDEM

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$



MODELO DE REGRESSÃO EXPONENCIAL

$$Y = \beta_0 * e^{(\beta_1 X)} \; ou \; \ln(Y) = \ln(\beta_0) + (\beta_1 X)$$



ANÁLISE DE SÉRIES (HISTÓRICAS) TEMPORAIS

Uma série histórica, também denominada série temporal, é uma seqüência de observações obtidas em intervalos regulares de tempo, durante um período específico. Este conjunto pode ser obtido através de amostras periódicas do evento de interesse, ou cumulativamente. Denomina-se trajetória de um processo, a curva obtida no gráfico da série histórica. O conjunto de todas possíveis trajetórias é denominado um processo estocástico, sendo a série temporal uma amostra deste processo.

DEFINIÇÕES

série temporal (Z)

É um conjunto de observações ordenadas no tempo. Essas observações podem ser discretas ou contínuas.

discreta: $Z_t \rightarrow$ t = 1, 2, ...n

- valores semanais do número de casos de Aids em São Paulo
- coeficientes de mortalidade (mensais, anuais)

contínua: Z(t) $\rightarrow$ $t \in [0, T]$

- o registro de um eletrocardiograma de uma pessoa.
- o movimento da costa terrestre, obtido através de um sismógrafo.

Essas observações podem ser obtidas através de amostras periódicas ou cumulativamente.

trajetória do processo

É a curva obtida no gráfico das observações no tempo.

processo estocástico

É um conjunto de todas as possíveis trajetórias que poder-se-ia observar. Cada trajetória é chamada de uma série temporal.

processo estocástico → população
série temporal → amostra

ciclo:

É o tempo que um determinado fenômeno leva para ter um comportamento periódico. Diz-se, nesse caso, que o fenômeno é cíclico. Esse ciclo pode ou não ser conhecido *a priori*. Um fenômeno cíclico envolve um comportamento sazonal. Porém o inverso não é verdadeiro.

estacionariedade:

Uma série é considerada estacionária quando as suas observações ocorrem, aleatoriamente, ao redor de uma média constante. Essa é a suposição de grande parte dos modelos. Quando isso não ocorre é necessário que se façam transformações nos dados e/ou se utilizem modelos adequados. A não aleatoriedade é um fenômeno freqüente.

Diz-se que uma série é estacionária quando, para qualquer instante de tempo t e para qualquer m, tem-se:

$$f(Z_t) = f(Z_{t+m}) \ , \ \ m = \pm 1, \pm 2, \ldots\ldots$$

$$\Rightarrow \begin{cases} E(Z_t) = E(Z_{t+m}) = \mu, \ \ \forall\, t \\ e \\ Var(Z_t) = Var(Z_{t+m}) = \sigma^2 \ , \ \forall\, t \end{cases}$$

Exemplo 1.1. A Figura 1.6 mostra a série de Índices de Produto Industrial (IPI) do Brasil, composta de 139 observações mensais (de janeiro de 1969 a julho de 1980). Nesta série, além de uma periodicidade (sazonal) aparente de doze meses, notamos que ela apresenta uma tendência crescente, sendo portanto não-estacionária (Tabela A.1).

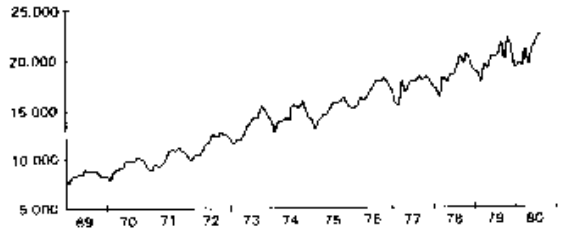


Fig. 1.6. *Série de Índice de Produto Industrial do Brasil*

Exemplo 1.2. Na Tabela A.2 temos 96 observações mensais (em graus centígrados) de temperaturas da cidade de São Paulo (de julho de 1949 a junho de 1957). Analisando os dados e a Figura 1.7, vemos que as temperaturas oscilam entre um mínimo, que ocorre geralmente em julho,

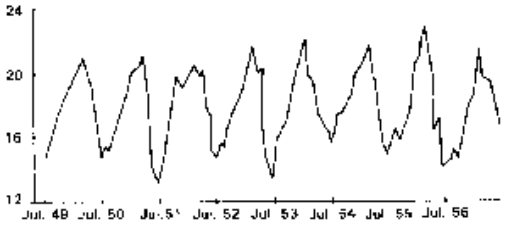


Fig. 1.7. *Série de temperaturas da cidade de São Paulo*

COMPONENTES DE UMA SÉRIE TEMPORAL

Uma série histórica pode ser decomposta em 3 componentes não observáveis: tendência (Tt), sazonalidade (St) e a variação aleatória denominada de ruído branco (at).

