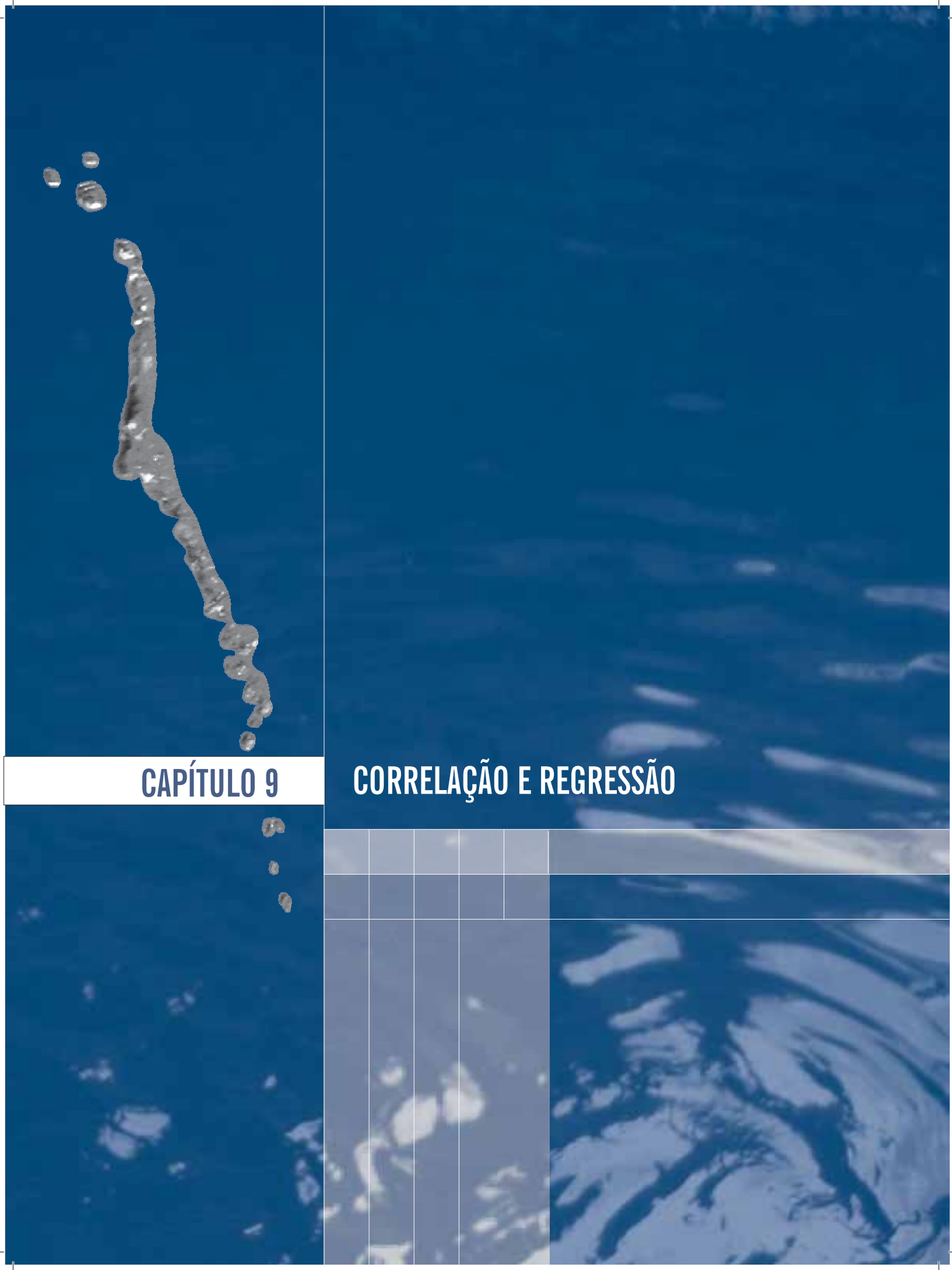

CAPÍTULO 9

CORRELAÇÃO E REGRESSÃO

CAPÍTULO 9 CORRELAÇÃO E REGRESSÃO

Existe um conjunto de métodos estatísticos que visam estudar a associação entre duas ou mais variáveis aleatórias. Dentre tais métodos, a teoria da regressão e correlação ocupa um lugar de destaque por ser o de uso mais difundido. Neste capítulo serão abordados os fundamentos dos métodos estatísticos da correlação e regressão, com vistas à sua aplicação em hidrologia. O objetivo deste capítulo é o de apresentar os conceitos básicos que permitam ao leitor realizar estudos de correlação e regressão linear entre duas ou mais variáveis aleatórias hidrológicas.

Na engenharia de recursos hídricos, algumas questões referem-se ao conhecimento da associação e do grau de associação entre duas ou mais variáveis, como por exemplo, as relações (i) entre as intensidades, as durações e as freqüências das precipitações intensas (ii) entre as vazões médias anuais e as áreas de drenagem ou (iii) entre as alturas anuais de precipitação e as altitudes dos postos pluviométricos. Nesses estudos, o primeiro objetivo é o de analisar o comportamento simultâneo das variáveis, tomadas duas a duas, verificando se a variação positiva (ou negativa) de uma delas está associada a uma variação positiva (ou negativa) da outra, ou mesmo, se não há nenhuma forma de dependência entre elas. Nesse sentido, uma primeira abordagem exploratória é a elaboração de um diagrama de dispersão entre as observações simultâneas das variáveis. O diagrama de dispersão permite visualizar o grau de associação entre as variáveis e a tendência de variação conjunta que apresentam. A Figura 9.1 apresenta alguns exemplos de variação conjunta entre duas variáveis.

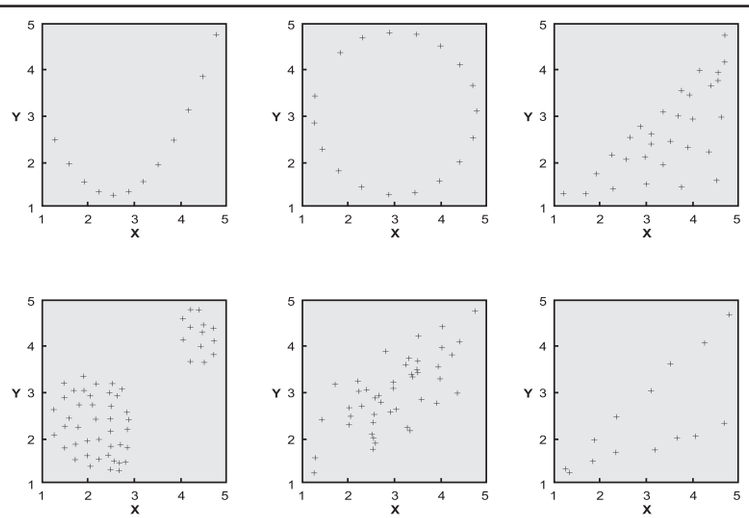

Figura 9.1 – Exemplos de relacionamento (Adaptado de Helsel e Hirsh, 1992)

A medida da variação conjunta das variáveis ou co-variação observada em um diagrama de dispersão é a correlação entre as duas variáveis. Essa medida é realizada numericamente por meio dos coeficientes de correlação que representam o grau de associação entre duas variáveis continuas. As medidas genéricas de correlação, freqüentemente são designadas por ρ, são adimensionais e variam entre -1 e +1. No caso de $\rho = 0$, não existe correlação entre as duas variáveis. Quando $\rho > 0$, a correlação é positiva e uma variável aumenta quando a outra cresce. A correlação é negativa, $\rho < 0$,quando as variáveis variam em direções opostas.

A correlação é chamada de monotônica se uma das variáveis aumenta ou diminui sistematicamente quando a outra decresce, com associações que podem ter forma linear ou não linear. A Figura 9.2 apresenta exemplos de correlações monotônicas não lineares e não monotônicas.

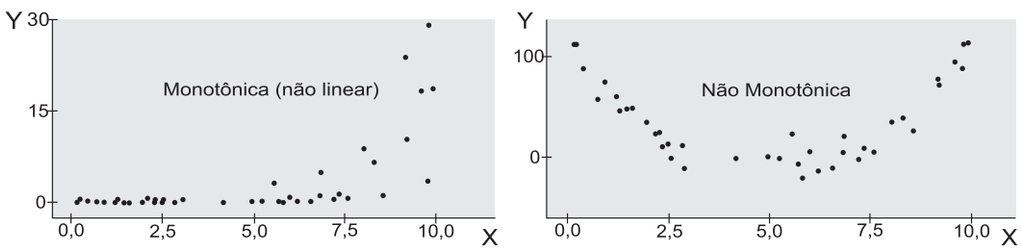

Figura 9.2 – Exemplos de correlações (Adaptado de Helsel e Hirsh, 1992)

É importante salientar que variáveis altamente correlacionadas *não* apresentam necessariamente qualquer relação de causa e efeito. A correlação representa simplesmente a tendência que as variáveis apresentam quanto à sua variação conjunta. Assim, a medida da correlação não indica necessariamente que há evidências de relações causais entre duas variáveis. As evidências de relações causais devem ser obtidas a partir do conhecimento dos processos envolvidos. Obviamente haverá casos em que uma variável está na origem da outra, tais como aqueles que associam a precipitação e o escoamento superficial em uma dada bacia. Entretanto, existirão situações em que as variáveis apresentam a mesma causa, como, por exemplo, a eventual forte correlação entre as vazões médias mensais de duas bacias vizinhas não significa que a mudança da vazão de uma delas é causada pela alteração da outra; certamente, as alterações são causadas por fatores comuns às duas bacias.

9.1 – Coeficiente de Correlação Linear de Pearson

Duas variáveis apresentam uma correlação linear quando os pontos do diagrama de dispersão se aproximam de uma reta. Essa correlação pode ser positiva (para valores crescentes de *X*, há uma tendência a valores também crescentes de *Y*) ou negativa (para valores crescentes de *X*, a tendência é observarem-se valores decrescentes de *Y*). As correlações lineares positivas e negativas encontram-se ilustradas na Figura 9.3.

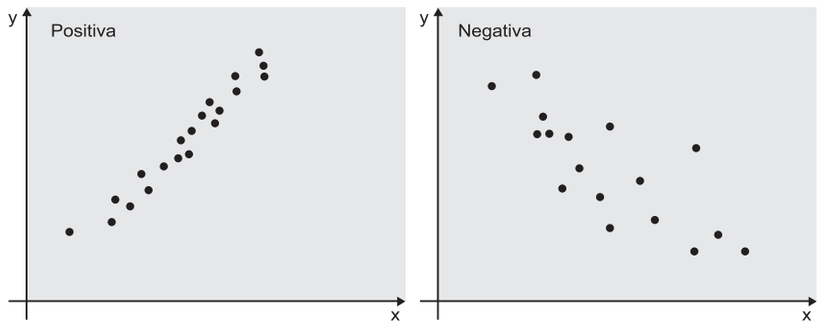

Figura 9.3 – Correlações Lineares Positivas e Negativas

O coeficiente de correlação linear, também chamado de covariância normalizada e representado por ρ, é expresso por:

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \tag{9.1}$$

onde, $\sigma_{X,Y}$ é a covariância entre as variáveis *X* e *Y*; σ_X e σ_Y são os desvios-padrão das variáveis *X* e *Y*, respectivamente.

Quando duas variáveis, *X* e *Y*, são estatisticamente independentes, o coeficiente de correlação linear é igual a zero, $\rho = 0$. Entretanto a recíproca não é verdadeira, ou seja, se o coeficiente de correlação linear é igual a zero, $\rho = 0$, isso não significa que as variáveis são independentes. Trata-se de uma decorrência do fato de que o coeficiente de correlação linear, ρ, é uma medida da dependência linear entre as variáveis *X* e *Y*, e, em algumas situações, *X* e *Y* podem apresentar dependência funcional não linear.

A covariância entre duas variáveispode ser estimada pela equação 9.2 e representa uma medida possível do grau e do sinal da correlação.

$$s_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{9.2}$$

onde, s_{xy} é a covariância amostral entre as variáveis X e Y; $\bar{x}$ e $\bar{y}$ são as médias aritméticas de cada uma das variáveis; n é o tamanho da amostra; x_i e y_i são as observações simultâneas das variáveis.

Entretanto, admitindo-se que a distribuição conjunta das variáveis X e Y é uma distribuição normal bivariada, torna-se conveniente utilizar, como medida da correlação, o chamado coeficiente de correlação linear de Pearson cujo estimador é apresentado a seguir:

$$r = \frac{s_{X,Y}}{s_X s_Y} \tag{9.3}$$

Na equação 9.3, r é coeficiente de correlação linear ($-1 \leq r \geq 1$), s_{XY} é covariância entre as variáveis, s_X e s_Y são os desvios-padrão das amostras calculados pelas equações:

$$s_X = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{9.4}$$

$$s_Y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} \tag{9.5}$$

O coeficiente de correlação linear de Pearson é adimensional e varia entre -1 e +1, o que não ocorre com a covariância. Assim, as unidades adotadas pelas variáveis não afetam o valor do coeficiente de correlação. Caso os dados se alinhem perfeitamente ao longo de uma reta com declividade positiva teremos a correlação linear positiva perfeita com o coeficiente de Pearson igual a 1. A correlação linear negativa perfeita ocorre quando os dados se alinham perfeitamente ao longo de uma reta com declividade negativa e o coeficiente de correlação de Pearson é igual a -1. O significado de valores intermediários é facilmente percebido. A Figura 9.4 apresenta alguns diagramas de dispersão com os respectivos valores do coeficiente de correlação.

