

Fundamentos IV

Ajuste de Curvas

Clarimar J. Coelho

Departamento de Computação

October 9, 2014

Diferença entre interpolação e ajuste de curvas

Interpolação

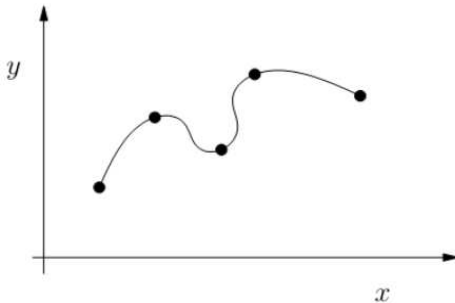
- A variação das leituras de uma variável ou fatores externos aos experimentos podem muitas vezes levar a interpolação a gerar um polinômio de grau elevado para modelar sistemas que na verdade são lineares ou de grau bem mais baixo

O ajuste de curvas

- Nestes casos devemos usar o ajuste de curvas para determinar o melhor polinômio de grau mais baixo que se encaixe nos dados apresentados

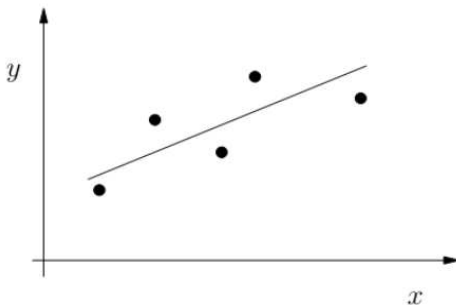
Na interpolação

- O polinômio gerado passa por todos os pontos da tabela utilizada no cálculo, com um polinômio de grau $(n - 1)$



No ajuste de curvas

- O polinômio gerado passa pelo melhor caminho entre os pontos da tabela, e não sobre eles
- O ajuste de curvas normalmente utiliza polinômios de grau menor



Tipos de relação entre variáveis

Relação entre variáveis

- As relações entre as variáveis envolvidas em um experimento podem ser classificadas em três tipos
 - Determinísticas
 - Semideterminísticas
 - Empíricas

Relações determinísticas

- As variáveis estão relacionadas entre si por algum tipo de lei que pode ser expressa por meio de uma fórmula matemática

$$saldo = saldo_inicial \times (1 + juros)^{meses}$$

- Qualquer variação nas observações é atribuída a erros experimentais

Relações semideterminísticas

- Existe uma expressão matemática que relaciona as variáveis
- Mas nem todos os seus parâmetros são conhecidos e preciso estimá-los
 - A concentração de uma substância depois de um tempo t depende de uma constante de velocidade da reação específica k , obtida experimentalmente

$$c = c_0 e^{-kt}$$

Relações empíricas

- Em muitas outras situações, a relação entre as variáveis é desconhecida
- Procura-se então expressar uma possível relação entre elas através da determinação de uma equação que melhor se ajuste aos pontos experimentais
 - A relação entre a produtividade na agricultura e a quantidade de adubo utilizada na lavoura
 - Existem diversos fatores que podem contribuir para a produtividade, mas temos interesse em somente um deles

Análise de regressão

- Estimativa de parâmetros, análise de variância e de resíduos, testes de hipótese
- Formulação de modelos matemáticos que descrevem as relações entre variáveis
- Uso destes modelos com o propósito de predição e outras inferências estatísticas
- A análise de regressão é estudada na estatística
- Nesse curso, vamos estudar apenas a determinação de parâmetros de modelos semideterminísticos

Regressão linear simples (RLS)

- A relação mais simples entre duas variáveis são as relações lineares
- A variável independente ou explicativa x é relacionada com a variável dependente ou resposta y através de um modelo linear

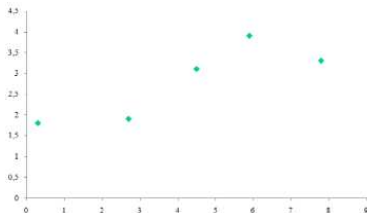
$$y = b_0 + b_1x$$

Diagrama de dispersão

- Um passo importante antes de analisar a relação entre duas variáveis é o esboço dos dados em um gráfico de coordenadas cartesianas chamado diagrama de dispersão

x	0.3	2.7	4.5	5.9	7.8
y	1.8	1.9	3.1	3.9	3.3

- O diagrama de dispersão mostra uma relação quase linear entre as variáveis explicativas x e as respostas y



