

3 O MODELO DE REGRESSÃO LINEAR MÚLTIPLA

O modelo de regressão linear múltipla implica na existência de mais de uma variável preditora. Supondo a disponibilidade de k variáveis preditoras, x_1, \dots, x_k a equação de regressão múltipla pode ser representada na forma:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \epsilon_i \quad (3.1)$$

Neste caso, a representação matricial apresenta grandes vantagens como a representação compacta do modelo. Na forma matricial, o modelo de regressão linear múltipla pode ser definido como:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_k \end{bmatrix}$$

Como consequência, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ representa a forma matricial da equação: $y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \epsilon_i, \quad i=1, \dots, n$, e $\mathbf{X} = [x_0 \ x_1 \ \dots \ x_k]$, onde \mathbf{x} é um vetor coluna, sendo $x_{0i} = x_{0i} = 1$, ou: $\mathbf{X} = [1 \ x_1 \ \dots \ x_k]$.

Dessa forma, tem-se $\mathbf{Y}_{n \times 1}$, $\mathbf{X}_{n \times p}$, onde $p = k + 1$, $\boldsymbol{\beta}_{p \times 1}$, e $\boldsymbol{\epsilon}_{n \times 1}$ é o vetor de erros aleatórios com $E(\boldsymbol{\epsilon}) = 0$ e $Cov(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$, onde $\mathbf{I}_{n \times n}$ é a matriz identidade.

3.1 ESTIMADORES DE MÍNIMOS QUADRADOS NA FORMA MATRICIAL

A equação de mínimos quadrados na forma matricial é escrita como:

$$SQE(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \quad (3.2)$$

Derivando com relação ao vetor de parâmetros:

$$\begin{aligned} \frac{\partial SQE}{\partial \boldsymbol{\beta}} &= -2\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= 0 \end{aligned} \quad (3.3)$$

obtém-se a seguinte solução:

$$\begin{aligned} \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{X}^T\mathbf{Y} \\ \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \end{aligned} \quad (3.4)$$

É importante observar que $SQE = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$, uma vez que $\boldsymbol{\epsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$.

3.2 MATRIZ DE PROJEÇÃO

A matriz de projeção, H , é definida como:

$$\begin{aligned}\hat{Y} &= X\hat{\beta} \\ &= X(X^T X)^{-1} X^T Y \\ &= HY\end{aligned}\tag{3.5}$$

Ou seja, $\hat{Y} = HY$. H é uma matriz simétrica, $H_{n \times n}$, onde $H = H^T$, e idempotente: $H^2 = HH^T = H^T H = H$. A matriz H também é conhecida como matriz *chapéu* ou matriz *hat*. Conforme a definição, é a matriz H que coloca o *chapéu* no vetor Y .

3.3 VETOR DE RESÍDUOS

O vetor de resíduos, r , é a estimativa do vetor de erros, ϵ :

$$\begin{aligned}r &= \hat{\epsilon} \\ &= Y - X\hat{\beta} \\ &= Y - HY \\ &= (I - H)Y\end{aligned}\tag{3.6}$$

3.3.1 PROPRIEDADES DO VETOR DE RESÍDUOS

1. r e \hat{Y} ($\hat{Y} = \hat{\mu} = X\hat{\beta}$) são ortogonais:

$$\begin{aligned}r^T \hat{Y} &= Y^T (I - H)HY \\ &= Y^T (H - H^2)Y \\ &= 0\end{aligned}\tag{3.7}$$

pois $H^2 = H$.

2. A soma dos quadrados das observações pode ser decompostas em duas partes: a soma dos quadrados dos valores ajustados, $\hat{\mu}^T \hat{\mu}$ ou $\hat{Y}^T \hat{Y}$, e a soma dos quadrados dos resíduos, $r^T r$, também escrita na forma:

$$\sum_{i=1}^n y_i^2 = \sum_{i=1}^n \hat{\mu}_i^2 + \sum_{i=1}^n \hat{\epsilon}_i^2$$

Demonstração:

$$\begin{aligned}r^T r &= (Y - \hat{Y})^T (Y - \hat{Y}) \\ &= Y^T (I - H)^T (I - H) Y \\ &= Y^T (I - H) Y \\ &= Y^T Y - Y^T H Y \\ &= Y^T Y - (H Y)^T (H Y) \\ &= Y^T Y - \hat{Y}^T \hat{Y}\end{aligned}$$

lembrando que $r_i = \hat{\epsilon}_i$.

3.4 PROPRIEDADES DO ESTIMADOR DE MÍNIMOS QUADRADOS NA FORMA MATRICIAL

1. $E(\hat{\beta}) = \beta$:

$$\begin{aligned}
 E(\hat{\beta}) &= E[(X^T X)^{-1} X^T Y] \\
 &= E[(X^T X)^{-1} X^T (X\beta + \epsilon)] \\
 &= E[(X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T \epsilon] \\
 &= \beta + (X^T X)^{-1} X^T E(\epsilon) \\
 &= \beta
 \end{aligned} \tag{3.8}$$

onde $E(\epsilon) = 0$.

2. $Cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$:

$$\begin{aligned}
 Cov(\hat{\beta}) &= Cov[(X^T X)^{-1} X^T Y] \\
 &= (X^T X)^{-1} X^T Cov(Y) X (X^T X)^{-1} \\
 &= (X^T X)^{-1} X^T Cov(\epsilon) X (X^T X)^{-1} \\
 &= (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1} \\
 &= \sigma^2 (X^T X)^{-1}
 \end{aligned} \tag{3.9}$$

onde: $Cov(AY) = ACov(Y)A^T$. Quando o modelo é linear simples, então:

$$\sigma^2 (X^T X)^{-1} = \begin{bmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) \end{bmatrix}$$

3. $Cov(\hat{Y}) = \sigma^2 H$:

$$\begin{aligned}
 Cov(\hat{Y}) &= Cov(\hat{\mu}) \\
 &= Cov(HY) \\
 &= HCov(Y)H^T \\
 &= H\sigma^2 I H^T \\
 &= \sigma^2 HH^T \\
 &= \sigma^2 H^2 \\
 &= \sigma^2 H
 \end{aligned} \tag{3.10}$$