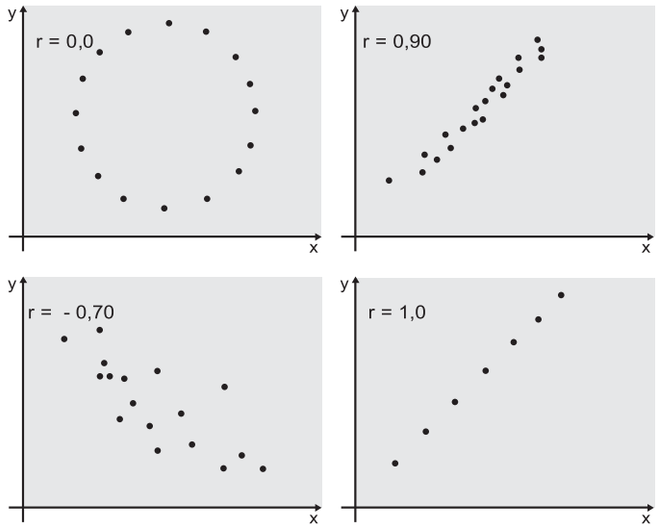

Figura 9.4 – Exemplos de coeficientes de correlação

Ressalta-se, novamente, que um valor do coeficiente de correlação alto, embora estatisticamente significativo, não implica necessariamente numa relação de causa e efeito, mas, simplesmente indica a tendência que aquelas variáveis apresentam quanto à sua variação conjunta.

Outro cuidado que se deve tomar na análise de duas variáveis é com a ocorrência de correlações espúrias, ou seja, qualquer correlação aparente entre duas variáveis que não são correlacionadas de fato. As causas mais freqüentes da ocorrência dessas correlações são: a distribuição não equilibrada dos dados, como está apresentada na Figura 9.5; a relação entre quocientes de variáveis que apresentam o mesmo denominador, ilustrado na Figura 9.6, e a relação de variáveis que foram multiplicadas por uma delas, tal como mostrado na Figura 9.7.

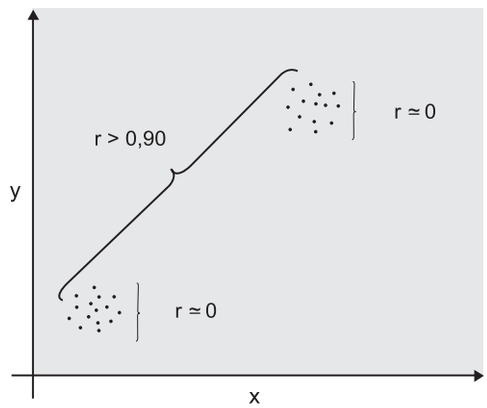

Figura 9.5 – Distribuição não equilibrada dos dados

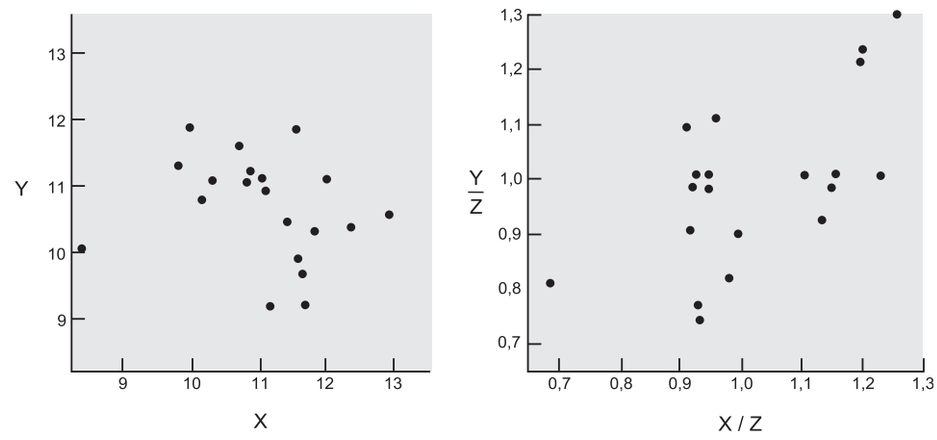

Figura 9.6 – Correlação entre quocientes de variáveis

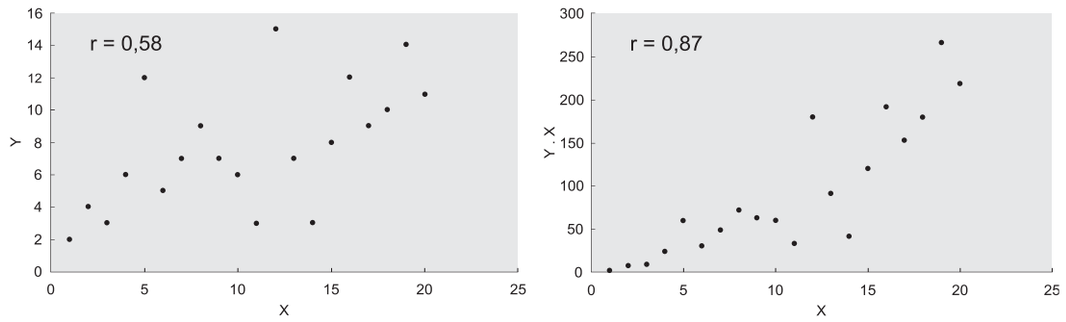

Figura 9.7 – Correlação entre produto de variáveis

9.1.1 – Testes de Hipóteses sobre o Coeficiente de Correlação

É possível testar a hipótese de que o coeficiente de correlação linear é igual a zero, ou seja:

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

Como decorrência de algumas hipóteses distributivas, a estatística apropriada para esse teste é a seguinte:

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad (9.6)$$

onde, t_0 é a estatística do teste; n é o tamanho da amostra e r é a estimativa do coeficiente de correlação linear.

A estatística do teste, t_0 , segue uma distribuição t de Student com $(n-2)$ graus de liberdade, sob a plausibilidade da hipótese nula H_0 : $\rho = 0$. A hipótese nula é rejeitada se:

$$|t_0| > t_{\alpha/2,n-2} \tag{9.7}$$

onde, $t_{\alpha/2,n-2}$ é o valor crítico para a estatística do teste bilateral para um nível de significância α, com $(n-2)$ graus de liberdade.

Testar hipóteses para o coeficiente de correlação, ρ_0 , diferente de zero, conforme apresentado a seguir, é um pouco mais complicado.

$$H_0 : \rho = \rho_0$$
$$H_1 : \rho \neq \rho_0$$

Segundo Montgomery e Peck (1992), para amostras de tamanho razoável ($n \geq 25$), a estatística:

$$Z = arctan\, h(r) = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right) \tag{9.8}$$

é aproximadamente normalmente distribuída com média

$$\mu_Z = arctan\, h(\rho) = \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right) \tag{9.9}$$

e variância

$$\sigma_Z^2 = (n-3)^{-1} \tag{9.10}$$

Para testar a hipótese nula, $\rho = \rho_0$, pode ser calculada a estatística

$$Z_0 = [arctan\, h(r) - arctan\, h(\rho_0)](n-3)^{1/2} \tag{9.11}$$

A hipótese nula será rejeitada se:

$$|Z_0| > Z_{\alpha/2} \tag{9.12}$$

onde, $Z_{\alpha/2}$ é o valor crítico para a estatística do teste bilateral, a qual é dada pela

variável central reduzida da distribuição normal padrão associada a um nível de significância α.

Segundo os mesmos autores, também é possível construir um intervalo de confiança, 100(1-α), para ρ utilizando a transformação obtida pela equação (9.8). Nesse caso, o intervalo de confiança é dado por

$$tanh\left[arctan\,h(r) - \frac{Z_{\alpha/2}}{\sqrt{n-3}} \right] \leq \rho \leq tanh\left[arctan\,h(r) + \frac{Z_{\alpha/2}}{\sqrt{n-3}} \right] \qquad (9.13)$$

onde *r* é o coeficiente de correlação estimado, $Z_{\alpha/2}$ é o quantil da distribuição normal padronizada com um nível de significância α, *n* é tamanho da amostra e

$$tanh(u) = \frac{\left(e^{u} - e^{-u}\right)}{\left(e^{u} + e^{-u}\right)} \qquad (9.14)$$

9.2 – Regressão Linear Simples

Muitas vezes, a simples visualização do diagrama de dispersão sugere a existência de uma relação funcional entre as duas variáveis. Essa observação introduz o problema de se determinar uma função que exprima esse relacionamento. A análise de regressão é uma técnica estatística cujo escopo é investigar e modelar a relação entre variáveis.

Considerando que exista um relacionamento funcional entre os valores *Y* e *X*, responsável pelo aspecto do diagrama, essa função deverá explicar parcela significativa da variação de *Y* com *X*. Contudo, uma parcela da variação permanece inexplicada e deve ser atribuída ao acaso. Colocando em outros termos, admite-se a existência de uma função que explica, em termos médios, a variação de uma das variáveis com a variação da outra. Freqüentemente, os pontos observados apresentarão uma variação em torno da linha da função de regressão, devido à existência de uma variação aleatória adicional denominada de *variação residual*. Portanto, essa equação de regressão fornece o valor médio de uma das variáveis em função da outra. Obviamente, caso se suponha conhecida a forma do modelo de regressão, a análise será facilitada. O problema, então, estará restrito à estimação dos parâmetros do modelo de regressão. Esse caso ocorrerá se existirem razões teóricas que permitam saber previamente que modelo rege a associação entre as variáveis. Geralmente, a forma da linha de regressão fica aparente na própria análise do diagrama de dispersão.

Admitindo ser uma reta a linha teórica de regressão, a função entre X e Y é a seguinte:

$$Y = \alpha + \beta X + e \tag{9.15}$$

onde, Y é a variável dependente, X é a variável independente, α *e* β são os coeficientes do modelo e e denota os erros ou resíduos da regressão.

Os coeficientes α *e* β da reta teórica são estimados através dos dados observados fornecidos pela amostra, obtendo uma reta estimativa na forma

$$\hat{y}_i = a + bx_i \tag{9.16}$$

onde a é a estimativa do coeficiente α ($\hat{\alpha} = a$); b é a estimativa de β ($\hat{\beta} = b$); $\hat{y}_i$ é o valor estimado da variável dependente e x_i é o valor observado da variável independente.

Existem vários métodos para a obtenção da reta desejada. O mais simples de todos, que podemos chamar de "método do ajuste visual", consiste simplesmente em traçar diretamente a reta, com auxílio de uma régua, no diagrama de dispersão, procurando fazer, da melhor forma possível, com que essa reta passe por entre os pontos. Entretanto, esse procedimento subjetivo, somente será razoável se a correlação linear for muito forte.

Um dos procedimentos objetivos mais adequados é a aplicação do método dos *mínimos quadrados*, segundo o qual a reta a ser adotada deverá ser aquela que torna mínima a soma dos quadrados dos erros ou resíduos da regressão.

9.2.1 – Método dos Mínimos Quadrados

O objetivo do método dos mínimos quadrados é encontrar a função de regressão que minimize a soma das distâncias entre a função ajustada e os pontos observados como apresentado na Figura 9.8. Adotando um modelo linear como da equação 9.15, os coeficientes α *e* β da reta teórica podem ser estimados através dos pontos experimentais fornecidos pela amostra, obtendo uma reta estimativa na forma da equação 9.16.

A distância, e_i, entre o valor observado e o valor estimado pela reta de regressão é dado por:

$$e_i = y_i - \hat{y}_i \tag{9.17}$$

onde y_i é o valor observado da variável dependente e $\hat{y}_i$ é o valor estimado da variável dependente.

Substituindo na equação 9.17 o valor estimado pela equação 9.16, $\hat{y}_i$, obtém-se:

$$e_i = y_i - a - bx_i \tag{9.18}$$

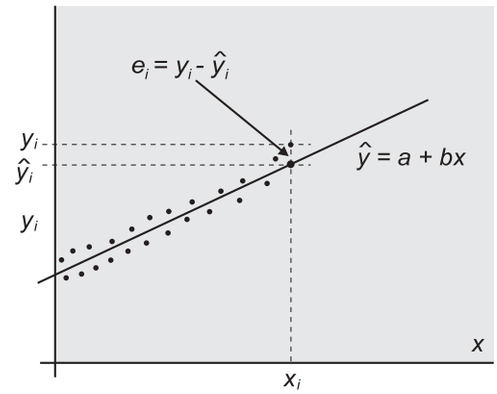

Figura 9.8 – Linha de Regressão

O método dos mínimos quadrados consiste em minimizar o somatório dos quadrados dos desvios entre o valor observado y_i e o valor estimado $\hat{y}_i$. Para o ponto indexado por *i*, o desvio quadrático é dado por

$$e_i^2 = \left(y_i - a - bx_i\right)^2 = y_i^2 - 2y_i a - 2y_i bx_i + a^2 + 2abx_i + b^2 x_i^2 \tag{9.19}$$

Para todos os *n* elementos da amostra,

$$Z = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} y_i^2 - 2a\sum_{i=1}^{n} y_i - 2b\sum_{i=1}^{n} x_i y_i + na^2 + 2ab\sum_{i=1}^{n} x_i + b^2 \sum_{i=1}^{n} x_i^2 \tag{9.20}$$

Como $Z = f\left(a,b\right)$, os valores de *a* e *b* que minimizam a equação acima são aqueles obtidos calculando-se as derivadas parciais, em relação a *a* e *b*, e igualando-as a zero,

$$\text{Mínimo de } Z \begin{cases} \dfrac{\partial Z}{\partial a} = 0 \\ \dfrac{\partial Z}{\partial b} = 0 \end{cases} \tag{9.21}$$

Calculando as derivadas para 9.20, obtém-se o seguinte sistema de equações

$$\begin{cases} \dfrac{\partial Z}{\partial a} = -2\sum_{i=1}^{n} y_i + 2na + 2b\sum_{i=1}^{n} x_i = 0 \\ \dfrac{\partial Z}{\partial b} = -2\sum_{i=1}^{n} x_i y_i + 2a\sum_{i=1}^{n} x_i + 2b\sum_{i=1}^{n} x_i^2 = 0 \end{cases} \tag{9.22}$$

Multiplicando as equações do sistema acima por (-1/2) encontra-se as equações normais da regressão linear simples:

$$\begin{cases} \sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 = 0 \end{cases} \tag{9.23}$$

A resolução do sistema de equações normais permite a estimativa dos parâmetros do modelo de regressão linear simples a partir dos dados amostrais:

$$a = \frac{\sum_{i=1}^{n} y_i}{n} - b\frac{\sum_{i=1}^{n} x_i}{n} = \bar{y} - b\bar{x} \tag{9.24}$$

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{9.25}$$

9.3 – Coeficiente de Determinação

Após a estimativa dos coeficientes da reta de regressão, é necessário verificar se os dados amostrais são descritos pelo modelo da equação 9.16 e, além disso, determinar a parcela da variabilidade amostral que foi, de fato, explicada pela reta de regressão. Essas questões podem ser analisadas considerando a Figura 9.9, a qual possibilita a dedução da seguinte relação simples:

$$y_i = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) + \bar{y} \tag{9.26}$$

A partir dessa equação, é possível demonstrar que

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{9.27}$$

O primeiro membro da equação 9.27 pode ser interpretado como proporcional à variância total de *Y*, enquanto o segundo membro reflete a soma de termos

proporcionais às suas variâncias residual e explicada pelo modelo de regressão. Em outros termos,

$$SQT = SQ\,Res + SQ\,Reg \tag{9.28}$$

onde SQT é a soma quadrática total; $SQ\,Res$ é soma dos quadrados dos resíduos e $SQ\,Reg$ é a soma dos quadrados devidos à regressão.