Retas de regressão

- Um modelo simples que relaciona duas variáveis x e y é

$$y = \beta_0 + \beta_1 x + \epsilon$$

- onde β_0 e β_1 são os parâmetros a serem estimados e ϵ é o erro aleatório do modelo
- Nosso problema é calcular os estimadores de β_0 e β_1 , b_0 e b_1

Modelo 1 - primeira tentativa

- Vamos tentar usando um polinômio interpolador de grau 1
- O diagrama de dispersão mostra que não é possível traçar uma reta única que passe por todos os pontos simultaneamente
- Então, a reta deve ser esboçada a partir pontos, por exemplo o primeiro e o último

x	0.3	7.8
y	1.8	3.3

Equação da reta

- A equação da reta $u(x)$ que passa por estes dois pontos, usando

$$P_1(x) = y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x - x_0)$$

- É

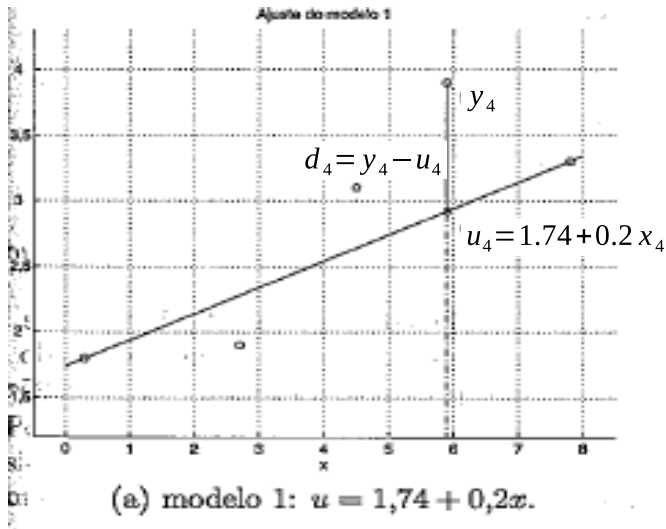
$$u(x) = y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x - x_0)$$

$$= 1.8 + \frac{3.3 - 1.8}{7.8 - 0.3}(x - 0.3)$$

$$= 1.8 + 0.2(x - 0.3)$$

$$\rightarrow u(x) = 1.74 + 0.2x$$

Ajustes lineares



Distância do i -ésimo ponto

- Na Figura (a) anterior a reta $u = 1.74 + 0.2x$ traçada entre os pontos do diagrama de dispersão
- A distância vertical d_i entre o i -ésimo ponto dado y_i e o ponto $u_i = 1.74 + 0.32x_i$ de mesma abscissa x_i e pertencentes à reta é

$$d_i = y_i - u_i$$

Qualidade do modelo

- Uma maneira de verificar a qualidade do ajuste é pelo cálculo de todas as n distâncias verticais de y_i aos pontos da reta $u_i = 1.74 + 0.42x$ para valores positivos de d_i

$$D(b_0, b_1) = \sum_{i=1}^n (y_i - u_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^n d_i^2$$

Resultado do ajuste pelo modelo 1

$$D(1.74; 0.2) = \sum_{i=1}^5 (y_i - 1.74 - 0.2x_i)^2$$

i	x_i	y_i	u_i	d_i
1	0.3	1.8	1.80	0.00
2	2.7	1.9	2.28	-0.38
3	4.5	3.1	2.64	0.46
4	5.9	3.9	2.92	0.98
5	7.8	3.3	3.30	0.00

- $D(1.74; 0.2) = 1.3164$

Modelo 2 - segunda tentativa

- Vamos traçar a reta por dois pontos quaisquer, que não pertencem necessariamente, ao diagrama de dispersão
- Vamos escolher os dois pontos

