CAPÍTULO 9 - Regressão linear e correlação

Veremos nesse capítulo os seguintes assuntos nessa ordem:

- Correlação amostral
- Regressão Linear Simples
- Regressão Linear Múltipla

Correlação Amostral

Serve para estudar o comportamento conjunto de duas variáveis quantitativas distintas. Ou, em outras palavras, mede o grau de associação entre duas variáveis aleatórias X e Y.

OBS.: não há, nesse caso, preocupação em apresentar alguma forma funcional entre as variáveis, se houver.

Exemplos: (apresentados em aula)

Para o estudo do comportamento conjunto de duas variáveis poderiam ser usados:

a) O Diagrama de dispersão

Representação gráfica do conjunto de dados. Nada mais é do que a representação dos pares de valores num sistema cartesiano. Veja exemplo a seguir.

Em síntese três situações marcantes poderiam acontecer:

- Se, quando uma das variáveis "cresce", a outra, em média, também "cresce", dizemos que entre as duas variáveis existe correlação positiva, tanto mais forte quanto mais perto de uma reta imaginária os pontos estiverem;
- Se, quando uma das variáveis "cresce", a outra, em média, também "decresce", dizemos que entre as duas variáveis existe correlação negativa, tanto mais forte quanto mais perto de uma reta imaginária os pontos estiverem;
- Se os pontos estiverem dispersos, sem definição de direção, dizemos que a correlação é muito baixa, ou mesmo nula. As variáveis nesse caso são ditas não correlacionadas.

b) O coeficiente de correlação

É um valor numérico, uma medida, para o grau de associação entre duas variáveis.

Se for observada uma associação entre as variáveis quantitativas (a partir de um diagrama de dispersão, por exemplo), é muito útil quantificar essa associabilidade.

Existem muitos tipos de associação possíveis, e aqui iremos apresentar o tipo de relação mais simples, que é o linear. Iremos julgar o quanto a nuvem de pontos do diagrama de dispersão se aproxima de uma reta.

Sejam duas amostras relativas às variáveis X e Y, dadas a seguir:

X_i	X_1	X_2	$\cdots$	X_n
Y_i	Y_1	Y_2	$\cdots$	Y_n

O coeficiente de correlação entre os valores de X e Y é dado por:

$$r_{XY} = \frac{C\hat{O}V(X,Y)}{\sqrt{\hat{V}(X)\cdot\hat{V}(Y)}} = \frac{\frac{SPD_{XY}}{n-1}}{\sqrt{\frac{SQD_X}{n-1}\cdot\frac{SQD_Y}{n-1}}} = \frac{SPD_{XY}}{\sqrt{SQD_X . SQD_Y}} \quad , \quad -1 \le r_{XY} \le 1$$

em que:

$$SPD_{XY} = \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}$$

$$SQD_X = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n} \quad \text{e} \quad SQD_Y = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

Para o exemplo:

Amostra A	4	8	3	9	7	5
Amostra B	1	5	2	14	3	11

$$SPD_{AB} = \sum_{i=1}^{n} A_i B_i - \frac{\left(\sum_{i=1}^{n} A_i\right)\left(\sum_{i=1}^{n} B_i\right)}{n} = 252 - \frac{(36)(36)}{6} = 36$$

$$SQD_A = \sum_{i=1}^{n} A_i^2 - \frac{\left(\sum_{i=1}^{n} A_i\right)^2}{n} = 244 - \frac{(36)^2}{6} = 28$$

$$SQD_B = \sum_{i=1}^{n} B_i^2 - \frac{\left(\sum_{i=1}^{n} B_i\right)^2}{n} = 356 - \frac{(36)^2}{6} = 140$$

$$r_{AB} = \frac{SP_{AB}}{\sqrt{SQD_A . SQD_B}} = \frac{36}{\sqrt{(28)\,(140)}} = 0{,}5750$$

Regressão linear

A análise de regressão consiste na realização de uma análise estatística com o objetivo de verificar a existência de uma relação funcional entre uma variável dependente com uma ou mais variáveis independentes. Em outras palavras consiste na obtenção de uma equação que tenta explicar a variação da variável dependente pela variação do(s) nível(is) da(s) variável(is) independente(s).