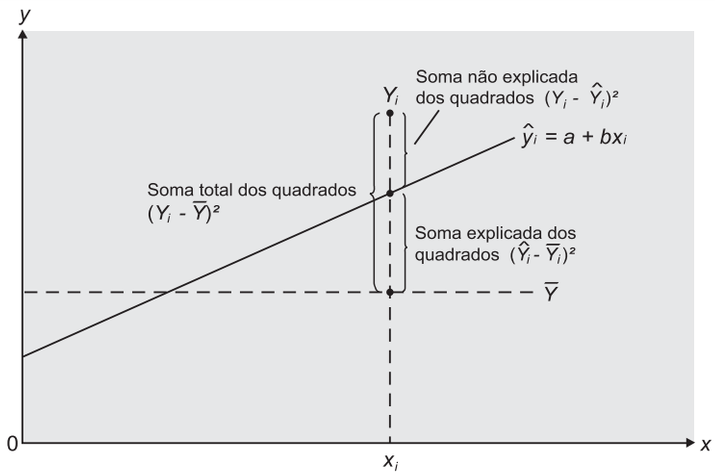

Figura 9.9 – Componentes de *Y*

O coeficiente de determinação é dado pela relação entre a soma dos quadrados devidos à regressão ($SQ\,Reg$) e a soma total dos quadrados (SQT), ou seja

$$r^2 = \frac{Variância\quad Explicada}{Variância\quad Total} = \frac{SQ\,Reg}{SQT} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{9.29}$$

onde r^2 é o coeficiente de determinação $\left(0 \le r^2 \le 1\right)$, y_i é o valor observado da variável dependente, $\hat{y}_i$ é o valor estimado da variável dependente e $\bar{y}$ é a média da variável dependente.

O coeficiente de determinação é sempre positivo e deve ser interpretado como a proporção da variância total da variável dependente *Y* que é explicada pelo modelo de regressão e que também pode ser estimado por:

$$r^2 = b^2 \frac{s_X^2}{s_Y^2} \tag{9.30}$$

onde s_X^2 é a variância amostral de X; s_Y^2 é a variância amostral de Y e b é o coeficiente angular da reta de regressão calculado pela equação 9.25.

O coeficiente de correlação amostral, r, está relacionado ao coeficiente de determinação, r^2, através da seguinte equação:

$$r = \pm\sqrt{r^2} \tag{9.31}$$

onde o sinal de r é o mesmo do de b.

9.4 – Hipóteses Básicas da Análise de Regressão Linear Simples (RLS)

As principais hipóteses da análise de regressão linear simples são a linearidade, a normalidade e a homoscedasticidade dos resíduos. A hipótese de linearidade define que a relação entre as variáveis analisadas deve ser linear, enquanto que o pressuposto de normalidade estabelece que os valores de Y são normalmente distribuídos para cada valor de X, conforme ilustrado na Figura 9.10.

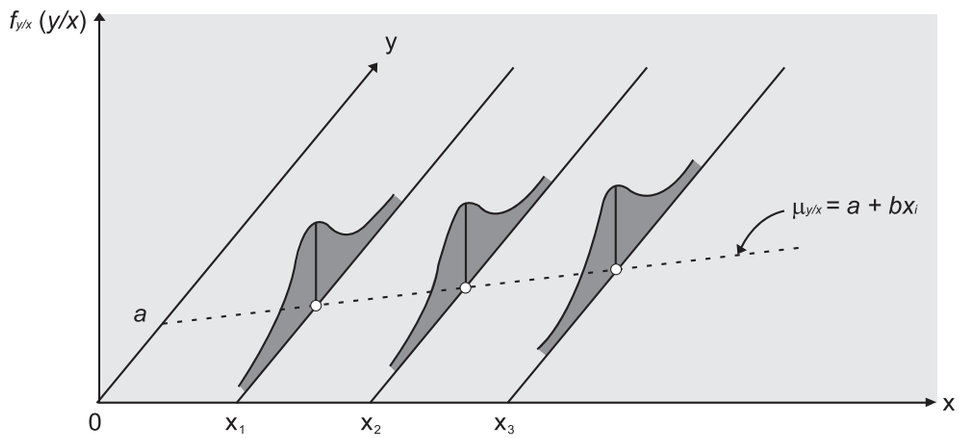

Figura 9.10 – Hipótese de normalidade

A hipótese de homoscedasticidade estabelece que os resíduos ou erros e_i, $e_i = y_i - (\alpha + \beta x_i)$, são realizações de uma variável aleatória independente e normalmente distribuída, com média zero e variância constante σ_e^2. A hipótese de homoscedasticidade dos resíduos implica nas seguintes afirmações:

a) O valor esperado da variável erro e_i é igual a zero, $E(e_i) = 0$
b) A correlação entre e_i e e_j com $i \neq j$ é é igual a zero

c) Como $Var(e_i) = Var(e_j)$, para $i \neq j$, a $Var(e_i)$ não varia com x_i, ou seja, a variância dos resíduos é constante.

9.4.1 – Erro Padrão da Estimativa

O modelo de regressão linear simples será perfeito se todos os pontos da amostra utilizados na estimativa dos parâmetros estiverem sobre a reta ajustada. Entretanto, a ocorrência de um modelo perfeito dificilmente será observada. A regressão linear simples possibilita uma estimativa aproximada de um valor de *Y* para um dado valor de *X*. Sendo assim, é importante uma medida da variabilidade dos pontos amostrais acima e abaixo da reta de regressão, tal como a dispersão esquematicamente ilustrada na Figura 9.8. Intrinsecamente ao processo de estimação dos parâmetros da reta de regressão, foi feita a premissa de que os erros são realizações de uma variável aleatória independente e normalmente distribuída com média zero, ou seja, $E(e_i) = 0$, e variância σ_e^2. Como $E(e_i) = 0$, a variância dos erros ou resíduos e_i será:

$$Var(e_i) = \sigma_e^2 = E(e_i^2) - E^2(e_i) = E(e_i^2) \tag{9.32}$$

Uma estimativa não enviesada da variância dos resíduos em torno da reta de regressão pode ser obtida por:

$$\hat{\sigma}_e^2 = s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} \tag{9.33}$$

A raiz quadrada da variância dos resíduos e_i é chamada de erro padrão da estimativa, σ_e, e mede a dispersão dos resíduos em torno da reta de regressão. O erro padrão da estimativa pode ser estimado por

$$\hat{\sigma}_e = s_e = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}} \tag{9.34}$$

9.5 – Teste de Hipóteses e Intervalos de Confiança para os Coeficientes da RLS

Devido à variabilidade amostral, a reta de regressão obtida da amostra extraída da população é uma das muitas retas possíveis. Os valores calculados para a e b

são estimativas pontuais dos parâmetros populacionais α e β. As retas da população e da amostra são paralelas quando $b = \beta$ e terão apenas um ponto necessariamente coincidente, a saber, a média da amostra x e a média da amostra y, quando $b \neq \beta$.

Os intervalos de confiança para os coeficientes α e β da reta de regressão são estimados por

$$a - t_{1-\frac{\alpha}{2},n-2} s_a \leq \alpha \leq a + t_{1-\frac{\alpha}{2},n-2} s_a \tag{9.35}$$

$$b - t_{1-\frac{\alpha}{2},n-2} s_b \leq \beta \leq b + t_{1-\frac{\alpha}{2},n-2} s_b \tag{9.36}$$

onde $t_{1-\frac{\alpha}{2},n-2}$ é valor do t de Student para (1- α/2) e $(n-2)$ graus de liberdade; a e b são os estimadores dos parâmetros da reta de regressão; s_a é o desvio-padrão da estimativa do parâmetro a e indica quão afastado o parâmetro estimado está do parâmetro populacional. A equação utilizada para o cálculo de s_a é dada por:

$$s_a = \sqrt{s_e^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)} \tag{9.37}$$

s_b é desvio-padrão da estimativa de b, calculado por:

$$s_b = \sqrt{\frac{s_e^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{9.38}$$

no cálculo de s_a e s_b tem-se:

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} \tag{9.39}$$

onde $e_i = y_i - \hat{y}_i$; n é o tamanho da amostra; $\bar{x}$ é a média da variável independente; e x_i é o valor observado da variável independente.

9.5.1 – Intervalos de Confiança para a Linha de Regressão Linear Simples

A reta obtida por mínimos quadrados é uma estimativa da função de regressão dada pela equação 9.15. De forma que, para um valor fixo x', o $\hat{y}'$ calculado pela relação $a + bx'$, corresponde a uma estimativa do valor que seria obtido pelo modelo de regressão linear, $y = \alpha + \beta x'$.

A construção de um intervalo de confiança para $\alpha + \beta x'$ pode se basear em sua estimativa, $\hat{y}'$. Considerando um valor x' que não foi utilizado no cálculo dos parâmetros da reta de regressão, demonstra-se que:

$$\mu(\hat{y}') = \alpha + \beta x' \tag{9.40}$$

$$\hat{\sigma}^2(\hat{y}') = \hat{\sigma}_e^2 \left[\frac{1}{n} + \frac{(x' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \tag{9.41}$$

O intervalo de confiança para a reta de regressão é dado por:

$$\hat{y}' \pm t_{1-\frac{\alpha}{2},n-2} s_e \sqrt{\frac{1}{n} + \frac{(x' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{9.42}$$

onde $\hat{y}' = a + bx'$, $t_{1-\frac{\alpha}{2},n-2}$ é valor do t de Student, para (1-α/2) e $(n-2)$ graus de liberdade; e s_e é calculado pela equação 9.34.

Analisando a equação 9.42, observa-se que a amplitude do intervalo será mínima quando x' for igual ao valor médio da amostra utilizada na definição da equação de regressão. Além disso, percebe-se que quanto mais distante x' estiver da média mais amplo será o intervalo. O limite inferior e superior do intervalo de confiança define a região de confiança em torno da reta de regressão, ou seja, tem-se um nível de confiança, 1 - α, de que a reta teórica, $y = \alpha + \beta x$, estará contida dentro dessa região. A Figura 9.11 ilustra a região de confiança em torno da reta de regressão.

9.5.2 – Intervalos de Confiança para um Valor Previsto pela RLS

Também é interessante estimar um intervalo com nível de confiança 1 - α, no qual estará contido um valor previsto de *y*, calculado para um certo valor especificado de *x'*. Os intervalos de confiança para um valor da variável dependente a ser previsto, $\hat{y}'$,utilizando um valor x' , são estimados por:

$$\hat{y}' - t_{1-\frac{\alpha}{2},n-2}\, s_e \sqrt{1+\frac{1}{n}+\frac{(x'-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \le \hat{y}' \le \hat{y}' + t_{1-\frac{\alpha}{2},n-2}\, s_e \sqrt{1+\frac{1}{n}+\frac{(x'-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \quad (9.43)$$

onde $\hat{y}' = a + bx'$, $t_{1-\frac{\alpha}{2},n-2}$ é valor do *t* de Student para (1- α/2) e $(n-2)$ graus; e s_e é calculado pela equação 9.34.

Variando *x'* na equação 9.43 obtêm-se a região de previsão para *y'*. Comparando as equações 9.42 e 9.43 verifica-se que o intervalo de confiança para um valor previsto é mais amplo que o estimado para a reta de regressão, como pode ser visualizado na Figura 9.11.