4. $Cov(r) = \sigma^2 (I - H)$

$$\begin{aligned}
 Cov(r) &= Cov(\hat{\epsilon}) \\
 &= Cov(Y - \hat{Y}) \\
 &= Cov(Y - HY) \\
 &= Cov[(I - H)Y] \\
 &= (I - H)Cov(Y)(I - H)^T \\
 &= (I - H)\sigma^2 I (I - H)^T \\
 &= \sigma^2 (I - H)(I - H)^T \\
 &= \sigma^2 (I - H)
 \end{aligned} \tag{3.11}$$

Neste caso, uma vez que a matriz H é idempotente, então é possível afirmar que:

$$(I - H)^2 = (I - H)(I - H)^T = (I - H)$$

É possível identificar uma grande diferença entre as propriedades do vetor de erros ϵ e o vetor dos resíduos r : $Cov(\epsilon) = \sigma^2 I$ ao passo que $Cov(r) = \sigma^2(I - H)$. Ou seja, o vetor dos erros é composto por variáveis aleatórias independentes ao passo que o vetor dos resíduos é composto por variáveis aleatórias dependentes e heterocedásticas, $Var(r_i) = \sigma^2[I - H]_{ii}$. Mesmo sendo o vetor dos resíduos um estimador do vetor dos erros, as propriedades de covariância são completamente distintas.

5. $Cov(\hat{\beta}, r) = 0$

$$\begin{aligned} Cov(\hat{\beta}, r) &= Cov\left((X^T X)^{-1} X^T Y, (I - H)Y\right) \\ &= (X^T X)^{-1} X^T Cov(Y)(I - H)^T \\ &= (X^T X)^{-1} X^T \sigma^2 I (I - H)^T \\ &= \sigma^2 \left[(X^T X)^{-1} X^T - (X^T X)^{-1} X^T X (X^T X)^{-1} X^T \right] \\ &= \sigma^2 0 \\ &= 0 \end{aligned} \tag{3.12}$$

3.5 O ESTIMADOR NÃO VICIADO PARA σ^2

Seja o vetor de variáveis aleatórias $Y_{n \times 1}$: $E(Y) = \mu$ e $Cov(Y) = \Sigma$. Inicialmente, é possível mostrar que $E(Y^T AY) = \text{tr}(A\Sigma) + E(\mu^T A\mu)$, onde $A_{n \times n}$ é uma matriz simétrica:

$$\begin{aligned} E(Y^T AY) &= \text{tr}(E[Y^T AY]) \\ &= E[\text{tr}(Y^T AY)] \\ &= E[\text{tr}(AYY^T)] \\ &= \text{tr}[E(AYY^T)] \\ &= \text{tr}[AE(YY^T)] \\ &= \text{tr}[A(Cov(Y) + \mu\mu^T)] \\ &= \text{tr}(A\Sigma) + \text{tr}(A\mu\mu^T) \\ &= \text{tr}(A\Sigma) + \text{tr}(\mu^T A\mu) \\ &= \text{tr}(A\Sigma) + \mu^T A\mu \end{aligned}$$

onde $\text{tr}(A)$ é traço da matriz A ou a soma dos termos da diagonal. Algumas propriedades interessantes:

(i) $\text{tr}(A+B) = \text{tr}(A) + \text{tr}(B)$

(ii) $\text{tr}(AC) = \text{tr}(CA)$

Vale lembrar também que: $Cov(Y) = E(YY^T) - \mu\mu^T$. No caso univariado: $Var(Y) = E(Y^2) - [E(Y)]^2$.

Vamos considerar a soma dos quadrados dos resíduos escrita na forma: $SQ_{Res} = (Y - \hat{\mu})^T (Y - \hat{\mu})$, onde $\hat{\mu} = HY$, ou seja, $SQ_{Res} = Y^T (I - H)Y$, $E(Y) = X\beta$ e $Cov(Y) = \sigma^2 I$.

Podemos então escrever a $E(SQ_{Res})$ na forma:

$$\begin{aligned} E(SQ_{Res}) &= \text{tr}[(\mathbf{I} - \mathbf{H})\sigma^2 \mathbf{I}] + \hat{\beta}^T \mathbf{X}^T (\mathbf{I} - \mathbf{H}) \mathbf{X} \hat{\beta} \\ &= \sigma^2 \text{tr}(\mathbf{I} - \mathbf{H}) + \hat{\beta}^T \mathbf{X}^T (\mathbf{I} - \mathbf{H}) \mathbf{X} \hat{\beta} \\ &= \sigma^2 \text{tr}(\mathbf{I} - \mathbf{H}) + 0 \\ &= \sigma^2 [\text{tr}(\mathbf{I}) - \text{tr}(\mathbf{H})] \\ &= \sigma^2(n - p) \end{aligned}$$

onde:

$$\begin{aligned} \text{tr}(\mathbf{I}) - \text{tr}(\mathbf{H}) &= n - \text{tr}(\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \\ &= n - \text{tr}(\mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}) \\ &= n - \text{tr}(\mathbf{I}_{p \times p}) \\ &= n - p \end{aligned}$$

Então, uma vez que $E(SQ_{Res}) = \sigma^2(n - p)$, um estimador não-viesado para σ^2 pode então ser definido como $\hat{\sigma}^2 = \frac{SQ_{Res}}{n-p}$, onde $n - p$ é o termo de correção de vício do estimador soma dos quadrados dos resíduos.