Para tentar estabelecer uma equação que representa o fenômeno em estudo pode-se fazer um gráfico, chamado de diagrama de dispersão, para verificar como se comportam os valores da variável dependente (Y) em função da variação da variável independente (X).

O comportamento de Y em relação a X pode se apresentar de diversas maneiras: linear, quadrático, cúbico, exponencial, logarítmico, etc... . Para se estabelecer o modelo para explicar o fenômeno, deve-se verificar qual tipo de curva e equação de um modelo matemático que mais se aproxime dos pontos representados no diagrama de dispersão.

Contudo, pode-se verificar que os pontos do diagrama de dispersão, não vão se ajustar perfeitamente à curva do modelo matemático proposto. Haverá na maior parte dos pontos, uma distância entre os pontos do diagrama e a curva do modelo matemático. Isto acontece, devido ao fato do fenômeno que está em estudo, não ser um fenômeno matemático e sim um fenômeno que está sujeito a influências que acontecem ao acaso. Assim, o objetivo da regressão é obter um modelo matemático que melhor se ajuste aos valores observados de Y em função da variação dos níveis da variável X.

No entanto o modelo escolhido deve ser coerente com o que acontece na prática. Para isto, deve-se levar em conta as seguintes considerações no momento de se escolher o modelo:

-o modelo selecionado deve ser condizente tanto no grau como no aspecto da curva, para representar em termos práticos, o fenômeno em estudo;

-o modelo deve conter apenas as variáveis que são relevantes para explicar o fenômeno;

Como foi dito anteriormente, os pontos do diagrama de dispersão ficam um pouco distantes da curva do modelo matemático escolhido. Um dos métodos que se pode utilizar para obter a relação funcional, se baseia na obtenção de uma equação estimada de tal forma que as distâncias entre os pontos do diagrama e os pontos da curva do modelo matemático, no todo, sejam as menores possíveis. Este método é denominado de Método dos Mínimos Quadrados (MMQ). Em resumo por este método a soma de quadrados das distâncias entre os pontos do diagrama e os respectivos pontos na curva da equação estimada é minimizada, obtendo-se, desta forma, uma relação funcional entre X e Y, para o modelo escolhido, com um mínimo de erro possível.

MODELO LINEAR DE 1º GRAU (Regressão Linear Simples)

O modelo estatístico para esta situação seria:

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

em que:

$Y_i =$ valor observado para a variável dependente Y no i-ésimo nível da variável independente X.

$\beta_0 =$ constante de regressão. Representa o intercepto da reta com o eixo dos Y.

$\beta_1 =$ coeficiente de regressão. Representa a variação de Y em função da variação de uma unidade da variável X.

$X_i =$ i-ésimo nível da variável independente X $(i = 1, 2, \ldots, n)$

$e_i =$ é o erro que está associado à distância entre o valor observado Y_i e o correspondente ponto na curva, do modelo proposto, para o mesmo nível i de X.

Para se obter a equação estimada, vamos utilizar o MMQ, visando a minimização dos erros. Assim, tem-se que:

$$e_i = Y_i - \beta_0 - \beta_1 X_i$$

elevando ambos os membros da equação ao quadrado,

$$e_i^2 = \left[Y_i - \beta_0 - \beta_1 X_i\right]^2$$

aplicando o somatório,

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[Y_i - \beta_0 - \beta_1 X_i\right]^2 \qquad (1)$$

Por meio da obtenção de estimadores de β_0 e β_1, que minimizem o valor obtido na expressão anterior (1), é possível alcançar a minimização da soma de quadrados dos erros.