x	2	6
y	2	3

A reta

$$u(x) = y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x - x_0)$$

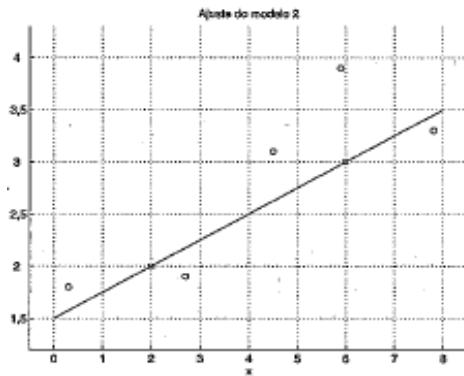
$$= 2 + \frac{3-2}{6-2}(x - 2)$$

$$= 2 + 0.25(x - 2)$$

$$\rightarrow u(x) = 1.5 + 0.25x$$

Ajustes lineares

- A Figura mostra a reta $u(x) = 1.5 + 0.25x$ esboçada no diagrama de dispersão



(b) modelo 2: $u = 1.5 + 0.25x$

Resultados do ajuste pelo modelo 2

i	x_i	y_i	u_i	d_i
1	0.3	1.8	1.575	0.225
2	2.7	1.9	2.175	-0.275
3	4.5	3.1	2.625	0.475
4	5.9	3.9	2.975	0.925
5	7.8	3.3	3.450	-0.150

- $D(1.5; 0.25) = 1.2300 < D(1.74; 0.2) = 1.3164$

Métodos dos quadrados mínimos

Métodos dos quadrados mínimos

- As Tabelas dos resultados anteriores mostra que a qualidade do ajuste depende da equação da reta escolhida
- O fato da reta não passar por dois pontos entre aqueles do diagrama de dispersão produz um resultado melhor
- Por onde deve passar a reta de modo a obter o menor valor do desvio D

Estimativa

- O método dos quadrados mínimos consiste em encontrar uma estimativa da reta

$$u = \beta_0 + \beta_1 x$$

- De modo a produzir o menor valor possível do desvio

$$D(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - u_i)^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

- Com as derivadas parciais

$$\frac{\partial D(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial D(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i$$

Estimadores de β

- Os valores para os quais a função $D(\beta_0, \beta_1)$ possui um mínimo são aquelas onde as derivadas parciais se anulam

- Se $D(b_0, b_1)$ for o ponto mínimo de $D(\beta_0, \beta_1)$

$$-2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) = 0 \rightarrow \sum_{i=1}^n b_0 + \sum_{i=1}^n b_1 x_i = \sum_{i=1}^n y_i$$

$$-2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) x_i = 0 \rightarrow \sum_{i=1}^n b_0 x_i + \sum_{i=1}^n b_1 x_i^2 = \sum_{i=1}^n x_i y_i$$

Forma matricial

- Escrevendo na forma matricial e trocando a notação $\sum_{i=1}^n$ para $\sum$

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} \quad (1)$$

Finalmente os estimadores de β

- Os valores de $D(\beta_0, \beta_1)$ apresenta um mínimo e é obtido pelo sistema linear (1) que denominado de equações normais
- Usando operações elementares, obtemos o sistema linear equivalente

$$\begin{bmatrix} n & \sum x_i \\ 0 & -\frac{1}{n}(\sum x_i)^2 + \sum x_i^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ -\frac{1}{n} \sum x_i \sum y_i + \sum x_i y_i \end{bmatrix}$$

- Cuja solução é

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

$$b_0 = \frac{\sum y_i - b_1 \sum x_i}{n}$$

Exemplo 1

- Calcular a reta de quadrados mínimos usando os dados da Tabela

i	x_i	y_i	x_i^2	$x_i y_i$	y_i^2
1	0.3	1.8	0.09	0.54	3.24
2	2.7	1.9	7.29	5.13	3.61
3	4.5	3.1	20.25	13.95	9.61
4	5.9	3.9	34.81	23.01	15.21
5	7.8	3.3	60.84	25.74	10.58
$\sum$	21.2	14.0	123.28	68.37	42.56

$$b_1 = \frac{\sum x_i \sum y_i - n \sum x_i y_i}{(\sum x_i)^2 - n \sum x_i^2} = \frac{21.2 \times 14.0 - 5 \times 68.37}{(21.2)^2 - 5 \times 123.28}$$