3.6 DECOMPOSIÇÃO DA SOMA DOS QUADRADOS DOS RESÍDUOS

A soma dos quadrados totais (SQ_T) é definida como a soma dos quadrados das diferenças entre os valores observados no vetor \mathbf{Y} e o vetor média amostral, $\bar{\mathbf{y}}$,

$$SQ_T = \sum_i^n (y_i - \bar{y})^2$$

É possível mostrar que SQ_T pode ser decomposta por uma componente *absorvida* pelo modelo de regressão e uma componente não *absorvida* pelo modelo de regressão, ou seja:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3.13)$$

onde $\hat{y}_i = \hat{\mu}_i = \mathbf{x}_i \hat{\beta}$. Define-se então a soma dos quadrados de regressão (SQ_{Reg}) como: $SQ_{Reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$, e a soma dos quadrados dos resíduos (SQ_{Res}) como: $SQ_{Res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$. Dessa forma, segundo a equação 3.13,

$$SQ_T = SQ_{Reg} + SQ_{Res}$$

Prova:

$$\begin{aligned} y_i - \bar{y} &= y_i - \bar{y} + \hat{\mu}_i - \hat{\mu}_i \\ &= (\hat{\mu}_i - \bar{y}) + (y_i - \hat{\mu}_i) \end{aligned}$$

elevando ambos os termos ao quadrado e aplicando a soma:

$$\begin{aligned} (y_i - \bar{y})^2 &= (\hat{\mu}_i - \bar{y})^2 + (y_i - \hat{\mu}_i)^2 + 2(\hat{\mu}_i - \bar{y})(y_i - \hat{\mu}_i) \\ \sum_i (y_i - \bar{y})^2 &= \sum_i (\hat{\mu}_i - \bar{y})^2 + \sum_i (y_i - \hat{\mu}_i)^2 + 2 \sum_i (\hat{\mu}_i - \bar{y})(y_i - \hat{\mu}_i) \end{aligned} \quad (3.14)$$

Definindo $\mathbf{1}$ como vetor unitário de dimensão n ($\mathbf{1}_i = 1$), podemos desenvolver o último termo da equação 3.14:

$$\begin{aligned} \sum_i (\hat{\mu}_i - \bar{y})(y_i - \hat{\mu}_i) &= (\hat{\mu} - \bar{y}\mathbf{1})^T(\mathbf{Y} - \hat{\mu}) \\ &= (\mathbf{Y}^T\mathbf{H} - \bar{y}\mathbf{1}^T)(\mathbf{I} - \mathbf{H})\mathbf{Y} \\ &= \mathbf{Y}^T\mathbf{H}\mathbf{Y} - \mathbf{Y}^T\mathbf{H}^2\mathbf{Y} - \bar{y}\mathbf{1}^T\mathbf{I}\mathbf{Y} + \bar{y}\mathbf{1}^T\mathbf{H}\mathbf{Y} \\ &= -\bar{y}\mathbf{1}^T\mathbf{Y} + \bar{y}\mathbf{1}^T\mathbf{H}\mathbf{Y} \\ &= 0 \end{aligned}$$

isso se deve ao fato de \mathbf{H} ser idempotente: $\mathbf{Y}^T\mathbf{H}\mathbf{Y} = \mathbf{Y}^T\mathbf{H}^2\mathbf{Y}$, e porque $\mathbf{1}^T\mathbf{H} = \mathbf{1}^T$, esta última propriedade se aplica desde que o intercepto seja mantido no modelo. **Demonstração:** se $\mathbf{X} = [\mathbf{1} \ \mathbf{X}_k]$ e $\mathbf{X}^T\mathbf{H} = \mathbf{X}^T$ então $[\mathbf{1} \ \mathbf{X}_k]^T\mathbf{H} = [\mathbf{1} \ \mathbf{X}_k]^T$, logo $\mathbf{1}^T\mathbf{H} = \mathbf{1}$.

3.7 A TABELA DE ANÁLISE DE VARIÂNCIA

Vamos considerar inicialmente a seguinte propriedade: $\mathbf{Y} \sim N_n(\mu, \Sigma)$ onde Σ é uma matriz definida positiva, então:

$$(\mathbf{Y} - \mu)^T \Sigma^{-1} (\mathbf{Y} - \mu) \sim \chi_n^2$$

no caso de uma variável aleatória com distribuição normal, $Y \sim Normal(\mu, \sigma^2)$, temos que $\left(\frac{Y-\mu}{\sigma}\right)^2 \sim \chi_1^2$. Ou seja, uma variável aleatória normal padronizada elevada ao quadrado se comporta segundo uma distribuição Chi-quadrado (χ_1^2) com um grau de liberdade.

Vamos definir também os quadrados médios dos resíduos (QM_{RES}) como: $QM_{RES} = \frac{SQ_{RES}}{n-p}$. É possível mostrar que:

1. $E[SQ_{RES}] = \sigma^2(n-p)$
2. $\frac{SQ_{RES}}{\sigma^2} \sim \chi_{n-p}^2$

Demonstração:

$$\begin{aligned} \frac{\epsilon^T \epsilon}{\sigma^2} &= \frac{(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)}{\sigma^2} \\ &= \frac{[\mathbf{r} + \mathbf{X}(\hat{\beta} - \beta)]^T[\mathbf{r} + \mathbf{X}(\hat{\beta} - \beta)]}{\sigma^2} \\ &= \frac{\mathbf{r}^T \mathbf{r} + (\hat{\beta} - \beta)^T \mathbf{X}^T \mathbf{X}(\hat{\beta} - \beta) + (\hat{\beta} - \beta)^T \mathbf{X}^T \mathbf{r} + \mathbf{r}^T \mathbf{X}(\hat{\beta} - \beta)}{\sigma^2} \\ &= \frac{\mathbf{r}^T \mathbf{r} + (\hat{\beta} - \beta)^T \mathbf{X}^T \mathbf{X}(\hat{\beta} - \beta)}{\sigma^2} \end{aligned} \tag{3.15}$$

a solução anterior é possível ao assumir que $\mathbf{r}^T \mathbf{X} = 0$. Para isso, basta considerar que $\mathbf{r} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$, então podemos desenvolver o seguinte raciocínio:

$$\begin{aligned} \mathbf{r}^T \mathbf{X} &= \mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{X} \\ &= \mathbf{Y}^T(\mathbf{X} - \mathbf{H}\mathbf{X}) \\ &= \mathbf{Y}^T(\mathbf{X} - \mathbf{X}) \\ &= 0 \end{aligned} \tag{3.16}$$

onde $\mathbf{H}\mathbf{X} = \mathbf{X}$ pois $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