Para se encontrar o mínimo para uma equação, deve-se derivá-la em relação à variável de interesse e igualá-la a zero. Derivando então a expressão (1) em relação a β_0 e β_1, e igualando-as a zero, poderemos obter duas equações que, juntas, vão compor o chamado sistemas de equações normais. A solução desse sistema fornecerá:

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}} = \frac{SPD_{xy}}{SQD_x} \quad \text{e} \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Uma vez obtidas estas estimativas, podemos escrever a equação estimada:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Exemplos:

1) Para verificar se existe relação linear de primeiro grau entre umidade relativa (UR) do ar de secagem de sementes e a germinação das mesmas, um pesquisador realizou um experimento com 4 valores diferentes para a %UR do ar, obtendo-se os seguintes dados (dados hipotéticos)

% UR	20	30	40	50
% germinação	94	96	95	97

a) Verificar se existe efeito da UR do ar de secagem na % de germinação. Usar α = 5%.

b) Qual seria a % de germinação esperada quando UR = 45 %?

c) Como poderia ser apresentada, num relatório técnico, a equação de regressão ajustada para esse exemplo?

R.: a) $\hat{\beta}_0$ = 92,7; $\hat{\beta}_1$ = 0,08. F = 3,55; t = 1,88. b) 95,5 %

2) Foi realizado uma análise de regressão para investigar a existência de ralação linear simples entre a temperatura superficial de uma estrada (X) medida em graus F e a deformação da pavimentação (Y) medida segundo uma técnica especial. Baseado nas seguintes informações pede-se:

n = 20; $\sum y_i$ = 12,75; $\sum y_i^2$ = 8,86; $\sum x_i$ = 1478; $\sum x_i^2$ =143215,8; e $\sum x_i y_i$ = 1083,67

a) Calcule as estimativas dos parâmetros da regressão. Apresente a equação ajustada num gráfico;

b) Use a equação para estimar qual deformação haveria na pavimentação quando a temperatura superficial fosse de 85 graus F.

c) Qual seria a mudança esperada na deformação da pavimentação para uma mudança de 1^o F na temperatura superficial?

d) Suponha que a temperatura seja medida em graus C ao invés de graus F. Qual seria a nova equação ajustada resultante? Lembre-se: C = 5(F – 32)/9.

e) Qual seria a mudança esperada na deformação da pavimentação para uma mudança de 1^o C na temperatura superficial?

Exercício Proposto

Os dados a seguir provêm de um experimento para testar o desempenho de uma máquina industrial. O experimento utilizou uma mistura de óleo diesel e gás, derivados de materiais destilados orgânicos. O valor da capacidade da máquina em cavalo vapor (HP) foi coletado a diversas velocidades medidas em rotações por minuto (rpm $\times$ 100).

X	Y	X	Y	X	Y	X	Y
22,0	64,03	15,0	46,85	18,0	52,90	15,0	45,79
20,0	62,47	17,0	51,17	16,0	48,84	17,0	51,17
18,0	54,94	19,0	58,00	14,0	42,74	19,0	56,65
16,0	48,84	21,0	63,21	12,0	36,63	21,0	62,61
14,0	43,73	22,0	64,03	10,5	32,05	23,0	65,31
12,0	37,48	20,0	62,63	13,0	39,68	24,0	63,89

X = velocidade Y = capacidade

Admitindo-se que as variáveis X e Y estão relacionadas de acordo com o modelo $Y_i = \beta_0 + \beta_1 X_i + e_i$, pede-se:

(a) Obter a equação ajustada e traçar seu gráfico. Mostre também o diagrama de dispersão;
(b) Calcule o coeficiente de determinação e interprete;
(c) Verifique que $\sum_{i=1}^{n} \hat{e}_i = 0$;
(d) Verifique que $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$;
(e) Interprete a estimativa obtida para β_1;
(f) Determine a estimativa de Y para X = 15,5.

COEFICIENTE DE DETERMINAÇÃO

O coeficiente de determinação, também conhecido como R^2, ou simplesmente r^2 para o caso de regressão linear simples, fornece uma informação auxiliar ao resultado da análise de variância da regressão (apresentado a seguir), como uma maneira de se verificar se o modelo proposto é adequado ou não para descrever o fenômeno.