T_t → tendência

Esse é um componente não aleatório que, muitas vezes, só consegue ser medido e/ou detectado em longas séries de tempo.

S_t → componente sazonal

Ocorre quando duas observações no tempo são correlacionadas, ou seja, não são independentes. Para se avaliá-lo é necessário analisar as funções de auto-covariância e de auto-correlação da série.

a_t → ruído branco

também conhecido como resíduo. Supõe-se que esse seja um componente aleatório, com média zero e variância constante (em toda a série).

Modelo aditivo: Zt=Tt+St+at

(pode haver , também, o modelo multiplicativo, que ao se realizar a transformação log, ele se transforma no modelo log-linear).

Ao ser feita a análise de uma série histórica, deve-se estudar cada um destes componentes separadamente, retirando-se o efeito dos outros.

TENDÊNCIA

Para analisar a tendência os 2 métodos mais utilizados são: a) ajustar uma função polinomial do tempo ou b) analisar o comportamento da série ao redor de um ponto, estimando a tendência naquele ponto.

Após a estimativa da tendência, uma série "livre de tendência" seria a série (Zt-Tt).

SAZONALIDADE

Esta parte da série histórica é difícil de ser estimada, compatibilizando a questão conceitual do fenômeno em estudo, com a questão estatística. Se houver uma sazonalidade dita determinística pode-se utilizar modelos de regressão que incorporem funções do tipo seno ou cosseno à variável tempo.

Para se retirar o efeito da sazonalidade de uma série, pode-se fazer a média móvel centrada no número de períodos que compõem uma repetição (por exemplo, para sazonalidade anual, seria utilizada a média móvel de 12 meses), ou, então, poderia-se trabalhar com a diferença entre a série original (Zt) e o polinômio estimado para a sazonalidade.

Exercicios

Complica

Inserir para o exercício complica:

$\hat{Y} = b_0 + b_1 X$

Como $b_0 = \bar{Y} - b_1 \bar{X}$

$\hat{Y} = \bar{Y} - b_1 \bar{X} + b_1 X$

$\hat{Y} = \bar{Y} - b_1 (X - \bar{X})$

Ao centralizar a variável ano, o b_0 será a média de Y do período

REGRESSÃO LOGÍSTICA

- Variável dependente é qualitativa dicotômica (presença/ausência)

- Objetivo principal do estudo é estudar os fatores associados à presença do evento de interesse.

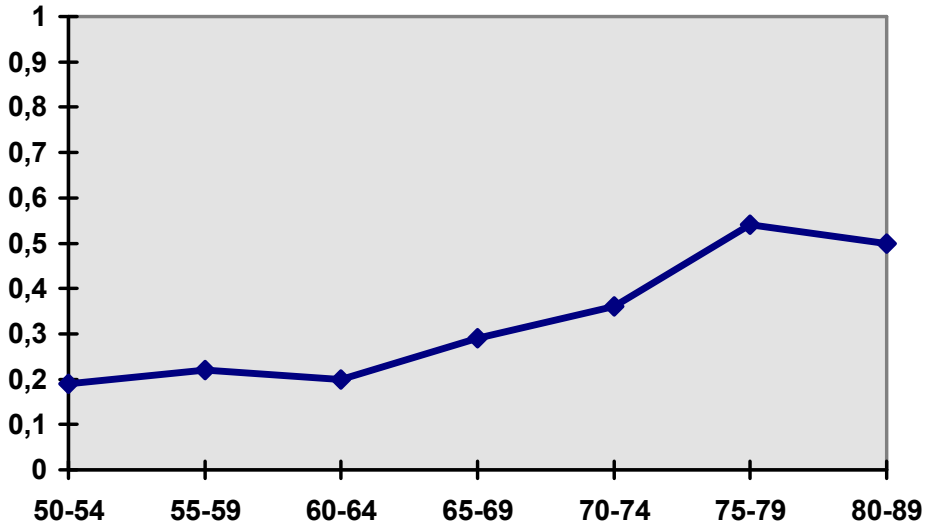
EXEMPLO

Y= doença coronariana(DC) $\begin{cases} Y = 1 \Rightarrow DC = sim \\ Y = 0 \Rightarrow DC = nao \end{cases}$

IDADE	DC			
	SIM	NÃO	TOTAL	p=% de sim
20 - 29	1	9	10	0.10
30 - 34	2	13	15	0.13
35 - 39	3	9	12	0.25
40 - 44	5	10	15	0.33
45 - 49	6	7	13	0.46
50 - 54	5	3	8	0.63
55 - 59	13	4	17	0.76
60 - 69	8	2	10	0.80
Total	43	57	100	0.43

Fonte: Kleimbaum,Klein, 2002.