A partir dos resultados demonstrados, podemos afirmar que:

$$\frac{\mathbf{r}^T \mathbf{r}}{\sigma^2} = \frac{\boldsymbol{\epsilon}^T \boldsymbol{\epsilon}}{\sigma^2} - \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2}$$

Observe que $\frac{\boldsymbol{\epsilon}^T \boldsymbol{\epsilon}}{\sigma^2}$ é uma soma de quadrados de n variáveis aleatórias independentes com distribuição normal padrão, $N(0, 1)$, e portanto tem distribuição χ_n^2 . Por outro lado, $\hat{\boldsymbol{\beta}} \sim \text{Normal}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$ e portanto $\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2} \sim \chi_p^2$. Também é conhecido que $\chi_{n-p}^2 = \chi_n^2 - \chi_p^2$, ou seja, a diferença entre duas variáveis aleatórias do tipo Chi-quadrado também é uma variável aleatória Chi-quadrado cujo grau de liberdade é a diferença dos graus de liberdade. Como consequência, $\frac{\mathbf{r}^T \mathbf{r}}{\sigma^2} \sim \chi_{n-p}^2$.

3.7.1 TESTE DE HIPÓTESES UTILIZANDO A DECOMPOSIÇÃO DA SOMA DOS QUADRADOS DOS RESÍDUOS

Utilizando as propriedades da decomposição da soma dos quadrados dos resíduos, é possível estruturar um teste de hipótese global para o modelo de regressão linear múltiplo. Seja o modelo de regressão linear múltipla, $y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \epsilon$, e a estatística F definida na forma:

$$F = \frac{QM_{Reg}}{QM_{Res}}$$

Assumindo a seguinte hipótese nula,

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

é possível mostrar que:

$$F|H_0 \sim F_{k,n-p}$$

onde $p = k + 1$ é o número total de parâmetros no modelo de regressão: $\beta_0, \beta_1, \dots, \beta_k$.

Demonstração:

De modo geral, desejamos testar a seguinte hipótese nula $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{c}$, onde

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}_{k \times p}$$

e $\mathbf{c} = \mathbf{1}_{k \times 1}$. Uma estatística de teste conveniente é $\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c}$. Então, a hipótese nula será rejeitada se $\mathbf{A}\hat{\boldsymbol{\beta}}$ for estatisticamente diferente de \mathbf{c} . Neste caso, é possível mostrar que quando a hipótese nula é verdadeira então:

$$\frac{(SQ_{Res}^{H_0} - SQ_{Res}) / k}{SQ_{Res} / (n - p)} \sim F_{k,n-p} \quad (3.17)$$

onde $SQ_{Res}^{H_0}$ é a soma dos quadrados dos resíduos assumindo que a hipótese nula é verdadeira. Neste caso, o modelo se reduz a $y = \beta_0 + \epsilon$, e a solução de mínimos quadrados é $\hat{\beta}_0 = \bar{y}$. Portanto, $SQ_{Res}^{H_0} = SQ_T - SQ_{Res}$ e, como consequência,

$$SQ_{Res}^{H_0} - SQ_{Res} = SQ_T - SQ_{Res} = SQ_{Reg}$$

ou seja, se a hipótese nula é verdadeira, então:

$$\frac{(SQ_{Reg}) / k}{SQ_{Res} / (n - p)} \sim F_{k, n-p} \quad (3.18)$$

A Tabela de Análise de Variância apresenta em detalhes a decomposição da soma dos quadrados dos resíduos e a razão entre a soma dos quadrados de regressão e a soma dos quadrados dos resíduos.

Tabela 3.1: Tabela de análise de variância (ANOVA) para o modelo de regressão linear múltipla

Fonte	graus de liberdade	Soma dos Quadrados	Quadrados Médios	Estatística F	valor-P
Regressão	k	$SQ_{Reg} = \sum_i (\hat{y}_i - \bar{y})^2$	$QM_{Reg} = \frac{SQ_{Reg}}{k}$	$F = \frac{QM_{Reg}}{QM_{Res}}$	-
Erro	n-p	$SQ_{Res} = \sum_i (y_i - \hat{y}_i)^2$	$QM_{Res} = \frac{SQ_{Res}}{n-p}$		
Total	n-1	$SQ_T = \sum_i (y_i - \bar{y})^2$	$QM_T = \frac{SQ_T}{n-1}$		

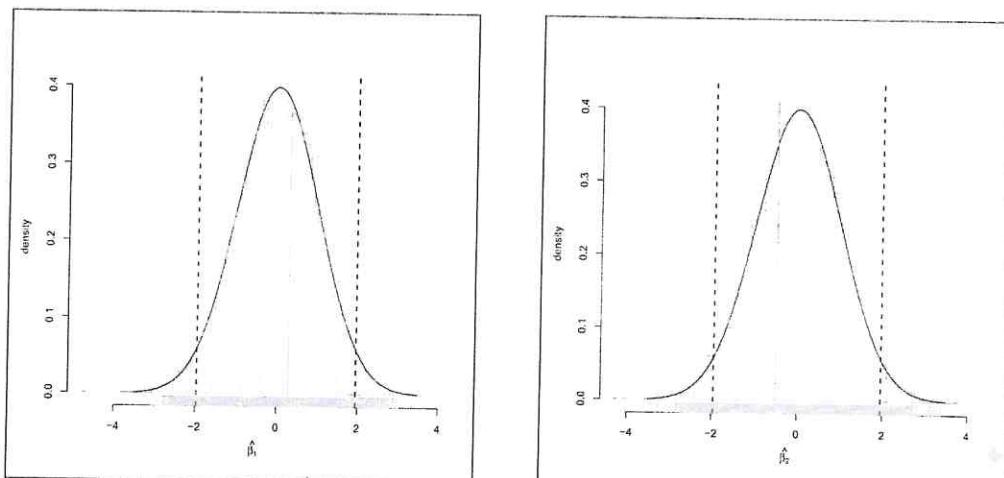
Vale destacar que, quando a hipótese nula é verdadeira, $E[SQ_{Reg}|H_0] = \sigma^2 \times k$. Ou seja, a soma dos quadrados de regressão dividido por k também é um estimador não viciado para σ^2 . Então, caso a hipótese nula seja verdadeira $F_{obs} \approx 1$. Maiores detalhes das propriedades do teste F podem ser encontrados em Seber e Lee (2012), pg. 99.