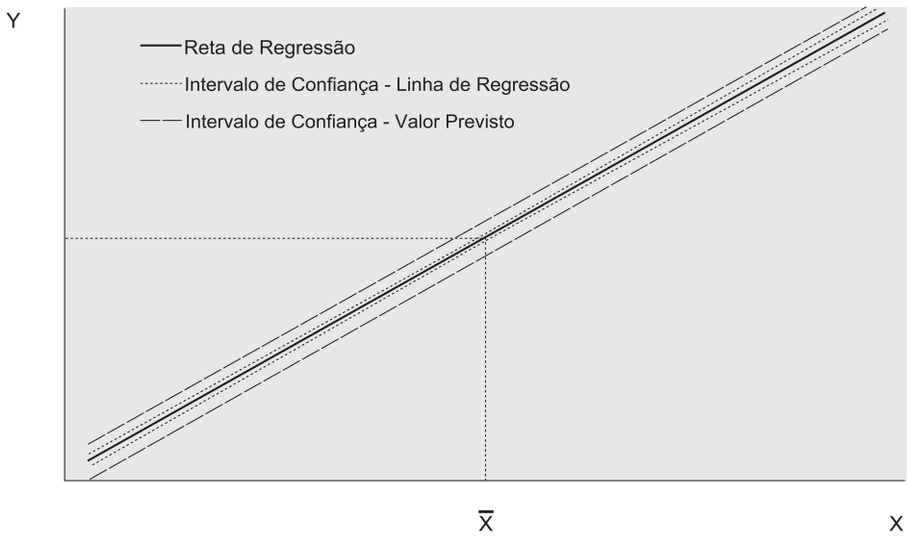

Figura 9.11 – Intervalos e Confiança

9.6 – Avaliação da Regressão Linear Simples

A análise de regressão é uma das técnicas mais úteis na hidrologia, mas exige certo cuidado na sua aplicação. Inicialmente devem ser verificadas as hipóteses da regressão, ou seja, avaliar a linearidade entre as variáveis X e Y, a independência dos resíduos e se estes seguem uma distribuição normal com média zero e variância constante σ_e^2.

A linearidade pode ser avaliada por meio do gráfico de dispersão entre as variáveis X e Y e pelo exame do valor da estimativa do coeficiente de correlação de Pearson. A existência de relação linear entre as variáveis X e Y também pode ser avaliada a partir de um teste de hipótese sobre o coeficiente angular β da equação 9.15. As hipóteses nula e alternativa podem ser expressas da seguinte forma:

$H_0 : \beta = 0$ (não existe relação linear)
$H_0 : \beta \neq 0$ (existe relação linear)

A estatística do teste, t, é igual a diferença entre a inclinação estimada a partir dos dados amostrais, b, e a inclinação da população, β, dividida pelo erro padrão da inclinação, s_b, calculado pela equação 9.38, ou seja,

$$t = \frac{b - \beta}{s_b} \tag{9.44}$$

No caso da plausibilidade da hipótese nula, $H_0 : \beta = 0$, obtém-se

$$t = \frac{b}{s_b} \tag{9.45}$$

A hipótese nula, H_0, é rejeitada se $|t| > t_{1-\alpha/2,n-2}$, onde $t_{1-\frac{\alpha}{2},n-2}$ é valor do t de Student para um nível de significância α (teste bilateral) e $(n-2)$ graus de liberdade.

Outra maneira de se avaliar a existência de uma relação linear entre as variáveis é realizada a partir do intervalo de confiança do parâmetro β, cuja estimativa foi detalhada no item 9.5. O teste consiste em verificar se o valor zero está contido dentro do intervalo de confiança de β. Se o valor zero estiver contido dentro do intervalo de confiança, não existe relação linear entre as variáveis.

A independência dos resíduos pode ser verificada com gráficos dos resíduos em relação à variável prevista, Y. A Figura 9.12 ilustra duas situações: uma onde se

verifica a independência dos resíduos e a outra na qual se observa a ocorrência de dependência.

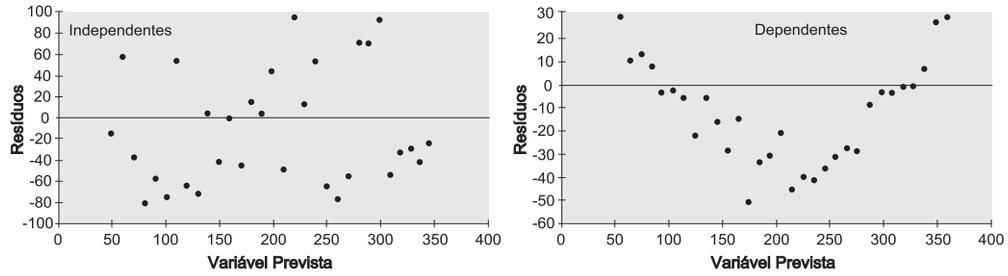

Figura 9.12 – Verificação da independência

Os métodos de análise de freqüência, descritos no capitulo 8, assim como a elaboração de gráficos de probabilidade Normal dos resíduos possibilitam a verificação da hipótese de normalidade. Contudo, para amostras pequenas, as definições sobre a normalidade dos resíduos geralmente não são conclusivas.

No caso da homoscedasticidade, a hipótese de média nula para os resíduos é garantida por construção. Entretanto, a hipótese de variância constante, σ_e^2, deve ser verificada por meio de análise gráfica entre os resíduos e a variável dependente X. A Figura 9.13 apresenta situações de verificação e violação de variância constante.

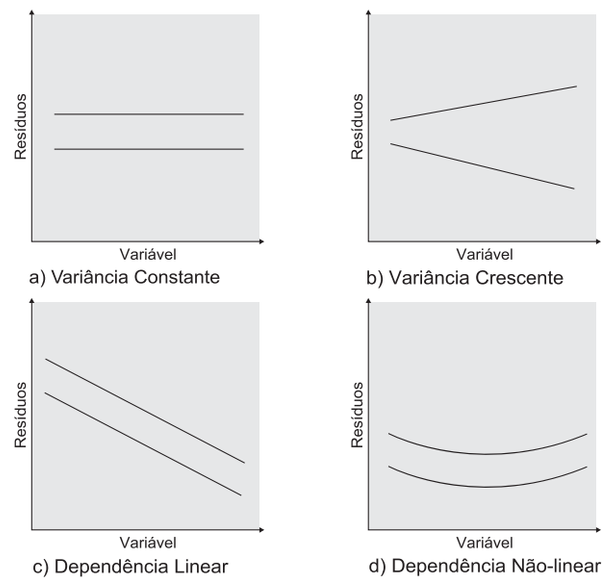

Figura 9.13 – Verificação da variância dos resíduos

Uma medida da qualidade da regressão pode ser obtida pela comparação do erro padrão da estimativa, s_e, com o desvio padrão da variável dependente Y, s_Y. Ambos, s_Y e s_e, apresentam as mesmas unidades e são, portanto, diretamente comparáveis, embora s_e tenha apenas n - 2 graus de liberdade e s_Y tenha n - 1. Caso a equação de regressão se ajuste bem aos dados amostrais, o erro padrão da estimativa se aproxima de zero. Entretanto, se o erro padrão da estimativa tiver valor próximo do desvio padrão de Y, o ajuste entre os dados amostrais e a equação de regressão será muito ruim. Assim, o erro padrão da estimativa deve ser comparado em seus extremos, a saber, zero e s_Y. Além disso, deve ser avaliado o coeficiente de determinação r^2, que expressa a proporção da variância total da variável dependente Y que é explicada pela equação de regressão.

Outro aspecto importante no uso de modelos de regressão é a sua extrapolação. De uma forma geral, não é recomendada a extrapolação da equação de regressão para além dos limites dos dados amostrais utilizados na estimativa dos parâmetros do modelo de regressão linear. O desestímulo à extrapolação apresenta basicamente dois motivos. O primeiro está associado ao fato do intervalo de confiança sobre a linha de regressão alargar, à medida que os valores da variável independente X se afastam da média, como pode ser visto na Figura 9.11. A outra razão é que a relação entre as variáveis X e Y pode não ser linear para valores que extrapolam os dados utilizados na regressão, como ilustrado na Figura 9.14.

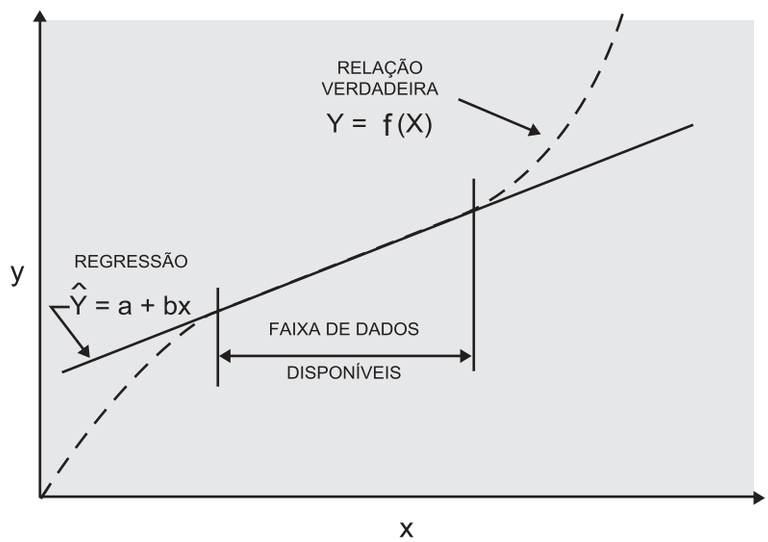

Figura 9.14 – Extrapolação do modelo de regressão

9.7 – Regressão Não-Linear com Funções Linearizáveis

Algumas funções podem ser linearizadas mediante o uso de transformações adequadas permitindo a aplicação da regressão linear simples. Um exemplo pode ser a função potencial a seguir:

$$y = ax^b \tag{9.46}$$

Realizando a anamorfose logarítmica dessa função, obtém-se:

$$ln\, y = ln\left(ax^b\right) \tag{9.47}$$

$$ln\, y = ln\, a + ln\left(x^b\right) \tag{9.48}$$

$$ln\, y = ln\, a + b\, ln\, x \tag{9.49}$$

Alterando as variáveis de forma que $z = ln\, y$,$k = ln\, a$ e $v = ln\, x$, a equação 9.49 se transforma na equação da reta:

$$z = k + bv \tag{9.50}$$

Trabalhando com as variáveis transformadas $z = ln\, y$ e $v = ln\, x$, é possível estimar os parâmetros *k* e *b* com as equações 9.24 e 9.25, respectivamente. Calculando o antilogaritmo de *k* estima-se o parâmetro *a* da equação 9.46.

De forma análoga, a função $y = ab^x$ pode ser resolvida utilizando as variáveis *x* e a transformada $ln\, y$. Existem muitas outras funções linearizáveis, como por exemplo, $y = \left(a + b.x\right)^{-2}$, que estão listadas no Anexo 10. Porém, como o processo de linearização pode envolver a transformação da variável dependente *Y*, em alguns casos as hipóteses da regressão podem não ser atendidas, após a modificação, prejudicando a aplicação dos testes estatísticos descritos anteriormente.

Exemplo 9.1 – Na Tabela 9.1 estão apresentados os valores médios de vazões máximas anuais e as respectivas áreas de drenagem de 22 estações fluviométricas que compõem uma região homogênea de um estudo de regionalização de vazões máximas da bacia do alto São Francisco no qual foi aplicado o método *index-flood*, ou cheia-índice, a ser descrito no capítulo 10. Nesse estudo as médias das vazões máximas anuais foram utilizadas como fator de adimensionalização das séries. Estabelecer uma regressão entre as médias das vazões máximas anuais e as áreas de drenagem, de

forma a permitir a estimativa da cheia-índice (ou *index-flood*) em locais que não possuam estações fluviométricas.

Tabela 9.1 – Área de drenagem e médias das vazões máximas anuais

Est.	1	2	3	4	5	6	7	8	9	10	11
Área (Km²)	269,1	481,3	1195,8	1055,0	1801,7	1725,7	1930,5	2000,2	1558,0	2504,1	5426,3
Q (m³/s)	31,2	49,7	100,2	109,7	154,3	172,8	199,1	202,2	207,2	263,8	483,8
ln A	5,59508	6,17649	7,08657	6,96130	7,49649	7,45339	7,56553	7,60100	7,35116	7,82568	8,59901
ln Q	3,44074	3,90560	4,60707	4,69784	5,03857	5,15190	5,29376	5,30906	5,33364	5,57500	6,18161
Est.	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
Área (Km²)	7378,3	9939,4	8734,0	8085,6	8986,9	11302,2	10711,6	13881,8	14180,1	16721,9	26553,0
Q (m³/s)	539,4	671,4	690,1	694,0	742,8	753,5	823,3	889,4	1032,4	1336,9	1964,8
ln A	8,90630	9,20426	9,07498	8,99784	9,10352	9,33275	9,27908	9,53833	9,55959	9,72447	10,18690
ln Q	6,29038	6,50941	6,53685	6,54241	6,61043	6,62469	6,71336	6,79050	6,93964	7,19810	7,58312

Solução: Inicialmente é elaborado um diagrama de dispersão, conforme está apresentado na Figura 9.15.