O R^2 é obtido por:

$$R^2 = \frac{SQ\,\mathrm{Re}\,g}{SQTotal}$$

O valor de R^2 varia no intervalo de 0 a 1. Valores próximos de 1 indicam que o modelo proposto é adequado para descrever o fenômeno.

O R^2 indica a proporção (ou porcentagem) da variação de Y que é "explicada" pela regressão, ou quanto da variação na variável dependente Y está sendo "explicada" pela variável independente X.

TESTE DE HIPÓTESE NA REGRESSÃO LINEAR SIMPLES

Após ajustar uma equação de regressão devemos verificar sua adequabilidade, por meio de testes de hipóteses para os parâmetros do modelo e/ou a construção de intervalos de confiança. Para tal intento precisamos da pressuposição adicional de que os erros tenham distribuição normal.

Como temos dois parâmetros no modelo $Y_i = \beta_0 + \beta_1 X_i + e_i$, poderíamos realizar os seguintes testes:

a) H_0: $\beta_1 = \beta_1^*$ versus H_a: $\beta_1 \neq \beta_1^*$
b) H_0: $\beta_0 = \beta_0^*$ versus H_a: $\beta_0 \neq \beta_0^*$

Em cada caso a estatística do teste e as conclusões seriam:

a) $t_{calc} = \dfrac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\hat{V}(\hat{\beta}_1)}}$, onde $\hat{V}(\hat{\beta}_1) = \dfrac{\hat{\sigma}^2}{SQD_x}$

- regra de decisão: Se $|\, t_{calc} \,| \geq t_{(\alpha/2,\, n-2)} \Rightarrow$ rejeita H_0

b) $t_{calc} = \dfrac{\hat{\beta}_0 - \beta_0^*}{\sqrt{\hat{V}(\hat{\beta}_0)}}$, onde $\hat{V}(\hat{\beta}_0) = \hat{\sigma}^2\left(\dfrac{1}{n} + \dfrac{\overline{X}^2}{SQD_x}\right)$

- regra de decisão: Se $|\, t_{calc} \,| \geq t_{(\alpha/2,\, n-2)} \Rightarrow$ rejeita H_0

OBS.: $\hat{\sigma}^2$ = estimativa da variância dos erros = $\dfrac{SQ\,\mathrm{Res}}{n-2} = \dfrac{SQD_y - \hat{\beta}_1 SPD_{xy}}{n-2}$

Um caso especial muito importante seria: H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$. Essas hipóteses estão relacionadas com a significância da regressão. Não rejeitar H_0 é equivalente a concluir que não há relação linear entre X e Y. Por outro lado, se H_0: $\beta_1 = 0$ for rejeitado indicaria que X é importante para explicar a variabilidade em Y. Veja ilustrações apresentadas em aula.

De maneira alternativa poderíamos testar a significância da regressão pelo método da Análise de Variância (ANOVA).

O método da ANOVA consiste em fazer uma partição da variabilidade total da variável resposta Y em outros componentes de acordo com o modelo e o teste a ser feito. Assim a seguinte identidade pode ser verificada:

$$\sum (Y_i - \overline{Y})^2 = \sum (\hat{Y}_i - \overline{Y})^2 + \sum (Y_i - \hat{Y})^2,$$

ou, em outra palavras,

SQTotal = SQRegressão + SQResíduo.

Onde

SQTotal = variação total em Y = SQD_Y

SQRegressão = variação em Y explicada pela regressão ajustada = $\hat{\beta}_1 SPD_{XY}$

de modo que

SQResíduo = SQRes = variação não explicada pela regressão = SQD_Y - $\hat{\beta}_1 SPD_{XY}$

Baseado nessa identidade o seguinte quadro pode ser montado:

FV	GL	SQ	QM	F
Regressão	1	SQReg	QMReg = SQReg	$\dfrac{QM\,\mathrm{Reg}}{QM\,\mathrm{Res}}$
Resíduo, ou Independente da Regressão	n – 2	SQRes	QMRes = $\dfrac{SQ\,\mathrm{Res}}{n-2}$	-
Total	n – 2	SQTotal		

A estatística F obtida no quadro acima serve para testar a significância da regressão, ou seja, testar H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$.