3.8 INTERVALOS DE CONFIANÇA PARA O VETOR $\hat{\beta}$

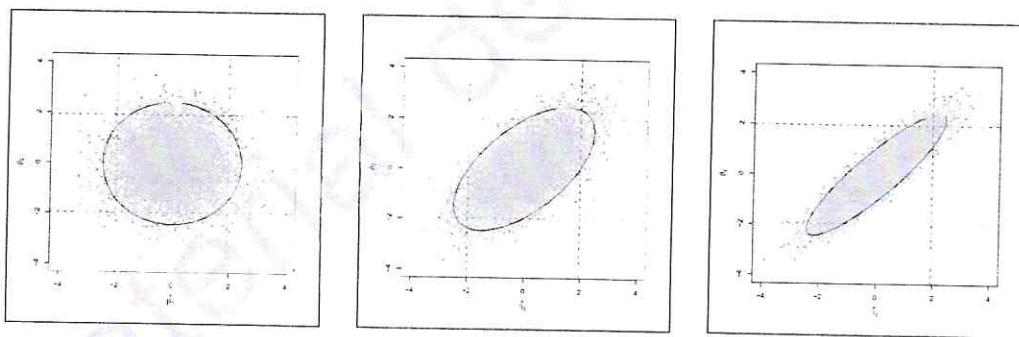
Antes de considerar o caso geral para o vetor $\hat{\beta}$ vamos abordar o seguinte caso: seja $\hat{\beta}_1$ e $\hat{\beta}_2$ duas variáveis aleatórias normais (gaussianas) identicamente distribuídas com média zero, e variância unitária. A princípio, a suposição de independência não será assumida. Seja definido o vetor $\hat{\beta}^T = [\hat{\beta}_1, \hat{\beta}_2]$. Então $\hat{\beta} \sim Normal(\mathbf{0}, \Sigma)$, onde $\mathbf{0}$ é um vetor bidimensional de zeros e Σ é a matriz de covariância,

$$\Sigma = \begin{bmatrix} Var(\hat{\beta}_1) & Cov(\hat{\beta}_1, \hat{\beta}_2) \\ Cov(\hat{\beta}_2, \hat{\beta}_1) & Var(\hat{\beta}_2) \end{bmatrix}$$

onde $Cov(\hat{\beta}_1, \hat{\beta}_2) = Cov(\hat{\beta}_2, \hat{\beta}_1)$. Faixas de referência podem ser construídas individualmente para as variáveis aleatórias $\hat{\beta}_1$ e $\hat{\beta}_2$ utilizando os elementos da diagonal da matriz Σ , como ilustrado na figura 3.1.

(a) Intervalo de confiança para $\hat{\beta}_1$.(b) Intervalo de confiança para $\hat{\beta}_2$.Figura 3.1: Intervalos de confiança construídos a partir dos elementos da diagonal da matriz Σ .

Ao se considerar a estrutura de correlação entre variáveis aleatórias na construção de faixas de referência, ou mesmo intervalos de confiança, sob as condições de normalidade, é possível afirmar que $(\hat{\beta} - E[\hat{\beta}])^T \Sigma^{-1} (\hat{\beta} - E[\hat{\beta}]) \sim \chi_p^2$ onde p é a dimensão do vetor $\hat{\beta}$. O intervalo de confiança no caso de variáveis aleatórias normais independentes e identicamente distribuídas onde $Cov(\hat{\beta}_2, \hat{\beta}_1) = 0$, pode ser construído utilizando os quartis da distribuição χ_p^2 . A figura 3.2 apresenta diferentes faixas de referência (bi-dimensionais) para as variáveis aleatórias normais definidas no vetor β , assumindo diferentes valores para a covariância. As linhas tracejadas representam as faixas de referência construídas de forma independente. As linhas contínuas delimitam a real região de confiança, considerando a matriz de covariância Σ .

(a) $Cov(\hat{\beta}_1, \hat{\beta}_2) = 0$ (b) $Cov(\hat{\beta}_1, \hat{\beta}_2) = 0.6$ (c) $Cov(\hat{\beta}_1, \hat{\beta}_2) = 0.9$ Figura 3.2: Comparação de intervalos de confiança marginais e intervalos de confiança considerando a matriz de correlação Σ .

Ou seja, a suposição de normalidade e independência de variáveis aleatórias não implica na independência de intervalos de confiança, como pode ser observado na figura 3.2 (a). O uso de faixas de referências independentes é aplicável, por exemplo, ao considerarmos variáveis aleatórias independentes com distribuição uniforme.

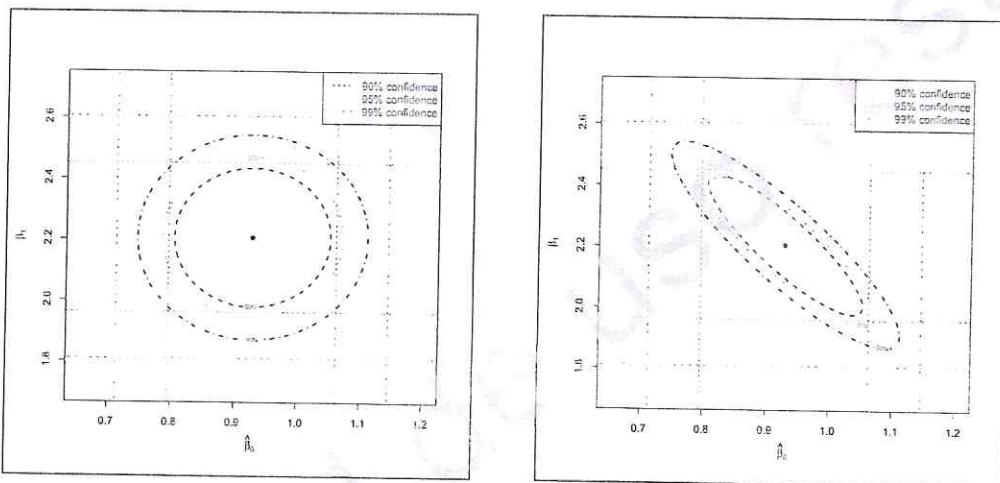
3.9 INTERVALOS DE CONFIANÇA PARA O VETOR $\hat{\beta}$ UTILIZANDO A ESTATÍSTICA F