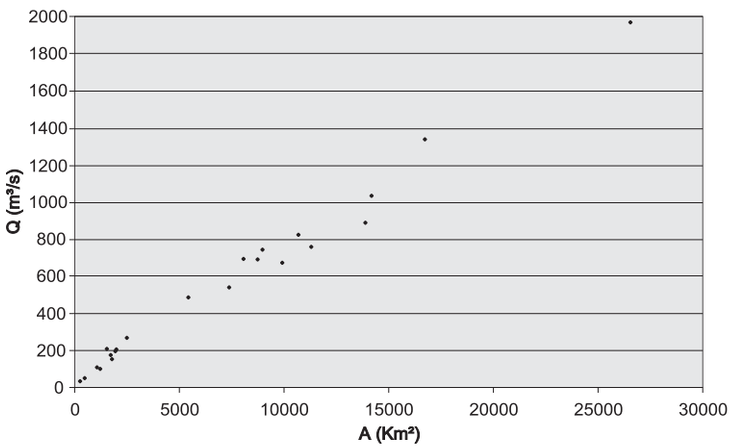

Figura 9.15 – Diagrama de dispersão

Analisando esse diagrama, percebe-se que a relação entre as variáveis área de drenagem e média da vazão máxima anual pode ser expressa por uma função potencial como a equação 9.46, ou seja,

$$Q = kA^b \tag{9.51}$$

Os parâmetros *k* e *b* podem ser estimados por meio da regressão linear simples, após a linearização da equação 9.51. A linearização é realizada

por anamorfose logarítmica como apresentado a seguir:

$$ln\,Q = ln\,k + b\,ln\,A \qquad (9.52)$$

Assim, para concretização da regressão linear simples é necessário calcular os logaritmos da área de drenagem e das médias das vazões máximas anuais, como apresentado na Tabela 9.1. A linearidade entre as variáveis, em coordenadas logarítmicas, pode ser visualizada na Figura 9.16.

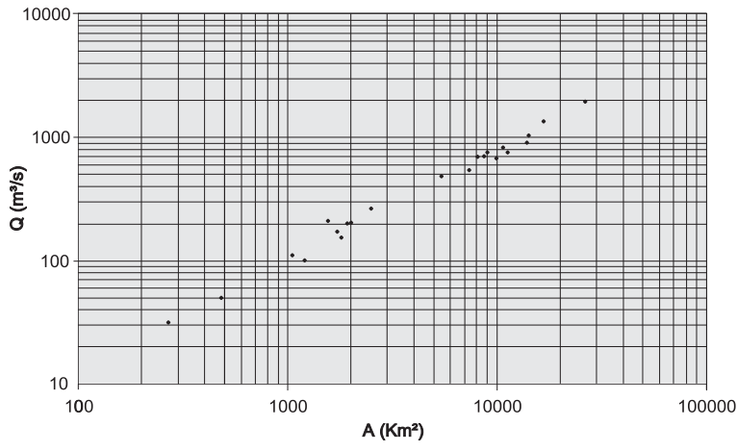

Figura 9.16 – Linearidade entre as variáveis

Utilizando as equações 9.24 e 9.25 e os logaritmos da Tabela 9.1, calcula-se os parâmetros da equação 9.52, b = 0,8751 e $a = ln(k)$ = - 1,4062 . A equação 9.52 é reescrita da seguinte forma:

$$ln\,Q = -1{,}4062 + 0{,}8751.ln\,A \qquad (9.53)$$

A equação 9.53 permite a estimativa de $ln\,Q$ em função do logaritmo da área de drenagem. O ajuste entre os logaritmos das médias das vazões máximas anuais e a reta de regressão da equação 9.53 está apresentado na Figura 9.17

As diferenças ou os resíduos entre os valores observados e os calculados pela reta de regressão estão na Tabela 9.2.

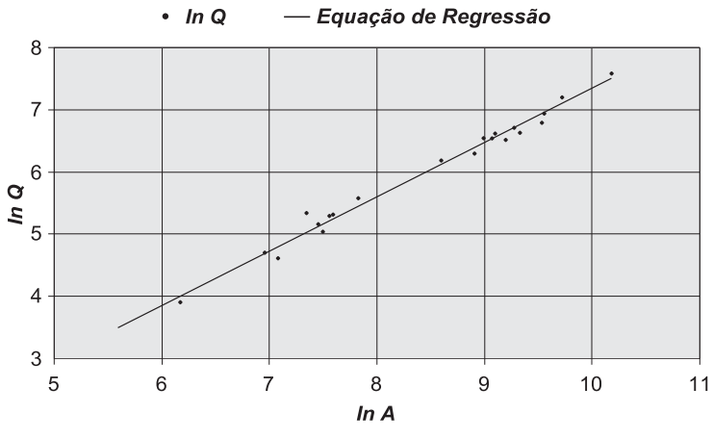

Figura 9.17 – Ajuste entre as observações e a reta de regressão

Tabela 9.2 – Resíduos

Est.	1	2	3	4	5	6	7	8	9	10	11
ln Q	3,4407	3,9056	4,6071	4,6978	5,0386	5,1519	5,2938	5,3091	5,3336	5,5750	6,1816
Previsto	3,4900	3,9988	4,7952	4,6856	5,1540	5,1162	5,2144	5,2454	5,0268	5,4420	6,1188
Res.	-0,0493	-0,0932	-0,1882	0,0122	-0,1154	0,0357	0,0794	0,0636	0,3069	0,1330	0,0628
Est.	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
ln Q	6,2904	6,5094	6,5369	6,5424	6,6104	6,6247	6,7134	6,7905	6,9396	7,1981	7,5831
Previsto	6,3877	6,6484	6,5353	6,4678	6,5603	6,7609	6,7139	6,9408	6,9594	7,1037	7,5083
Res.	-0,0973	-0,1390	0,0016	0,0746	0,0502	-0,1362	-0,0005	-0,1503	-0,0197	0,0944	0,0748

Os valores observados e os calculados com a equação de regressão permitem a estimativa dos termos da equação 9.27, ou seja, os somatórios dos quadrados total, dos resíduos e os devidos à regressão. Os valores desses somatórios estão apresentados na Tabela 9.3.

Tabela 9.3 – Somatórios dos Quadrados

	Graus de Liberdade	Somatórios dos Quadrados
Regressão	1	24,7726
Resíduo	20	0,2803
Total	21	25,0529

O coeficiente de determinação r^2 é calculado através da equação 9.29.

$$r^2 = \frac{SQ\,Reg}{SQT} = \frac{24{,}7726}{25{,}0529} = 0{,}989 \qquad (9.54)$$

O coeficiente de correlação, r, é igual a 0,994.

Após o cálculo dos parâmetros e dos resíduos é possível verificar as hipóteses da regressão. A seguir é verificada a hipótese de homoscedasticidade dos resíduos. Avaliando a Figura 9.18 observa-se que os resíduos parecem ser independentes e que a variância pode ser considerada aproximadamente constante.

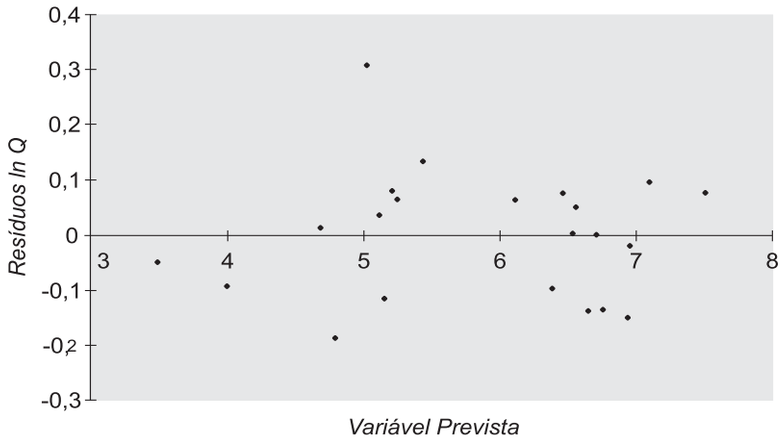

Figura 9.18 – Resíduos

Como o somatório dos resíduos é igual a zero, a sua média também é igual a zero. A raiz quadrada da variância dos resíduos ou o erro padrão da estimativa é calculado pela equação 9.34.

$$\hat{\sigma}_e = s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SQ\,Res}{n-2}} = \sqrt{\frac{0{,}2803}{20}} = 0{,}1184 \qquad (9.55)$$

A Figura 9.19 apresenta o ajuste entre os resíduos e uma distribuição normal de média zero e desvio padrão igual a 0,1184.

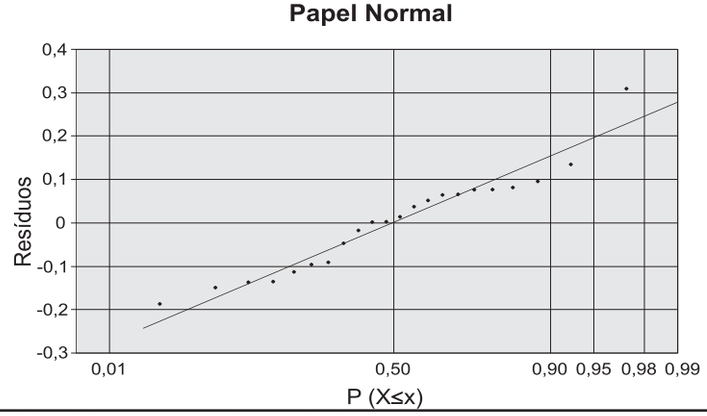

Figura 9.19 – Ajuste dos resíduos à distribuição normal

Os intervalos de confiança para os coeficientes α e β da reta de regressão são estimados com as equações 9.35 e 9.36. Adotando um nível de significância de 5% obtém-se:

$$-1{,}77045 \le \alpha \le -0{,}04196 \text{ e } 0{,}83168 \le \beta \le 0{,}91851$$

No calculo dos limites desses intervalos foram utilizadas os seguintes valores:

$$t_{1-\frac{\alpha}{2},n-2} = t_{0,975,21} = 2{,}086$$

$$s_a = \sqrt{s_e^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)} = 0{,}1746 \text{ e } s_b = \sqrt{\frac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 0{,}0208$$

A relação linear entre as variáveis *ln Q* e *ln A* também pode ser avaliada através de um teste de hipótese com o coeficiente angular da reta de regressão, como descrito no item 9.5. Neste exemplo, a estatística do teste é dada por:

$$t = \frac{b - \beta}{s_b} = \frac{0{,}8751 - 0}{0{,}0208} = 42{,}072 \tag{9.56}$$

Como $|t| > t_{1-\alpha/2,n-2}$, pois $t_{0,975,21} = 2{,}086$, a hipótese nula, $\beta = 0$, é rejeitada a um nível de significância de 5%, ou seja, a relação entre as variáveis pode ser considerada linear com uma confiança de 95%.

As etapas anteriores descreveram a regressão linear simples das variáveis transformadas, entretanto, para estimativa do fator "index-flood" utiliza-se a equação na forma potencial como descrito acima. Assim, o parâmetro *k* da equação 9.51 é definido da seguinte forma:

$$k = exp(a) = exp(-1{,}4062) = 0{,}2451 \tag{9.57}$$

A equação 9.51 é reescrita como:

$$Q = kA^b = 0{,}2451A^{0{,}8751} \tag{9.58}$$

Finalmente é realizada uma comparação entre os valores observados e os estimados com a equação 9.58 como está apresentado na Tabela 9.4 e Figura 9.20.

Tabela 9.4 – Desvios Percentuais (DP)

n	1	2	3	4	5	6	7	8	9	10	11
Qobs (m³/s)	31,2	49,7	100,2	109,7	154,3	172,8	199,1	202,2	207,2	263,8	483,8
Qcalc (m³/s)	32,8	54,5	120,9	108,4	173,1	166,7	183,9	189,7	152,4	230,9	454,3
DP (%)	5,1	9,8	20,7	-1,2	12,2	-3,5	-7,6	-6,2	-26,4	-12,5	-6,1
n	12	13	14	15	16	17	18	19	20	21	22
Qobs (m³/s)	539,4	671,4	690,1	694,0	742,8	753,5	823,3	889,4	1032,4	1336,9	1964,8
Qcalc (m³/s)	594,5	771,6	689,0	644,1	706,5	863,4	823,8	1033,6	1053,0	1216,4	1823,2
DP (%)	10,2	14,9	-0,2	-7,2	-4,9	14,6	0,1	16,2	2,0	-9,0	-7,2

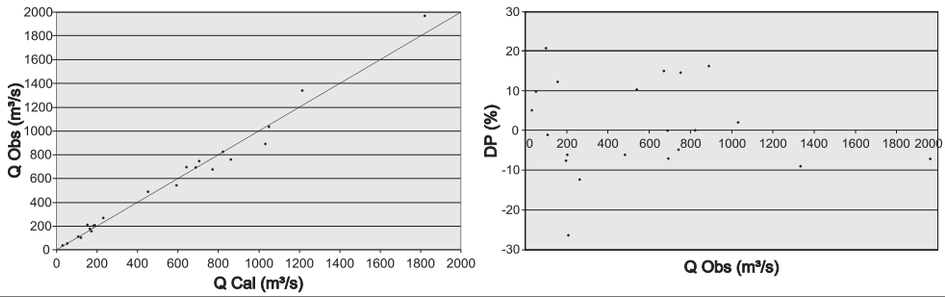

Figura 9.20 – Vazões calculadas versus observadas e desvio percentual

9.8 – Regressão Linear Múltipla

Na regressão múltipla estuda-se o comportamento de uma variável dependente Y em função de duas ou mais variáveis independentes X_i. Se a variável Y variar linearmente com as variáveis X_i, pode-se adotar um modelo geral com a seguinte forma:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P \tag{9.59}$$

onde Y é a variável dependente ou prevista; $X_1, X_2, \cdots, X_P$ são as variáveis independentes ou explicativas e $\beta_1, \beta_2, \cdots, \beta_P$ são os coeficientes de regressão.