FAZER O DIAGRAMA DE DISPERSÃO



$$Prob(Y=1) = p = \frac{1}{1+e^{-f(x)}}$$

Quando a f(x) é uma função linear, tem-se que

$$Prob(Y=1) = p = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

e

$$\text{Prob}(Y \neq 1) = \text{Prob}(Y = 0) = 1 - p = 1 - \frac{1}{1+e^{-(\beta_0+\beta_1 X)}} = \frac{e^{-(\beta_0+\beta_1 X)}}{1+e^{-(\beta_0+\beta_1 X)}}$$

	doente	não doente	TOTAL
EXPOSTO	a	b	a+b
NÃO EXPOSTO	c	d	c+d
TOTAL	a+c	b+d	N=a+b+c+d

Medidas de risco:

RP: razão de prevalências ➔ $RP = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$

RR: risco relativo ➔ $RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$

OR: *odds ratio* ➔ $OR = \frac{a/b}{c/d} = \frac{a.d}{b.c}$

densidade de incidência, incidência acumulada.

Y = variável dependente; variável categórica (0,1)

$$\begin{cases} \mathrm{Y} = 1 \\ \mathrm{Y} = 0 \end{cases} \Rightarrow Y \sim Bernoulli \Rightarrow \begin{cases} \mathrm{P}(\mathrm{Y} = 1) = \pi \\ \mathrm{P}(\mathrm{Y} = 0) = 1 - \pi \end{cases}$$

$$\mathrm{E}(\mathrm{Y}) = \sum_{i=1}^{2} y_i P(Y = y_i) = 1P(Y = 1) + 0P(Y = 0) = 1\pi + 0(1 - \pi) = \pi$$

O objetivo é escrever Y em função de X, porém, na regressão logística, se escreve a probabilidade de Y como função de X e não Y.

$$\pi(x) = E(Y / X = x)$$

$$\pi(x) = \frac{e^{f(x)}}{1 + e^{f(x)}}$$

Quando a f (x) e uma função linear, tem - se que

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} = \frac{1}{1 + e^{-\left(\beta_0 + \beta_1 X\right)}}$$

Fazendo - se a transformação para o logito de $\pi(x)$,

$$\ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 X$$

SUPOSIÇÕES

1. Y é uma variável dicotômica (0,1). A extensão para outras variáveis categóricas não será vista neste curso.

2. Os valores de Y são independentes.

3.

$$\mathrm{E}(\mathrm{Y}) = \pi(x) \Rightarrow \hat{\mathrm{E}}(\mathrm{Y}) = \hat{\pi}(x) + \varepsilon$$

ε = erro = resíduo

$$\varepsilon \sim Binomial \quad \text{, pois } \varepsilon = \begin{cases} 1 - \pi(x), \text{ se } \hat{\mathrm{E}}(\mathrm{Y}) = 1, \\ [\text{com prob. } \pi(x)] \\ \\ -\pi(x), \text{ se } \hat{\mathrm{E}}(\mathrm{Y}) = 0, \\ [\text{com prob. } 1 - \pi(x)] \end{cases}$$

.

$$\therefore \begin{cases} \bar{\varepsilon} = 0 \\ S_{\varepsilon}^2 = \{\pi(x)[1 - \pi(x)]\} \rightarrow \text{variância não é constante} \end{cases}$$

4. A covariância entre dois erros quaisquer é zero.

ESTIMATIVA DOS PARÂMETROS β_i

Na regressão logística é utilizado o Método da Máxima Verossimilhança para se estimar os parâmetros β_i .

De uma maneira genérica, pode-se dizer que o método da máxima verossimilhança fornece os valores para os parâmetros a serem estimados, os quais maximizam a probabilidade de se obter o conjunto de dados existente.