Uma característica importante para a estatística F é a possibilidade de definir intervalos de confiança para o vetor de estimadores $\hat{\beta}$ considerando a estrutura de covariância $\text{Cov}(\hat{\beta}) = \sigma^2(X^T X)^{-1}$. Como $\hat{\beta} \sim \text{Normal}(\beta, \sigma^2(X^T X)^{-1})$, então $\frac{(\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta)}{\sigma^2} \sim \chi_p^2$. Como σ^2 é desconhecido então:

$$\frac{\frac{(\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta)}{p\sigma^2}}{\frac{SQ_{Res}}{\sigma^2(n-p)}} \sim F_{p,n-p} \quad (3.19)$$

Como σ^2 é definido tanto no numerador quanto no denominador da equação 3.19, o mesmo é cancelado da equação. Portanto, a propriedade exibida na equação 3.19 não requer nenhuma estimativa de σ^2 .

Como exemplo, um intervalo de confiança de 95% para o vetor β pode ser definido utilizando o percentil $F_{(0.95,p,n-p)}$. A Figura 3.3 ilustra a região de confiança utilizando a estatística F . Na Figura 3.3 (a) os limites são calculados utilizando somente os elementos da diagonal da matriz $X^T X$ ao passo que na Figura 3.3 (b) todos os elementos da matriz $X^T X$ são considerados.



(a) Limites de confiança utilizando somente os elementos da diagonal da matriz $X^T X$. As linhas tracejadas representam os intervalos de 90% e 95% de confiança utilizando a estatística F .

(b) Limites de confiança utilizando todos os elementos da matriz $X^T X$. As linhas tracejadas representam os intervalos de 90% e 95% de confiança utilizando a estatística F .

Figura 3.3: Intervalos ou limites de confiança calculados a partir da estatística F e considerando somente os elementos da diagonal da matriz de covariância (a) e todos os elementos da matriz de covariância (b).

3.9.1 EXEMPLO DO USO DA MATRIZ DE CORRELAÇÃO DOS ESTIMADORES PARA TESTE DE HIPÓTESE

Suponha que, para o exemplo de regressão linear simples, desejamos testar a hipótese de que a empresa pratica uma política de salário inicial de R\$ Mil 1,70 com ganhos anuais de R\$ Mil 0,10. Ou seja, queremos testar a hipótese de que $\beta_0 = 1,70$ e $\beta_1 = 0,10$. Vamos supor um nível de significância de 5%. O resultado do ajuste do modelo, e as estimativas dos intervalos de confiança marginais são mostrados a seguir:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
β_0	1.700	0.050	34.00	0.000
β_1	0.100	0.010	10.00	0.000

```
(Intercept) 1.806330  0.081610  22.13  <2e-16 ***
experiencia 0.100759  0.005014  20.10  <2e-16 ***

> confint(modelo, level = 0.95)
              2.5 \% 97.5 \%
(Intercept) 1.6382511 1.9744096
experiencia 0.0904328 0.1110857
```

Uma análise dos resultados, indicam que os valores definidos na hipótese de interesse estão contidos nos intervalos de confiança e, portanto, há evidência de que a hipótese de interesse é verdadeira. Esta análise, embora comumente aplicada, está errada, pois não considera a estrutura de correlação entre os estimadores. Uma região de confiança para duas variáveis aleatórias correlacionadas e supondo uma distribuição gaussiana multivariada pode ser definida da seguinte forma: se $\beta \sim \text{Normal}(\mu, \Sigma)$, então $(\beta - \mu)^T \Sigma^{-1} (\beta - \mu) \sim \chi_p^2$. A figura 3.4 mostra a região de confiança para os parâmetros β_0 e β_1 utilizando a matriz de covariância amostral, S . Os resultados mostram que, considerando a estrutura de correlação entre os estimadores, a hipótese de interesse ($\beta_0 = 1,70$ e $\beta_1 = 0,10$) está fora da região de confiança e portanto deve ser rejeitada.

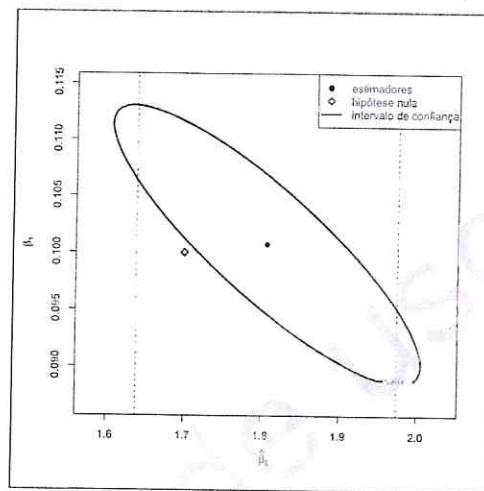


Figura 3.4: Intervalos de confiança utilizando a distribuição marginal e a matriz de covariância amostral.

É importante destacar que o exemplo proposto avalia os estimadores do intercepto (β_0) e da inclinação (β_1) para fins didáticos. A análises desses dois estimadores em particular é *incomum* pois, na prática, não há interesse em avaliar a correlação desses dois estimadores. Por outro lado, é de grande utilidade avaliar as propriedades de estimadores associados a diferentes variáveis preditoras no modelo de regressão múltipla, por exemplo $\hat{\beta}_2$ e $\hat{\beta}_3$.