A partir de um conjunto de *n* valores da variável *Y*, associados às *n* observações correspondentes das *P* variáveis independentes, e utilizando a equação 9.59, pode-se escrever

$$\begin{cases} Y_1 = \beta_1 X_{1,1} + \beta_2 X_{1,2} + \cdots + \beta_P X_{1,P} \\ Y_2 = \beta_1 X_{2,1} + \beta_2 X_{2,2} + \cdots + \beta_P X_{2,P} \\ \vdots \\ Y_n = \beta_1 X_{n,1} + \beta_2 X_{n,2} + \cdots + \beta_P X_{n,P} \end{cases} \tag{9.60}$$

no qual Y_i é o i-ésimo valor da variável dependente e $X_{i,j}$ é a i-ésima observação da j-ésima variável independente. O sistema de equações 9.60 pode ser representado na forma de matriz:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,P} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,P} \\ \vdots & \vdots & \cdots & \vdots \\ X_{n,1} & X_{n,2} & \cdots & X_{n,P} \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \tag{9.61}$$

ou em notação matricial,

$$[Y] = [X] \cdot [\beta] \tag{9.62}$$

onde $[Y]$ é um vetor (n x 1) das observações da variável dependente; $[X]$ é uma matriz (n x P) com as n observações de cada uma das P variáveis independentes, e $[\beta]$ é um vetor (P x 1) com os parâmetros desconhecidos. A equação 9.62 terá um termo de intercepto, β_1, se $X_{i,1} = 1$; doravante, no presente texto, adota-se a condição de $X_{i,1} = 1$ para i de 1 até n.

De maneira análoga à regressão linear simples, os coeficientes desconhecidos β_i podem ser estimados pela minimização do somatório dos erros quadráticos, $\sum_{i=1}^{n} e_i^2$, onde,

$$e_i = Y_i - \hat{Y}_i = Y_i - \sum_{i=1}^{P} \hat{\beta}_j X_{i,j} \tag{9.63}$$

Em representação matricial,

$$\sum e_i^2 = [e]^T [e] = \left([Y] - [X\hat{\beta}]\right)^T \left([Y] - [X\hat{\beta}]\right) \tag{9.64}$$

Diferenciando a equação 9.64, em relação a $\hat{\beta}$, e igualando a derivada parcial a zero, obtém-se o sistema

$$[X]^T [Y] = [X]^T [X\hat{\beta}] \tag{9.65}$$

que representa as equações normais de regressão. As soluções da equação 9.65 são encontradas pela multiplicação dois termos da equação por $\left([X]^T[X]\right)^{-1}$.

Desse modo, o vetor $[\hat{\beta}]$ pode ser estimado por:

$$[\hat{\beta}]=([X]^T[X])^{-1}[X]^{-1}[Y] \tag{9.66}$$

De maneira semelhante à regressão simples, o somatório total dos quadrados pode ser apresentado em três parcelas:

$$\sum Y_i^2 = n\overline{Y}^2 + \sum\left(Y_i - \hat{Y}_i\right)^2 + \sum\left(\hat{Y} - \overline{Y}\right)^2 \tag{9.67}$$

ou, em notação matricial, como:

$$[Y]^T[Y] = n\overline{Y}^2 + \left([\hat{\beta}]^T[X]^T[Y] - n.\overline{Y}^2\right) + \left([Y]^T[Y] - [\hat{\beta}]^T[X]^T[Y]\right) \tag{9.68}$$

Freqüentemente, essas parcelas dos somatórios dos quadrados são apresentadas na forma de uma tabela de análise de variância (ANOVA), tal como a ilustrada na Tabela 9.5. O quadrado médio, na Tabela 9.5, resulta da divisão do somatório dos quadrados pelo respectivo número de graus de liberdade.

Tabela 9.5 – Tabela ANOVA da regressão múltipla

Fonte	Graus de liberdade	Somatório dos quadrados	Quadrado médio
Regressão	P	$SQ\,Reg = [\hat{\beta}]^T[X]^T[Y] - n\overline{Y}^2$	$QM\,Reg = \frac{SQ\,Reg}{P}$
Resíduos	$n - P$ -1	$SQ\,Res = [Y]^T[Y] - [\hat{\beta}]^T[X]^T[Y]$	$QM\,Res = \frac{SQ\,Res}{n-P-1}$
Total	n - 1	$SQT = [Y]^T[Y] - n\overline{Y}^2$	

O coeficiente de determinação múltipla R^2 é definido pela seguinte relação:

$$R^2 = \frac{SQ\,Reg}{SQT} = \frac{[\hat{\beta}]^T[X]^T[Y] - n\overline{Y}^2}{[Y]^T[Y] - n\overline{Y}^2} \tag{9.69}$$

O coeficiente de determinação múltipla varia entre 0 a 1 e expressa a proporção da variância que é explicada pelo modelo de regressão. O coeficiente de correlação múltipla é calculado pela extração da raiz quadrada da equação 9.69.

Uma estimativa não enviesada da variância dos erros, $Var(\varepsilon)$ ou σ_e^2, é dada por s_e^2 que é calculada pelo quadrado médio dos resíduos, conforme está apresentado a seguir.

$$s_e^2 = QM\ Res = \frac{SQ\ Res}{n-P-1} = \frac{[Y]^T[Y] - \left[\hat{\beta}\right]^T[X]^T[Y]}{n-P-1} \tag{9.70}$$

O erro padrão da equação de regressão linear múltipla, σ_e, é estimado por s_e, o qual é calculado pela raiz quadrada da equação 9.70.

9.8.1 – Teste da Significância da Equação de Regressão Linear Múltipla

A existência de uma relação significativa entre a variável dependente e as variáveis independentes ou explicativas, pode ser avaliada pelo seguinte teste de hipóteses:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_n = 0$ (a relação entre as variáveis não é linear)
H_1: pelo menos um $\beta_i \neq 0$

Esse teste é conhecido como 'teste do *F* total', o qual é utilizado para testar a razão entre duas variâncias e, assim, pode ser empregado para verificar a hipótese nula. A estatística do teste é a relação entre a variância decorrente da regressão linear múltipla e variância dos resíduos, ou seja,

$$F = \frac{QM\ Reg}{QM\ Res} \tag{9.71}$$

Os quadrados médios da regressão e dos resíduos ($QM\ Reg$ e $QM\ Res$) podem ser calculados pelas equações apresentadas na Tabela 9.5. A hipótese nula será aceita se

$$F < F(\alpha, P, n-p-1) \tag{9.72}$$

onde α é o nível de significância, P e $n-P-1$ são os graus de liberdade da distribuição *F* de Snedecor, sendo que *P* é o número de variáveis independentes.

9.8.2 – Teste de Partes de um Modelo de Regressão Linear Múltipla

A contribuição de uma variável explicativa ao modelo de regressão múltipla pode ser determinada pelo critério do chamado 'teste do *F* parcial'. De acordo com esse critério, avalia-se a contribuição de uma variável explicativa para a soma dos quadrados devido a regressão, depois que todas as outras variáveis independentes foram incluídas no modelo. Sendo assim, a contribuição de uma variável X_k para a soma dos quadrados da regressão, $SQ\ Reg(X_k)$, considerando que as outras

variáveis estão incluídas, é estimada pela diferença dada por

$SQ\,Reg(X_k)$ = $SQ\,Reg$ (todas as variáveis com X_k) - $SQ\,Reg$ (todas as variáveis sem X_k) (9.73)

A verificação se a inclusão de uma variável X_k melhora significativamente o modelo de regressão é realizada por meio de um teste com as seguintes hipóteses nula e alternativa:

H_0 : a variável X_k não melhora significativamente o modelo
H_1 : a variável X_k melhora significativamente o modelo

A estatística do teste é dada pela relação entre a contribuição da variável X_k à soma dos quadrados devido a regressão, $SQ\,Reg(X_k)$, calculada pela equação 9.73, e a variância dos resíduos considerando o modelo com todas as variáveis inclusive X_k, que é estimada pelo quadrado médio dos resíduos apresentado na Tabela 9.5. Formalmente,

$$F_P = \frac{SQ\,Reg(X_k)}{QM\,Res} \tag{9.74}$$

A hipótese nula deve ser rejeitada se a estatística F_P for maior que o valor critico da distribuição F de Snedecor, com 1 e $n-P-1$ graus de liberdade, e nível de significância α, onde n é o tamanho da amostra e P é o número de variáveis explicativas incluindo X_k, ou seja, rejeita-se H_0 se

$$F_P > F(\alpha,1,n-p-1) \tag{9.75}$$

9.8.3 – Coeficiente de Determinação Parcial

O coeficiente de determinação múltipla, R^2, avalia a proporção da variância da variável dependente Y que é explicada pelas variáveis independentes X_i. Todavia, também é importante avaliar a contribuição de cada variável explicativa em relação ao modelo de regressão múltipla. A proporção da variância da variável dependente Y que é explicada por uma variável independente X_k, enquanto se mantém constante as outras variáveis explicativas, é estimada pelo coeficiente de regressão parcial $R^2_{Yk(P-k)}$. Para um modelo de regressão múltipla com P variáveis explicativas, o coeficiente de determinação parcial para a k-ésima variável é dado por:

$$R^2_{Yk(P-k)} = \frac{SQ\,Reg(X_k)}{SQT - SQ\,Reg + SQ\,Reg(X_k)} \tag{9.76}$$

onde SQT é a soma dos quadrados total, $SQ\,Reg$ é a soma dos quadrados da regressão com todas as variáveis inclusive X_k, ambos calculados pelas fórmulas apresentadas na Tabela 9.5, e $SQ\,Reg(X_k)$ é a contribuição da variável X_k para a soma dos quadrados da regressão estimada pela equação 9.73.

9.8.4 – Inferências sobre os Coeficientes da Regressão Linear Múltipla

Nesse item também serão admitidas as hipóteses que os resíduos ou erros e_i são independentes e normalmente distribuídos com média zero e variância σ_e^2. A variância de $\hat{\beta}_i$ é estimada pela seguinte relação:

$$\hat{V}ar(\hat{\beta}_i) = \hat{\sigma}^2_{\hat{\beta}_i} = S^2_{\hat{\beta}_i} = C^{-1}_{ii}\hat{\sigma}^2_e \tag{9.77}$$

onde C^{-1}_{ii} é o i-ésimo elemento da diagonal de $\left[X^T X\right]^{-1}$ e $\hat{\sigma}^2_e$ á estimativa de variância dos erros e_i.

Se o modelo estiver correto, então $\hat{\beta}_i / S_{\hat{\beta}_i}$ é distribuído conforme *t* de Student, com $n - P - 1$ graus de liberdade, onde $s_{\hat{\beta}_i}$ é uma estimativa de $\sigma_{\hat{\beta}_i}$ calculada por:

$$s_{\hat{\beta}_i} = \sqrt{C^{-1}_{ii} s^2_e} \tag{9.78}$$

s^2_e é uma estimativa da variância dos resíduos e_i, tal como calculada pela equação 9.70.

Um teste de hipótese para verificar se $\beta_i = \beta_0$, onde β_0 é um valor constante conhecido, pode ser implementado com as seguintes hipóteses nula e alternativa:

$$H_0 : \beta_i = \beta_0$$
$$H_1 : \beta_i \neq \beta_0$$

Para tais hipóteses, a estatística do teste é calculada pela relação:

$$t = \frac{\hat{\beta}_i - \beta_0}{s_{\hat{\beta}_i}} \tag{9.79}$$

A hipótese nula H_0 deve ser rejeitada se

$$|t| > t_{1-\alpha/2,\,n-P-1} \tag{9.80}$$

onde α é o nível de significância (teste bilateral), *n* é tamanho da amostra e *P* é número de variáveis independentes do modelo.

Um teste para a hipótese nula, $H_0 : \beta_i = 0$, e hipótese alternativa, $H_1 : \beta_i \neq 0$, é equivalente a testar a significância da i-ésima variável independente na explicação da variância da variável dependente. A estatística do teste é calculada pela equação 9.79 considerando $\beta_0 = 0$ e a verificação da hipótese é realizada com a equação 9.80. Caso a hipótese nula seja aceita, $\beta_i = 0$, sendo recomendável que a i-ésima variável explicativa seja retirada do modelo.

Verifica-se facilmente que a estatística do teste *F* parcial, equação 9.74, e a estatística *t*, equação 9.79, apresentam a seguinte relação:

$$F_{1,gl} = t_{gl}^2 \tag{9.81}$$

onde *gl* é são os graus de liberdade.

Os intervalos de confiança para os coeficientes da regressão, β_i , são dados por:

$$\hat{\beta}_i \pm t_{1-\frac{\alpha}{2},n-P-1} s_{\hat{\beta}_i} \tag{9.82}$$

9.8.5 – Intervalos de Confiança da Regressão Linear Múltipla

Os limites de confiança de Y_h, onde $Y_h = [X_h][\hat{\beta}]$, são definidos a partir da variância de $\hat{Y}_h$. Neste caso, $\hat{Y}_h$ é uma estimativa de Y (um escalar), no ponto $[X_h]$ (um vetor 1 x *P*) no espaço *P* dimensional e $[\hat{\beta}]$ é um vetor contendo as estimativas de β . A variância de $\hat{Y}_h$ é calculada por:

$$Var(\hat{Y}_h) = \sigma_e^2 [X_h][X^T X]^{-1} [X_h]^T \tag{9.83}$$

onde σ_e^2 é a estimado por s_e^2 através da equação 9.70.