- regra de decisão: Se $F_{calc} \geq F_{(\alpha,\, 1,\, n-2)} \Rightarrow$ rejeita H_0

OBS.: Para H_0: $\beta_1 = 0$ temos que $(t_{calc})^2 = F_{calc}$

A equação estimada obtida, apenas estabelece uma relação funcional, entre a variável dependente e a variável independente, para representar o fenômeno em estudo. Portanto a simples obtenção da equação estimada não responde ao pesquisador se a variação da variável independente influencia significativamente na variação da variável dependente.

Para se responder a esta pergunta, é necessário realizar um teste estatístico para as estimativas dos coeficientes da equação de regressão estimada. Um teste que pode ser realizado para verificar tal fato é o teste F da análise de variância. Portanto, é necessário realizar uma análise de variância dos dados observados, em função do modelo proposto.

O quadro para a análise de variância para a regressão é do seguinte tipo:

FV	GL	SQ	QM	F
Regressão	P	SQReg	$\frac{SQ\,\mathrm{Re}\,g}{p}$	$\frac{QM\,\mathrm{Re}\,gr}{QMInd}$
Independente da Regressão	n – 1 – p	SQInd	$\frac{SQInd}{n-1-p}$	-
Total	n – 1	SQTotal		

em que:

- p = nº de coeficientes de regressão (não inclui o β_0)

- n = nº de observações.

As fórmulas para a obtenção das somas de quadrados total e da soma de quadrados do independente da regressão são as mesmas, tanto para o modelo linear de 1º grau quanto para o de 2º grau, as quais são dadas a seguir:

$$SQTotal = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

SQIndependente da Regressão = SQTotal - SQRegressão

Já a soma de quadrados para a regressão varia de acordo com o modelo em teste. Assim tem-se que, para o modelo linear de 1º grau, a soma de quadrados da regressão é obtida por:

$$SQ\,\mathrm{Re}\,gressão = \hat{\beta}_0 \sum_{i=1}^{n} Y_i + \hat{\beta}_1 \sum_{i=1}^{n} Y_i X_i - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

Para o modelo linear de 2º grau, a soma de quadrados da regressão é dada por:

$$\text{SQ Regressão} = \hat{\beta}_0 \sum_{i=1}^{n} Y_i + \hat{\beta}_1 \sum_{i=1}^{n} Y_i X_i + \hat{\beta}_2 \sum_{i=1}^{n} Y_i X_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

As hipóteses estatísticas para o teste F, são as seguintes:

$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$, o que significa dizer que as p variáveis independentes não exercem influência na variável dependente, segundo o modelo proposto.

$H_a : \beta_i \neq 0$, para pelo menos um i, o que significa dizer que pelo menos uma das p variáveis independentes exerce influência na variável dependente, segundo o modelo proposto.

O valor de F da análise de variância, deve ser comparado, com o valor de F tabelado (F_{tab}), o qual se obtém na tabela da distribuição F de acordo com o nível de significância do teste, e o número de graus de liberdade para a regressão e independente da regressão, ou seja:

$F_{tab} = F_\alpha(p; n-1-p)$.

A regra decisória para o teste F é:

- Se $F \geq F_{tab} \Rightarrow$ Rejeita-se H_0 ao nível de significância que foi realizado o teste. Pode-se inferir que o modelo proposto é adequado para descrever o fenômeno.

- Se $F < F_{tab} \Rightarrow$ Não rejeita-se H_0 ao nível de significância que foi realizado o teste. Pode-se inferir que o modelo proposto não é adequado para descrever o fenômeno.