3.10 O COEFICIENTE DE DETERMINAÇÃO AJUSTADO R_{adj}^2

Em geral, o coeficiente de determinação R^2 sempre aumenta quando uma nova variável é adicionada ao modelo, independente da real contribuição da variável na soma dos quadrados de regressão. No caso do modelo de regressão múltipla é possível utilizar o coeficiente de determinação ajustado, R_{adj}^2 :

$$R_{adj}^2 = 1 - \frac{SQ_{Res}/(n-p)}{SQ_T/(n-1)} \quad (3.20)$$

O coeficiente de determinação ajustado utiliza os quadrados médios dos resíduos, $QM_{Res} = SQ_{Res}/(n-p)$, que é uma estimativa não viciada de σ^2 . Na prática o coeficiente de deter-

minação ajustados irá aumentar se a variável adicionada ao modelo contribuir para a redução do valor médio da soma dos quadrados dos resíduos. O coeficiente de determinação ajustado é penalizado nos casos em que são adicionadas ao modelo variáveis que não apresentam uma contribuição significativa. A figura 3.5 ilustra o comportamento de R^2 e R_{adj}^2 em um problema de regressão linear simples com tamanho de amostra $n = 500$, ao qual foram adicionados, gradualmente, 400 novos preditores gerados a partir de uma distribuição uniforme padrão. Em geral, o valor de R^2 aumenta *falsamente* indicando uma melhora no desempenho do modelo. O valor de R_{adj}^2 , em média, não se altera e começa a cair quando o número de variáveis se aproxima de 400.

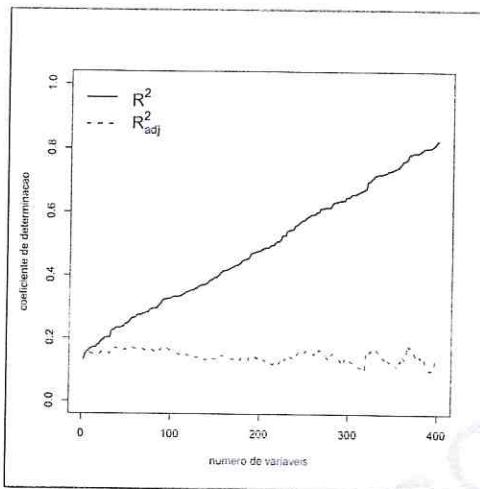


Figura 3.5: Comparação entre o coeficiente de determinação (R^2) e o coeficiente de determinação ajustado (R_{adj}^2). Em geral, R^2 aumenta à medida que novas variáveis são incorporadas ao modelo. O valor de R_{adj}^2 aumenta caso a variável seja capaz de aumentar os quadrados médios dos resíduos, o que não é o caso no exemplo.

3.11 COLINEARIDADE E MULTICOLINEARIDADE

Colinearidade, como o próprio nome indica, caracteriza o efeito nas estimativas e inferência do modelo de regressão decorrentes de uma forte correlação linear entre as variáveis regressoras ou preditoras. Para ilustrar a colinearidade uma base de dados é gerada utilizando o código a seguir.

```

rho <- 0
n <- 30
X <- rmvnorm(n, mean=c(0,0), sigma = matrix(c(1, rho, rho, 1), ncol=2))
x1 <- X[,1]
x2 <- X[,2]
Y <- -1.5*x1 + 1.5*x2 + rnorm(n, sd=0.8)
summary(lm(y ~ x1 + x2))

```

O código permite gerar uma amostra de tamanho $n = 30$ para duas variáveis preditoras: x_1 e x_2 . A variável resposta é definida pela equação $y = 0,5 - 1,5x_1 + 1,5x_2 + \epsilon$, onde ϵ é uma variável aleatória normal com média zero e variância $\sigma^2 = 0,8^2$. A correlação linear entre as variáveis preditoras é definida pelo parâmetro de correlação rho (ρ). Neste exemplo, a correlação é zero. A figura 3.6 mostra os gráficos de dispersão, par a par, das variáveis de regressão. o resultado do ajuste do modelo de regressão é mostrado a seguir.

```
> summary(lm(y ~ x1 + x2))
```

```

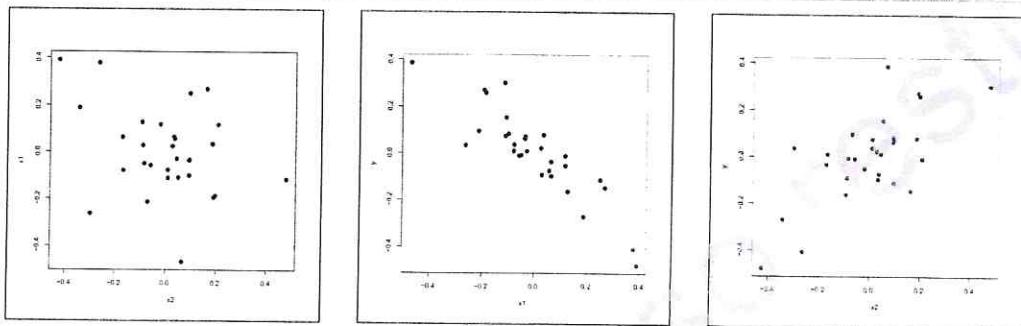
Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8498 -0.3830  0.1525  0.5901  1.1717 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.02764   0.14502  -0.191   0.85    
x1          -1.27288   0.16062  -7.925 1.61e-08 ***  
x2           1.27083   0.17628   7.209 9.40e-08 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7631 on 27 degrees of freedom
Multiple R-squared:  0.8654, Adjusted R-squared:  0.8554 
F-statistic: 86.8 on 2 and 27 DF,  p-value: 1.746e-12

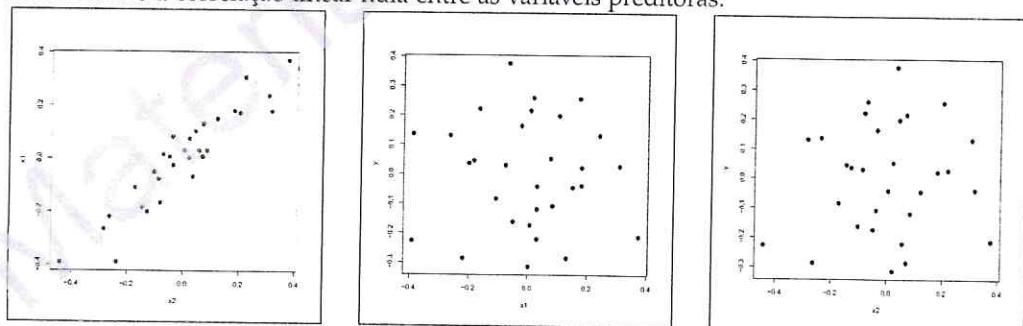
```



(a) Correlação x_1 versus x_2 . (b) Correlação y versus x_1 . (c) Correlação y versus x_2 .