Os limites de confiança de $\hat{Y}_h$ são estabelecidos por:

$$[X_h][\hat{\beta}] \pm t_{1-\frac{\alpha}{2},n-P-1} \sqrt{Var(\hat{Y}_h)} \tag{9.84}$$

Os intervalos de confiança de um valor individual previsto $\hat{Y}_h$ são estimados pela equação a seguir:

$$[X_h][\hat{\beta}] \pm t_{1-\frac{\alpha}{2},n-P-1} \sqrt{Var_i(\hat{Y}_h)} \tag{9.85}$$

onde $Var_i(\hat{Y}_h)$ é a variância de um valor individual previsto de *Y* calculado com

$[X_h]$, sendo estimada por:

$$\hat{V}ar_i\left(\hat{Y}_h\right) = \hat{\sigma}_e^2\left(1 + [X_h][X^T X]^{-1}[X_h]^T\right) \tag{9.86}$$

9.8.6 – Transformações de um Modelo de Regressão Múltipla

Em alguns casos, a violação do pressuposto de homoscedasticidade dos resíduos pode ser superada, por meio da transformação da variável dependente, das variáveis explicativas ou de ambas. Além disso, a transformação de variáveis pode permitir a linearização de uma relação não linear. De uma forma geral, a modificação das variáveis para alcançar os critérios de homoscedasticidade não é uma tarefa fácil. As transformações mais utilizadas são a de raiz quadrada, a logarítmica e a recíproca, conforme apresentado a seguir:

$$Y = \beta_0 + \beta_1\sqrt{X_1} + \beta_2\sqrt{X_2} + \cdots + \varepsilon \tag{9.87}$$

$$Y = \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \cdots + \varepsilon \tag{9.88}$$

$$Y = \beta_0 + \beta_1 \frac{1}{X_1} + \beta_2 \frac{1}{X_1} + \cdots + \varepsilon \tag{9.89}$$

As transformações de modelos não lineares podem ser obtidas por meio de anamorfose logarítmica, tal como exemplificado a seguir.

Modelo multiplicativo do tipo

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \varepsilon \tag{9.90}$$

Após a transformação obtêm-se:

$$\ln Y = \ln \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \ln \varepsilon \tag{9.91}$$

No caso de um modelo exponencial

$$Y = e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}\varepsilon \tag{9.92}$$

A transformação logarítmica resulta em:

$$\ln Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ln \varepsilon \tag{9.93}$$

9.8.7 – Comentários Sobre a Regressão Múltipla

Em situações onde as variáveis explicativas são fortemente correlacionadas podem ocorrer problemas na regressão múltipla. Variáveis colineares não fornecem novas informações, dificultando a interpretação dos coeficientes obtidos na regressão, pois em alguns casos o sinal do coeficiente de regressão pode ser o oposto do esperado. Por isso é fortemente recomendável a montagem de uma matriz de coeficientes de correlação simples entre as variáveis explicativas para verificar a existência de uma possível colinearidade entre essas variáveis. Um modo expedito de evitar a colinearidade é a eliminação de uma, entre cada conjunto de duas variáveis explicativas que apresentarem coeficientes de correlação superiores a 0,85. Desse modo, espera-se que as variáveis mantidas no modelo de regressão contribuam significativamente para explicar a variabilidade de *Y*.

O número de observações disponíveis para a análise de regressão deve ser no mínimo 3 a 4 vezes maior que o número de coeficientes da equação regressão que serão estimados. Esta regra procura evitar um falso ajuste causado pelas oscilações que podem ocorrer nas variáveis independentes e que são de difícil detecção nas amostras muito pequenas.

Existem alguns procedimentos que facilitam a elaboração dos modelos de regressão múltipla, do ponto de vista da seleção de variáveis explicativas. Dentre os vários métodos podem ser destacado o de todas as equações possíveis e o da regressão passo a passo.

As diferentes combinações das variáveis independentes permitem a construção de vários modelos de regressão. Caso as equações de regressão tenham um intercepto, β_1, podem ser definidos 2^{P-1} modelos, onde *P* é o número de variáveis independentes. A definição pelo melhor modelo está associada à análise de cada um separadamente.

A regressão passo a passo consiste na incorporação ao modelo de uma variável, a cada vez, com o objetivo de explicar a maior parte da variância que ainda não foi explicada pelo modelo. Esse método inicia-se com a variável independente que apresenta o maior coeficiente de correlação simples com a variável dependente. Em seguida, é acrescentada uma variável independente à equação, a cada passo, com a avaliação da significância do modelo elaborado e de suas variáveis explicativas, por meio do teste do *F* parcial. Se a contribuição de uma das variáveis explicativas não for considerada significativa, ela é retirada do modelo.

A definição sobre qual a melhor equação de regressão a ser adotada envolve

certa subjetividade. Entretanto, a avaliação da equação de regressão pode ser realizada objetivamente a partir das considerações descritas a seguir. O erro padrão da estimativa deve ser inferior ao desvio padrão da variável independente, $0 \leq S_e \leq S_Y$, pelos mesmos motivos apontados para a regressão linear simples. O coeficiente de determinação deve se aproximar de 1, pois quanto maior o valor desse coeficiente, maior será a proporção da variância explicada pelo modelo. Os testes *F* total, *F* parcial e o teste *t* dos coeficientes da regressão devem ser aplicados para avaliar a significância de cada preditor e do modelo. O sinal do coeficiente de correlação entre uma variável explicativa (X_i) e a variável dependente (Y) deve ser o mesmo do coeficiente da regressão associado a essa variável independente. Os resíduos devem ser examinados através de gráficos com as variáveis independentes e dependentes, para identificar deficiências na equação de regressão e conferir as hipóteses da regressão. E finalmente, comparar os valores previstos com a equação de regressão e dados observados.

Uma maneira de se avaliar os resultados da equação de regressão é verificar a capacidade do modelo prever a variável dependente a partir de observações das variáveis explicativas que não foram utilizadas na estimativa dos coeficientes da regressão. Obviamente, para se fazer essa avaliação é necessário que os dados observados sejam separados aleatoriamente em dois grupos, um para estimar os coeficientes da regressão e o outro para verificar o modelo. Entretanto, na maioria dos casos, o número reduzido de observações não permite esse procedimento.

Exemplo 9.2 – Em um estudo de regionalização de vazões mínimas com 7 dias de duração na bacia do rio Paraopeba, no qual foi aplicado o método *index-flood*, definiu-se uma região homogênea com 15 estações fluviométricas. Nesse estudo as médias das vazões mínimas anuais com 7 dias de duração foram utilizadas como fator de adimensionalização das séries. Defina um modelo de regressão que permita a estimativa do fator *index-flood* em locais que não possuam estações fluviométricas utilizando como prováveis variáveis explicativas as apresentadas na Tabela 9.6.

Tabela 9.6 – Vazões mínimas, área de drenagem, declividade e densidade de drenagem

Estação	1	2	3	4	5	6	7	8
Qmin méd (m³/s)	2,6	1,49	1,43	3,44	1,37	2,53	15,12	16,21
Área (Km²)	461	291	244	579	293	486	2465	2760
I equiv (m/km)	2,69	3,94	7,20	3,18	2,44	1,25	1,81	1,59
DD (Junções/Km²)	0,098	0,079	0,119	0,102	0,123	0,136	0,121	0,137
Estação	**9**	**10**	**11**	**12**	**13**	**14**	**15**	
Qmin méd (m³/s)	21,16	30,26	28,53	1,33	0,43	39,12	45	
Área (Km²)	3939	5414	5680	273	84	8734	10192	
I equiv (m/km)	1,21	1,08	1,00	4,52	10,27	0,66	0,60	
DD (Junções/Km²)	0,134	0,018	0,141	0,064	0,131	0,143	0,133	

Solução: Inicialmente avalia-se a existência de colinearidade entre as variáveis explicativas através da matriz de correlações como apresentado a seguir.

Tabela 9.7 – Matriz de correlações

	Qmin méd (m³/s)	Área (Km²)	I equiv (m/km)	DD (Junções/Km²)
Qmin méd (m³/s)	1			
Área (Km²)	0,992	1		
I equiv (m/km)	-0,625	-0,594	1	
DD (Junções/Km²)	0,141	0,186	-0,049	1

Analisando a Tabela 9.7 observa-se que não existe colinearidade entre as variáveis independentes e que aparentemente as médias das vazões mínimas com 7 dias de duração apresentam uma forte relação linear com a área de drenagem. Assim, para verificar a linearidade entre as variáveis e a possível ocorrência de correlações espúrias foram elaborados os diagramas de dispersão da Figura 9.21.

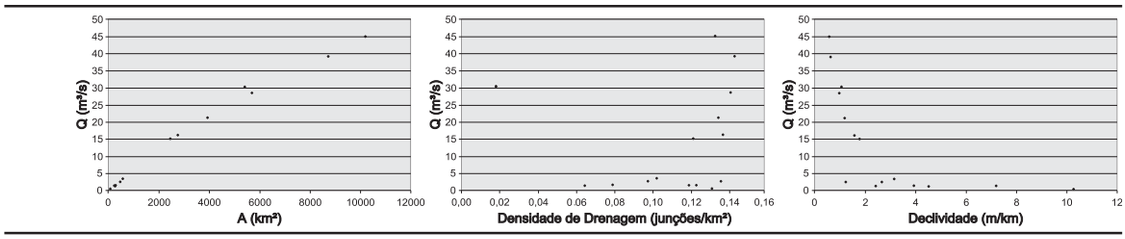

Figura 9.21 – Diagramas de dispersão

Os resultados da Tabela 9.7 e os gráficos da Figura 9.21 indicam que no modelo de regressão a ser adotado terá obrigatoriamente como uma das variáveis explicativas a área de drenagem. Sendo assim, o problema se restringe a avaliar se a inclusão de novas variáveis trará melhora significativa aos resultados do modelo. O modelo de regressão adotado será do tipo multiplicativo como apresentado a seguir:

$$Q = \beta_0 A^{\beta_1} X_2^{\beta_2} X_3^{\beta_3} \tag{9.94}$$

Após a transformação logarítmica obtêm-se:

$$ln\,Q = ln\,\beta_0 + \beta_1\, ln\,A + \beta_2\, ln\,X_2 + \beta_3\, ln\,X_3 \tag{9.95}$$

Assim, para calcular os parâmetros da equação 9.95 é necessário calcular os logaritmos das variáveis independentes e dependentes conforme está apresentado na Tabela 9.8

Tabela 9.8 – Logaritmos das variáveis

Estação	1	2	3	4	5	6	7	8
Qmin méd (m³/s)	0,9555	0,3988	0,3577	1,2355	0,3148	0,9282	2,7160	2,7856
Área (Km²)	6,1343	5,6737	5,4972	6,3604	5,6812	6,1870	7,8100	7,9230
I equiv (m/km)	0,9895	1,3712	1,9741	1,1569	0,8920	0,2231	0,5933	0,4637
DD (Junções/Km2)	-2,3276	-2,5382	-2,1299	-2,2829	-2,0977	-1,9974	-2,1095	-1,9908
Estação	9	10	11	12	13	14	15	
Qmin méd (m³/s)	3,0521	3,4098	3,3510	0,2852	-0,8440	3,6666	3,8067	
Área (Km²)	8,2787	8,5968	8,6448	5,6095	4,4296	9,0750	9,2293	
I equiv (m/km)	0,1906	0,0770	0,0000	1,5085	2,3292	-0,4155	-0,5108	
DD (Junções/Km²)	-2,0077	-4,0118	-1,9614	-2,7423	-2,0317	-1,9465	-2,0207	

A definição sobre quais serão as variáveis explicativas que comporão o modelo de estimativa das vazões mínimas é realizada através da análise das equações de regressão que contenham as seguintes variáveis independentes: somente a área de drenagem (QA); a área de drenagem e a declividade (QAI); a área de drenagem e densidade de drenagem (QADD); e área de drenagem, a declividade e a densidade de drenagem (QAIDD). A avaliação da inclusão de uma nova variável ao modelo QA é realizada através do teste da significância da equação de regressão linear múltipla e do teste de partes de um modelo de regressão linear múltipla.

Inicialmente analisa-se o modelo que utiliza somente a área de drenagem como variável independente, ou seja,

$$Q = \beta_0 A^{\beta_1} \tag{9.96}$$

$$ln\,Q = ln\,\beta_0 + \beta_1\, ln\,A \tag{9.97}$$

A Tabela 9.9 apresenta os somatórios dos quadrados e a estatística *F* do teste de significância da equação de regressão na forma de uma tabela ANOVA.

Tabela 9.9 – ANOVA modelo QA

	gl	*SQ*	*MQ*	*F*
Regressão	1	33,04321	33,04321	2915,798
Resíduo	13	0,147322	0,011332	
Total	14	33,19053		

O modelo QA é considerado significativo, pois a hipótese nula do teste, $\beta_1 = 0$, é rejeitada uma vez que:

$$(F = 2916) > [F(0{,}05;1;13) = 4{,}67] \tag{9.98}$$

Os parâmetros do modelo QA, o coeficiente de determinação e o erro padrão estão na Tabela 9.12. A inclusão da declividade como mais uma variável explicativa no modelo da equação 9.96 resulta em:

$$Q = \beta_0 A^{\beta_1} I^{\beta_2} \tag{9.99}$$

$$ln\,Q = ln\,\beta_0 + \beta_1\, ln\,A + \beta_2\, ln\,I \tag{9.100}$$

Os parâmetros do modelo QAI, o coeficiente de determinação e o erro padrão estão na Tabela 9.12. A estatística *F* do teste de significância da equação de regressão e os somatórios dos quadrados do modelo QAI estão na Tabela 9.10.

Tabela 9.10 – ANOVA modelo QAI

	gl	*SQ*	*MQ*	*F*
Regressão	2	33,07298	16,53649	1688,119
Resíduo	12	0,11755	0,009796	
Total	14	33,19053		

O modelo QAI também é considerado significativo pois a estatística do teste é maior que o valor de referência para um nível de significância de 5%, ou seja, $(F = 1688) > [F(0{,}05;2;12) = 3{,}89]$. A contribuição da variável declividade para a soma dos quadrados da regressão, $SQ\,Reg(X_I)$, considerando que a variável área de drenagem já está incluída, é estimada pela equação 9.73.

$$SQ\,Reg(X_I) = 33{,}07 - 33{,}04 = 0{,}03$$

A estatística do teste de partes de um modelo de regressão linear múltipla é calculada pela equação 9.74. Sendo assim,

$$F_P = \frac{SQ\,Reg(X_I)}{MQ\,Res} = \frac{0{,}03}{0{,}0098} = 3{,}04$$

Como $(F_P = 3{,}04) < [F(0{,}05;1;12) = 4{,}75]$, a inclusão da variável declividade não melhora significativamente o modelo quando se considera um nível de significância de 5%.