$$\rightarrow b_1 = 0.2698$$

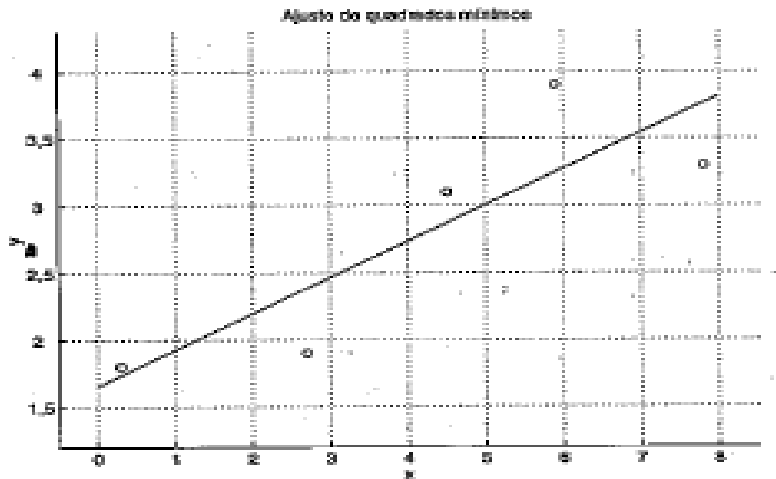
$$b_0 = \frac{\sum y_i - b_1 \sum x_i}{n} = \frac{14.0 - 0.2698 \times 21.2}{5} \rightarrow b_0 = 1.6560$$

Resultado do ajuste por quadrados mínimos

i	x_i	y_i	u_i	d_i
1	0.3	1.8	1.7369	0.0631
2	2.7	1.9	2.3845	-0.4845
3	4.5	3.1	2.8701	0.2299
4	5.9	3.9	3.2478	0.6522
5	7.8	3.3	3.7604	-0.4604

- $D(1.6560; 0.2698) = 0.9289$

Ajuste de quadrados mínimos $u = 1.6560 + 0.2698x$



Comparação entre os modelos

- Considerando

$$D(1.6560; 0.2698) = 0.9280 < D(1.5; 0.25) = 1.2300 < \\ D(1.74; 0.2) = 1.3164$$

- O ajuste de quadrados mínimos é o melhor dos três modelos propostos por que tem o menor desvio D

Qualidade do ajuste

Parâmetros

- Vamos estudar dois parâmetros para verificar a qualidade do ajuste da regressão linear
 - Coeficiente de terminação R^2
 - Variância residual σ^2

Coeficiente de determinação

- Seja

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

- A expressão para o i -ésimo ponto, onde

$$u_i = b_0 + b_1 x_i \text{ e } \bar{y} = \frac{1}{n} \left(\sum_{i=1}^n y_i \right)$$

- Tomando o quadrado em ambos os termos da igualdade, temos

$$(y_i - \bar{y})^2 = (y_i - u_i)^2 + (u_i - \bar{y})^2 + 2(y_i - u_i)(u_i - \bar{y})$$

Coeficiente de determinação, cont.

- Calculando o somatório para $i = 1, 2, \dots, n$, obtemos

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - u_i)^2 + \sum_{i=1}^n (u_i - \bar{y})^2 + 2 \sum_{i=1}^n (y_i - u_i)(u_i - \bar{y}) \quad (3)$$

- Logo,

$$\begin{aligned} \sum_{i=1}^n (y_i - u_i)(y_i - \bar{y}) &= \sum_{i=1}^n d_i(b_0 + b_1 x_i - \bar{y}) \rightarrow \\ \sum_{i=1}^n (y_i - u_i)(y_i - \bar{y}) &= (b_0 - \bar{y}) \sum_{i=1}^n d_i + b_1 \sum_{i=1}^n x_i d_i \end{aligned} \quad (4)$$

Coeficiente de determinação, cont.

- Por outro lado

$$\sum_{i=1}^n d_i = \sum_{i=1}^n (y_i - b_0 - b_1 x_i) = \sum_{i=1}^n y_i - nb_0 - b_1 \sum_{i=1}^n x_i$$

- A partir da expressão de b_0 da Equação (2)

$$\sum_{i=1}^n d_i = 0 \quad (5)$$

- Isto é, a soma dos desvios obtidos por quadrados mínimos é zero

Coeficiente de determinação, cont.

- Adicionalmente,

$$\sum_{i=1}^n x_i d_i = \sum_{i=1}^n (x_i (y_i - b_0 - b_1 x_i)) = \sum_{i=1}^n (x_i y_i - b_0 x_i - b_1 x_i^2)$$

- Substituindo o valor de b_0 dado em (2), temos

$$\begin{aligned}\sum_{i=1}^n x_i d_i &= \sum_{i=1}^n x_i y_i - \frac{1}{n} (\sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i) \sum_{i=1}^n x_i - b_1 \sum_{i=1}^n x_i^2 \\ &= \frac{1}{n} \left(n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i + b_1 (\sum_{i=1}^n x_i)^2 - n b_1 \sum_{i=1}^n x_i^2 \right) \rightarrow \\ \sum_{i=1}^n x_i d_i &= \frac{1}{n} \left(n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i + b_1 \left((\sum_{i=1}^n x_i)^2 - n \sum_{i=1}^n x_i^2 \right) \right)\end{aligned}$$

Coeficiente de determinação, cont.