Para se aplicar este método, em primeiro lugar precisa-se definir a função de verossimilhança. Na situação em que a variável dependente é dicotômica, tem-se:

Seja $Y = \begin{cases} 0 \\ 1 \end{cases} \Rightarrow$

$$\begin{cases} 1 - \pi(x) = \dfrac{e^{\beta_0 + \beta_1 X}}{1 + e^{-(\beta_0 + \beta_1 X)}} = P(Y = 0/X) \\ \pi(x) = \dfrac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} = P(Y = 1/X) \end{cases}$$

para um arbitrário valor de $\beta = (\beta_0, \beta_1) \Rightarrow$

A função de probabilid ades de Y é

$$f(Y_i) = \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i}, \quad \text{onde} \begin{cases} Y = 0,1 \\ i = 1,2,...n \end{cases}$$

Assim, para aqueles pares $(x_i, 1)$, a contribuiç ão para a função de verossimi lhança é $\pi(x)$ e naqueles onde $Y_i = 0$, a contribuiç ão é $1 - \pi(x)$.

A função de verossimi lhança é definida pelo produto dos termos dados acima, ié,

$$L(\beta) = \prod_{i=1}^{n} f(Y_i)$$

No entanto, e mais fácil maximizar o $\ln \left[L(\beta) \right]$.

$$\ln \left[L(\beta) \right] = \sum_{i=1}^{n} \left[y_i \ln(\pi_i) + (1 - y_i) \ln(1 - \pi_i) \right]$$

Para encontrar os valores dos βi que maximizam a função acima deve-se derivar $\left[\ln L\left(\underset{\sim}{\beta}\right)\right]$ em relação a cada um dos βi e igualar a zero. Como estas equações não são lineares, são necessários métodos interativos e sua solução não é fácil! Porém os *softwares* fazem isso por nós !!!!

As equações são:

$$\begin{cases} \sum_{i=1}^{n}\left[y_i - \pi(x_i)\right] = 0 \\ \text{e} \\ \sum_{i=1}^{n} x_i \left[y_i - \pi(x_i)\right] = 0 \end{cases}$$

Estas são as chamadas equações de verossimilhança.

Normalmente as saídas de computador fornecem não só os valores dos β_i, mas, também, os respectivos erros padrão ($SE_{\beta i}$). Os valores dos $SE_{\beta i}$ serão utilizados para os testes de significância dos coeficientes e para o cálculos dos respectivos intervalos de confiança.

No caso do pior modelo (modelo só com β_0), o logaritmo da função de verossimilhança pode ser calculado por:

$$\ln\left[\mathrm{L}\left(\beta_{\mathrm{o}}\right)\right] = n_1 . \ln\left(n_1\right) + n_0 \ln\left(n_0\right) - n \ln(n)$$

onde:n_1: número de casos de Y=1

n_0: número de casos de Y=0

$n=n_1+n_0$= total da amostra

TESTES DE HIPÓTESES

Na regressão logística a comparação entre o valor observado e o valor predito pela regressão não é feita através da ANOVA, mas é baseada no logarítmo da função de verossimilhança já definida $\left[\ln L\left(\underset{\sim}{\beta}\right)\right]$.

1. Teste da razão de verossimilhança

É feita a comparação entre a função de verossimilhança dos valores observados na amostra e a função de verossimilhança do modelo saturado. O modelo saturado é aquele que contém tantos parâmetros quanto o número de pontos da amostra (ex: ajustar uma linha reta com 2 pontos).

$$D = deviance$$

$$D = -2\left\{\ln\left[L(\text{modelo reduzido})\right] - \ln\left[L(\text{modelo saturado})\right]\right\} \Rightarrow$$

$$D = -2\ln\left[\underbrace{\frac{L(\text{modelo reduzido})}{L(\text{modelo saturado})}}_{\text{razão de verossimilhança}}\right]$$

Para verificar a significância de uma variável independente, compara-se o valor de D com e sem a variável independente na equação. A mudança de D devido à inclusão da variável independente é:

$$G = D(\text{para o modelo sem a variavel}) - D(\text{para o modelo com a variavel})$$

$$G = \left\{ -2\ln\left[\frac{L(\text{mod.sem variavel})}{L(\text{mod elo saturado})}\right] - 2\ln\left[\frac{L(\text{mod. com variavel})}{L(\text{mod elo saturado})}\right]\right\}$$

$$G = -2\ln\left[\frac{L(\text{modelo sem variável})}{L(\text{mod elo com variável})}\right]$$

$G \sim \chi^2_1 \rightarrow$ para o teste de significância de 1 variável com 2 categorias

no caso do modelo univariado, $H_0 : \beta_1 = 0$

2. Teste Wald (baixo poder)

$$H_0 : \beta_1 = 0 \Leftrightarrow H_0 : OR(X_1) = 1$$

$$W = \frac{\hat{\beta}_i}{SE_{\hat{\beta}_i}} \qquad , \qquad \text{onde } W_c \sim N(0,1)$$