Exercícios Propostos:

1) (questão de prova do II/2000) Para estudar a relação entre Y (número total de horas necessárias à montagem da parte de uma estrutura) e X (número total de operações de furar e rebitar), registraram-se os dados da tabela abaixo.

estudo	A	B	C	D	E	F	G	H	I
X	236	80	127	445	180	343	305	488	170
Y	5,1	1,7	3,3	6,0	2,9	5,9	7,0	9,4	4,8

Para facilitar seus cálculos considere as seguintes informações:

$\sum_i x_i$ = 2374; $\sum_i y_i$ = 46,1; $\sum_i x_i^2$ = 786368; $\sum_i y_i^2$ = 279,41; $\sum_i x_i y_i$ = 14512,6

também, SPD_{xy} = 2352,4444; SQD_x = 160159,5556

Pede-se:

a) Obter a equação de regressão ajustada para o modelo $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

R.: $\hat{Y}$ = 1,271 + 0,0146X

b) Interpretar as estimativas obtidas dos parâmetros da regressão.

c) Calcular o coeficiente de determinação para o modelo ajustado. Faça a interpretação apropriada para esse resultado. R.: 79,9%

d) A análise de variância (ANOVA) da regressão pode ser resumida no seguinte quadro

F.V.	g.l.	SQ	QM	F
Regressão	1	34,59	34,59	
Resíduo	7	8,68	1,24	
Total	8	43,27		

Uma maneira de verificar a significância da regressão ajustada é por meio da ANOVA apresentada acima. Apresente a hipótese a ser testada pela ANOVA e realize o teste apropriado (use $\alpha = 5\%$) para testar essa hipótese.

e) Se fosse concluído que podemos considerar $\beta_1 = 0$, como deveria ser reescrito o modelo ajustado? Justifique.

Regressão linear múltipla

A regressão múltipla envolve três ou mais variáveis, ou seja, uma única variável dependente (Y) e duas ou mais variáveis independentes ou explanatórias ou covariáveis ou regressoras (X_i, i = 1, 2, ...). A teoria é uma extensão da análise de regressão linear simples. De modo similar a análise tem por objetivo estabelecer uma equação que possa ser usada para predizer valores de Y para valores dados das diversas variáveis independentes. A finalidade das variáveis independentes adicionais é melhorar a capacidade de predição em confronto com a regressão linear simples. A técnica de cálculo é bastante complicada e pode ser facilitada com o auxílio de álgebra de matrizes.

O modelo

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

é chamado de modelo de regressão linear múltipla com k variáveis regressoras. Os parâmetros β_i (i = 1 a k) são chamados de coeficientes de regressão parciais.

Veremos dois exemplos envolvendo regressão linear múltipla.

MODELO LINEAR DE 2º GRAU

O modelo estatístico para esta situação seria:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + e_i$$

em que:

$Y_i =$ valor observado para a variável dependente Y no i-ésimo nível da variável independente X.

$\beta_0 =$ constante de regressão.

$\beta_1 =$ coeficiente de regressão.

$\beta_2 =$ coeficiente de regressão.

$X_i =$ i-ésimo nível da variável independente X $(i = 1,2,\ldots,n)$

$X_i^2 =$ i-ésimo nível da variável independente X, elevado ao quadrado

$e_i =$ é o erro que está associado à distância entre o valor observado Y_i e o correspondente ponto na curva para o mesmo nível i de X.

Utilizando o MMQ, no modelo de 2º grau, chegar-se-á ao seguinte sistema de equações normais, para se obter as estimativas de $\beta_0, \beta_1\ e\ \beta_2$:

$$\begin{cases} \sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2 \\ \sum_{i=1}^{n} Y_i X_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 + \hat{\beta}_2 \sum_{i=1}^{n} X_i^3 \\ \sum_{i=1}^{n} Y_i X_i^2 = \hat{\beta}_0 \sum_{i=1}^{n} X_i^2 + \hat{\beta}_1 \sum_{i=1}^{n} X_i^3 + \hat{\beta}_2 \sum_{i=1}^{n} X_i^4 \end{cases}$$

Uma vez obtidas estas estimativas, podemos escrever a equação estimada:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2$$