Acrescentando a densidade de drenagem como mais uma variável explicativa no modelo da equação 9.96 obtêm-se:

$$Q = \beta_0 A^{\beta_1} DD^{\beta_2} \tag{9.101}$$

$$ln\,Q = ln\,\beta_0 + \beta_1\, ln\, A + \beta_2\, ln\, DD \tag{9.102}$$

Os parâmetros do modelo QADD, o coeficiente de determinação e o erro padrão estão na Tabela 9.12. A estatística *F* do teste de significância da equação de regressão e os somatórios dos quadrados do modelo QADD estão na Tabela 9.11.

Tabela 9.11 – ANOVA modelo QADD

	gl	*SQ*	*MQ*	*F*
Regressão	2	33,04797	16,52399	1390,935
Resíduo	12	0,142557	0,01188	
Total	14	33,19053		

O teste da significância da equação de Regressão Linear Múltipla indicou que o modelo QADD pode ser considerado significativo para um nível de significância de 5%, uma vez que $\left(F = 1390{,}9\right) > \left[F\left(0{,}05;2;12\right) = 3{,}89\right]$.

A contribuição da variável densidade de drenagem para a soma dos quadrados da regressão, $SQ\,Re\,g\left(X_{DD}\right)$, considerando que a variável área de drenagem já está incluída, é estimada pela equação 9.73.

$$SQ\,Re\,g\left(X_{DD}\right) = 33{,}048 - 33{,}043 = 0{,}005$$

A estatística do teste de partes de um modelo de regressão linear múltipla é calculada pela equação 9.74. Sendo assim,

$$F_P = \frac{SQ\,Re\,g\left(X_1\right)}{MQ\,Res} = \frac{0{,}005}{0{,}01188} = 0{,}40$$

A inclusão da variável densidade de drenagem não melhora significativamente o modelo quando se considera um nível de significância de 5%, pois $\left(F_P = 0{,}40\right) < \left[F\left(0{,}05;1;12\right) = 4{,}75\right]$.

Acrescentando a densidade de drenagem como mais uma variável explicativa no modelo da equação 9.99 obtêm-se:

$$Q = \beta_0 . A^{\beta_1} . I^{\beta_2} DD^{\beta_3} \quad (9.103)$$

$$ln\, Q = ln\, \beta_0 + \beta_1\, ln\, A + \beta_2\, ln\, I + \beta_3\, ln\, DD \quad (9.104)$$

Os parâmetros do modelo QAIDD, o coeficiente de determinação e o erro padrão estão na Tabela 9.12. Entretanto, como a inclusão das variáveis declividade e densidade de drenagem mostrou-se não significativa, não é necessário avaliar o modelo a três variáveis explicativas, uma vez que teríamos um modelo significativo, mas com excesso de variáveis explicativas que não contribuem significativamente para a explicação da variância total da vazão mínima com 7 dias de duração.

Tabela 9.12 – Parâmetros dos modelos

Modelo	$ln\ (\beta_0)$	(β_1)	(β_2)	(β_3)	r^2	Erro Padrão
QA	-5,1696	0,9889			0,9956	0,1065
QAI	-5,7309	1,0551	0,1344		0,9965	0,0990
QADD	-5,24512	0,9884	-0,0348		0,9957	0,1090
QAIDD	-5,7579	1,05224	0,12930	- 0,0223	0,9965	0,1025

Analisando os resultados anteriores verifica-se que a inclusão das variáveis declividade e densidade de drenagem não traz ganhos significativos ao modelo de estimativa das vazões mínimas médias com 7 dias de duração. Dessa forma, o melhor modelo é o que adota somente a área de drenagem como variável explicativa, ou seja, a equação 9.97. A partir do comportamento dos resíduos na Figura 9.22 verifica-se que os resíduos são independentes e que a variância pode ser considerada aproximadamente constante. A Figura 9.22 apresenta o ajuste entre os resíduos e uma distribuição normal de média zero e desvio padrão igual a 0,1065.

A análise de regressão foi realizada com dados transformados, sendo assim, é necessário realizar a operação de inversão do parâmetro $ln(\beta_0)$ para definir o modelo na forma da equação 9.96.

$$\beta_0 = exp[ln(\beta_0)] = exp(-5{,}1696) = 0{,}00569$$

$$Q = 0{,}00596 A^{0{,}9889}$$

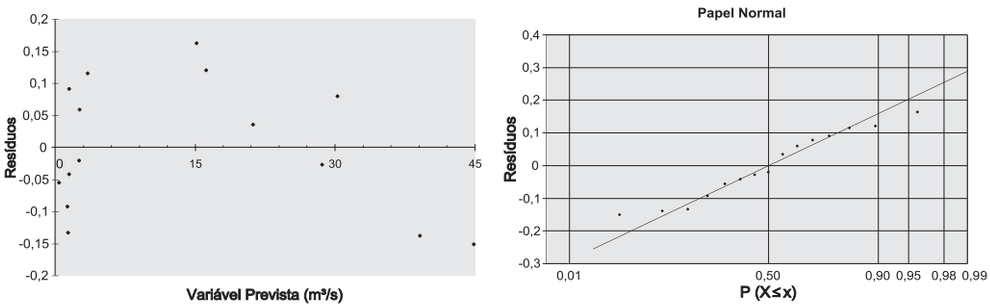

Figura 9.22 – Resíduos

Exercícios

1 – Deduzir a equação 9.28

2 – Mostrar que a correlação entre a variável independente, Y, e a sua estimativa, $\hat{Y}$, é equivalente ao coeficiente de correlação da regressão simples.

3 – A Tabela 9.13 apresenta os valores da área de drenagem e a vazão média de longo termo de 22 estações fluvioméricas da bacia do alto rio São Francisco. Estime a equação de regressão linear considerando a área de drenagem (km²) como a variável independente.
a) Verificar se os desvios atendem a hipótese de homoscedasticidade
b) Calcular o erro padrão e o coeficiente de determinação
c) Plotar os intervalos de confiança de 95% da linha de regressão e do valor previsto.

Tabela 9.13 – Áreas de drenagem e vazões médias de longo termo – Exercício 3

Estação	Área (km²)	Q_{mlt} (m³/s)	Estação	Área (km²)	Q_{mlt} (m³/s)	Estação	Área (km²)	Q_{mlt} (m³/s)
1	83,9	1,32	9	1206,9	19,3	17	5680,4	85,7
2	188,3	2,29	10	1743,5	34,2	18	8734	128
3	279,4	4,24	11	2242,4	40,9	19	10191,5	152
4	481,3	7,34	12	3727,4	65,3	20	13881,8	224
5	675,7	8,17	13	4142,9	75,0	21	14180,1	241
6	769,7	8,49	14	4874,2	77,2	22	29366,18	455
7	875,8	18,9	15	5235	77,5			
8	964,2	18,3	16	5414,2	86,8			

4 – (Adaptado de Haan,1979) Estime a equação de regressão do exercício 3 considerando a vazão média de longo termo como variável independente.
a) O modelo obtido concorda com o estimado no exercício anterior
b) Os modelos deveriam concordar? Por quê?

5 – Utilizando os dados da Tabela 9.13, estime a equação de regressão considerando uma relação potencial entre a vazão média de longo termo e a área de drenagem, ou seja, $Q = kA^C$. Compare os resultados do modelo com os obtidos no exercício 3.

6 – Em muitos casos é mais conveniente utilizar um modelo de regressão do tipo $Y = ax$, ou seja, a reta de regressão passa pela origem e o parâmetro b é igual a zero.
a) Deduza a equação normal para essa situação
b) Calcule a reta de regressão passando pela origem para os dados do exercício 3.

7) Deduzir as equações normais para o seguinte modelo parabólico $Q = a + bH + cH^2$, no qual Q denota as descargas e H os níveis d'água em uma estação fluviométrica.

8) A Tabela 9.14 apresenta uma lista de medições de descargas realizadas em um posto fluviométrico.

Tabela 9.14 – Lista de medições de descargas do exercício 8

H (m)	Q (m³/s)	H (m)	Q (m³/s)	H (m)	Q (m³/s)	H (m)	Q (m³/s)
0,0	20	1,91	170	4,73	990	8,21	2540
0,8	40	2,36	240	4,87	990	8,84	2840
1,19	90	2,70	300	5,84	1260	9,64	3320
1,56	120	4,07	680	7,19	1920	——	——

a) Faça um gráfico dos pontos cota-descarga com H em ordenadas e Q em abcissas.
b) Estime a relação cota-descarga (curva chave), usando os seguintes modelos de regressão:
- $Q = a + bH + cH^2$
- $Q = a\left(H - h_0\right)^n$ onde h_0 representa a cota para a vazão nula.

c) Desenhe no gráfico do item (a) as duas curvas ajustadas. Decida qual é o melhor modelo de regressão a partir da comparação da variância residual, dada pela fórmula $S_{res}^2 = \dfrac{\sum_{i=1}^{n}\left(Q_i^{obs} - Q_i^{est}\right)^2}{n-k-1}$, onde n é o tamanho da amostra, k é o número de variáveis explicativas e os índices *obs* e *est* referem-se aos valores observados e estimados, respectivamente.
d) Uma ponte será construída nesse local, o qual situa-se a cerca de 500 m a jusante de uma barragem. O tabuleiro dessa ponte deverá ter uma altura suficientemente grande para permitir a passagem da descarga de projeto do

vertedor da barragem que é de 5200 m^3/s. Determine a cota altimétrica mínima do tabuleiro da ponte, sabendo que o RN-2, de cota arbitrária 5,673 m em relação ao zero da régua, possui cota altimétrica 731,229 m.

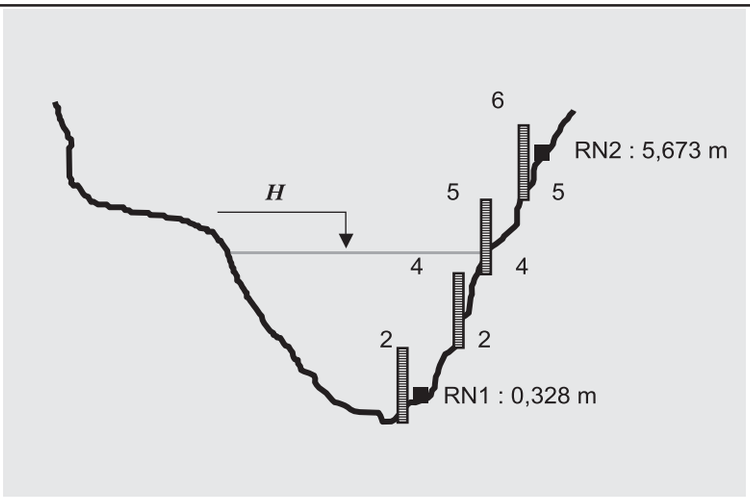

Figura 9.23 – Exercício 8

9 – A curva de dupla massa é muito utilizada em engenharia de recursos hídricos para detectar problemas na consistência de dados pluviométricos. Essa curva permite a comparação gráfica entre os valores acumulados das precipitações anuais (ou mensais) observadas na estação em análise e os valores acumulados das precipitações anuais (ou mensais) regionais, que são estimadas como as médias aritméticas de várias estações vizinhas. A Tabela 9.15 apresenta os totais anuais de uma estação em análise e da média regional. Grafe a precipitação acumulada regional no eixo das abscissas e a precipitação acumulada da estação em análise no eixo das ordenadas.

a) A partir de que ano parece haver uma mudança na inclinação da curva de dupla massa?

b) Calcule as inclinações das retas de regressão considerando dois cenários distintos. O primeiro, com os dados anteriores a aparente mudança de inclinação e o outro utilizando os dados posteriores a essa alteração.

c) Testar a hipótese das inclinações serem significativamente diferentes.

Tabela 9.15 – Dados do exercício 9

Ano	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970
Analisada (mm)	1700	1300	2100	1900	1800	1200	1450	1250	1710	1700	1400
Média Regional (mm)	1067	857	1440	1393	1233	980	1177	1043	1490	1450	1200

10 – Em um estudo de regionalização de vazões máximas, no qual foi aplicado o método *index-flood*, definiu-se uma região homogênea com 13 estações

fluviométricas. Nesse estudo as médias das vazões máximas foram utilizadas como fator de adimensionalização das séries. Defina um modelo de regressão que permita a estimativa do fator *index-flood* em locais que não possuam estações fluviométricas utilizando como possíveis variáveis explicativas as apresentadas na Tabela 9.16. Calcular o erro padrão e plotar os intervalos de confiança de 90% do plano de regressão e do valor previsto.

Tabela 9.16 – Dados do exercício 10

Estações	Q_{max} médio	Área (Km2)	P médio (m)	I equiv (m/km)	L (km)L (km)
1	12,6	83,9	1,436	10,27	18
2	29,8	188,3	1,460	3,1	26,4
3	30,4	244	1,466	7,2	18,3
4	35,5	273	1,531	4,52	40
5	31,5	291,1	1,462	3,94	32,7
6	64,7	461,4	1,400	2,69	52
7	86,9	486,4	1,369	1,25	47,3
8	78,2	578,5	1,464	3,18	41,6
9	74,5	675,2	1,485	2,96	53,8
10	241,6	2465,1	1,409	1,81	88,9
11	437,1	3939,2	1,422	1,21	187,4
12	541,7	5414,2	1,448	1,08	218,2
13	534,2	5680,4	1,449	1	236,33