3. Intervalo de Confiança

$$IC_{(1-\alpha)\%}(\beta_i) = \hat{\beta}_i \pm z_{1-\alpha} \; x \; SE_{\hat{\beta}_i}$$

4. Cálculo do RR

Vamos supor o caso mais simples em que a variável dependente X é dicotômica. Então,

$$RR = \frac{\Pr ob(Y = 1 / X = 1)}{\Pr ob(Y = 1 / X = 0)} = \frac{\dfrac{1}{1 + \exp^{-(\beta + \beta_1 x 1)}}}{\dfrac{1}{1 + \exp^{-(\beta_0 + \beta_1 x 0)}}} = \frac{1 + \exp^{-(\beta_0)}}{1 + \exp^{-(\beta_0 + \beta_1)}}$$

logo,

$$H_0 : \beta_1 = 0 \Leftrightarrow H_0 : OR(X_1) = 1 \Leftrightarrow H_0 : RR(X_i) = 1$$

5. Caso múltiplo

Utilizar o teste da razão de verossimilhança para verificar a adequação do modelo como um todo, ié:

$$\begin{cases} \mathrm{H_o} : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \\ \mathrm{H_a} : \text{o modelo é adequado, ie existe pelo um } \beta \neq 0 \end{cases}$$

$$G \sim \chi^2_k \text{ , onde k : número de } \beta\text{'s do modelo}$$

Para testar a significância de cada coeficiente, utilizar o teste Wald:

$$\begin{cases} \mathrm{H_0} : \hat{\beta}_i = 0 \Leftrightarrow H_0 : OR(X_i) = 1 \Leftrightarrow H_0 : RR(X_i) = 1 \\ \mathrm{H_a} : \hat{\beta}_i \neq 0 \Leftrightarrow H_0 : OR(X_i) \neq 1 \Leftrightarrow H_0 : RR(X_i) \neq 1 \end{cases}$$

$$W_i = \frac{\hat{\beta}_i}{SE_{\hat{\beta}_i}} \text{ , onde } W_{ic} \sim N(0,1)$$

Estimativa da *odds ratio* (OR) a partir do modelo de regressão logística múltipla

chance: $$\frac{\Pr ob(Y=1)}{\Pr ob(Y=0)} = \frac{p}{1-p}$$

$$OR\left(X_1\right) = \frac{\dfrac{p_{X_1=1}}{1-p_{X_1=1}}}{\dfrac{p_{X_1=0}}{1-p_{X_1=0}}} = \frac{e^{(\beta_0+\beta_1(X_1=1)+\beta_2X_2+.....+\beta_kX_k)}}{e^{(\beta_0+\beta_1(X_1=0)+\beta_2X_2+.....+\beta_kX_k)}} =$$

$$e^{(\beta_0+\beta_1+\beta_2X_2+.....+\beta_kX_k)-(\beta_0+\beta_2X_2+.....+\beta_kX_k)} = e^{\beta_1}$$

6.Análise de confusão e interação na regressão logística

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1.X_2$$

Outra maneira de testar interação: criar uma 3a. variável (Z), que é a combinação de X_1 e X_2.

X_1	X_2	Z	Z_1	Z_2	Z_3
1	1	3	0	0	1
1	0	2	0	1	0
0	1	1	1	0	0
0	0	0	0	0	0

MODELOS DE REGRESSÃO LOGÍSTICA

- Não condicional: estudos transversais, coorte e caso-controle não pareado

- Condicional: estudos caso-controle e outros onde haja pareamento. Nestes casos, no banco de dados deverá existir a variável "**par**".

ANÁLISE DOS RESÍDUOS

1. Estatística do χ^2 de Pearson

2. Teste de Hosmer-Lemeshow

---------- Hosmer and Lemeshow Goodness-of-Fit Test----------

	LOW = 0		LOW = 1		
Group	Observed	Expected	Observed	Expected	Total
1	35.000	34.180	3.000	3.820	38.000
2	25.000	26.537	9.000	7.463	34.000
3	29.000	29.743	10.000	9.257	39.000
4	16.000	14.736	6.000	7.264	22.000
5	10.000	9.460	7.000	7.540	17.000
6	8.000	9.877	12.000	10.123	20.000
7	7.000	5.466	12.000	13.534	19.000

	Chi-Square	df	Significance
Goodness-of-fit test	2.3862	5	.7935