- A partir do valor de b_1 em (2)

$$\sum_{i=1}^n x_i d_i = 0 \quad (6)$$

- Substituindo (5) e (6) em (4), temos

$$\sum_{i=1}^n (y_i - u_i)(u_i - \bar{y}) = 0$$

- Logo, (3) torna-se

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - u_i)^2 + \sum_{i=1}^n (u_i - \bar{y})^2$$

- Onde

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \text{SQTot (soma de quadrados total)}$$

$$\sum_{i=1}^n (y_i - u_i)^2 = \text{SQRes (soma de quadrados residual)}$$

$$\sum_{i=1}^n (u_i - \bar{y})^2 = \text{SQReg (soma de quadrados devido a regressão)}$$

Cálculo do R^2

- Uma maneira de avaliar a qualidade do ajuste do modelo $u = b_0 + b_1x$ aos dados é tomando a razão entre $SQReg$ e $SQTot$

$$R^2 = \frac{SQReg}{SQTot} = \frac{SQTot - SQRes}{SQTot} \rightarrow R^2 = 1 - \frac{SQRes}{SQTot}$$

- R^2 é denominado coeficiente de determinação $0 \leq R^2 \leq 1$
- Quanto mais próximo de 1 melhor é o ajuste

Proporção da variação total

- Considerando que

$$D(b_0, b_1) = \sum_{i=1}^n (y_i - u_i)^2 = \sum_{i=1}^n d_i^2 \quad (7)$$

- E que

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (y_i^2 - 2\bar{y} \sum_{i=1}^n y_i + n\bar{y}^2) \rightarrow \\ \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n y_i^2 - \frac{1}{n} (\sum_{i=1}^n y_i)^2, \end{aligned}$$

- Então

$$R^2 = 1 - \frac{D(b_0, b_1)}{\sum y_i^2 - \frac{1}{n} (\sum y_i)^2} \quad (8)$$

- R^2 é a proporção da variação total dos dados em torno da média $\bar{y}$ que é explicada pelo modelo de regressão

Exercício 1

- Calcular a reta de quadrados mínimos a partir dos dados da tabela

x	1.2	2.5	3.0	4.1	6.2	7.1	8.8	9.5
y	6.8	6.1	9.9	9.7	12.1	17.9	18.0	21.5

Variância residual

Variância residual

- Parâmetro importante para medir a qualidade do ajuste é a variância residual σ^2 definida por

$$\sigma^2 = \frac{D(\beta_0, \beta_1)}{n - p} \quad (9)$$

- onde $D(b_0, b_1)$ é dado por (7), n é o número de pontos e p é o número de parâmetros estimados
- No caso da regressão linear simples

$$u = b_0 + b_1 x$$

- Temos que $p = 2$

Efeito de mais parâmetros

- Tanto o numerador quanto o denominador de (9) vão diminuir se forem introduzidos mais parâmetros no modelo
- Porém, a redução global de σ^2 define se mais parâmetros devem ou não ser incorporados ao modelo

Exemplo 2

- Calcular o coeficiente de determinação e a variância residual, usando os dados das tabelas

i	x_i	y_i	x_i^2	$x_i y_i$	y_i^2
1	0.3	1.8	0.09	0.54	3.24
2	2.7	1.9	7.29	5.13	3.61
3	4.5	3.1	20.25	13.95	9.61
4	5.9	3.9	34.81	23.01	15.21
5	7.8	3.3	60.84	25.74	10.58
Σ	21.2	14.0	123.28	68.37	42.56

i	x_i	y_i	u_i	d_i
1	0.3	1.8	1.7369	0.0631
2	2.7	1.9	2.3845	-0.4845
3	4.5	3.1	2.8701	0.2299
4	5.9	3.9	3.2478	0.6522
5	7.8	3.3	3.7604	-0.4604

Solução

$$R^2 = 1 - \frac{0.9289}{42.56 - (11.0)^2/5} \rightarrow R^2 = 0,7235$$

- A variância residual é

$$\sigma^2 = \frac{0.9289}{5 - 2} \rightarrow \sigma^2 = 0,3096$$