Figura 3.6: Gráficos de correlação das variáveis dependentes e independentes assumindo a correlação entre as variáveis preditoras igual a zero ($\rho = 0$).

A seguir, o mesmo resultado é apresentado mas alterando o parâmetro de correlação para $\rho = 0.96$. É possível observar nos resultados apresentados na figura 3.7 e no resultado do ajuste do modelo (mostrado na sequência) que o P-valor para as estimativas dos coeficientes e os respectivos desvios padrão estão bem elevados quando comparados com os resultados obtidos considerando a correlação linear nula entre as variáveis preditoras.



(a) Correlação x_1 versus x_2 . (b) Correlação y versus x_1 . (c) Correlação y versus x_2 .

Figura 3.7: Gráficos de correlação das variáveis dependentes e independentes assumindo a correlação entre as variáveis preditoras igual a 0,96 ($\rho = 0.96$).

```
> summary(lm(y ~ x1 + x2))
```

Call:

```

lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.57258 -0.74231  0.02168  0.64805  1.62060 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.09054   0.18110   0.500   0.621    
x1          -0.83854   0.46738  -1.794   0.084 .  
x2          0.94147   0.51021   1.845   0.076 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8605 on 27 degrees of freedom
Multiple R-squared:  0.1132, Adjusted R-squared:  0.04752 
F-statistic: 1.723 on 2 and 27 DF,  p-value: 0.1975

```

Em um primeiro momento, é possível definir a colinearidade como o efeito da alta correlação linear entre duas variáveis preditoras que tem como consequência o aumento da variância dos estimadores e um comprometimento dos resultados de inferência estatística (teste de hipótese) dos estimadores.

Para caracterizar, matematicamente, o efeito de colinearidade, será considerado um exemplo em que as variáveis de regressão foram previamente padronizadas e o modelo regressão não contém o intercepto. Padronização, neste caso, é a diferença entre a variável de interesse e a sua média amostral, e posterior divisão do resultado pelo desvio padrão amostral, como mostrado no código a seguir.

```

s1 ← sum((x1 - mean(x1))^2)
x1 ← (x1 - mean(x1))/sqrt(s1)

s2 ← sum((x2 - mean(x2))^2)
x2 ← (x2 - mean(x2))/sqrt(s2)

sy ← sum((y - mean(y))^2)
y ← (y - mean(y))/sqrt(sy)

```

Seja, então, a equação de regressão linear $Y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$. O estimador na forma matricial é escrito por $(X^T X) \hat{\beta} = X^T Y$. Como as variáveis foram previamente padronizadas, o estimador pode ser reescrito como:

$$\begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix} \quad (3.21)$$

onde r_{12} é o coeficiente de correlação linear amostral entre as variáveis preditoras (x_1 e x_2) e r_{1y} é o coeficiente de correlação linear amostral entre a variável preditora x_1 e a variável resposta y . Como consequência, a matriz inversa que compõe o estimador, $\hat{\beta} = (X^T X)^{-1} X^T Y$, será definida na forma:

$$(X^T X)^{-1} = \begin{bmatrix} \frac{1}{1-r_{12}^2} & \frac{-r_{12}}{1-r_{12}^2} \\ \frac{-r_{12}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{bmatrix} \quad (3.22)$$

considerando que $r_{12} = r_{21}$. É importante destacar que a variância do estimador $\hat{\beta}$ depende da inversa da matriz, definida na equação 3.9, $Cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$. Uma importante característica da equação 3.22 é o fato de que quanto mais próximo da unidade for o valor absoluto da correlação entre as variáveis preditoras, $r_{12}^2 \approx 1$, maior será o inverso desse valor, $\frac{1}{1-r_{12}^2}$, que tenderá para o infinito. Ou seja, quanto mais próximo da unidade for a correlação linear entre as variáveis preditoras, maior será a variância dos estimadores. Em um efeito cascata, quanto maior a variância dos estimadores, menor será a chance de rejeitar a hipótese nula $H_0 : \beta_1 = 0$ ou $H_0 : \beta_2 = 0$. Ou seja, maior será o P-valor para ambas as hipóteses nulas.

Uma forma alternativa para exemplificar o efeito da colinearidade é o fato de que se duas variáveis preditoras apresentam correlação linear próxima da unidade, significa que ambas as variáveis contém, praticamente, a mesma informação. Então, em um modelo de regressão, ambas as variáveis preditoras carregam a mesma informação com relação à variável dependente Y . Neste caso, o ajuste do modelo não será capaz de distinguir qual é a variável mais relevante. Pelo contrário, existe a possibilidade do resultado final indicar que nenhuma das variáveis é estatisticamente relevante. Uma alternativa simples para solucionar este problema é eliminar uma das variáveis e ajustar novamente o modelo. Na prática, é comum eliminar a variável que apresenta o maior P-valor.

Para o caso de regressão múltipla, a colinearidade pode ser generalizada indicando dependência linear entre uma variável independente e uma combinação linear das demais variáveis independentes. É o caso, por exemplo, da seguinte estrutura de dependência:

$$x_1 = \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3 + \zeta \quad (3.23)$$

onde ζ é a componente da variável preditora x_1 que não é linearmente correlacionada com as variáveis x_2 e x_3 . O objetivo da análise de multicolinearidade não é identificar os parâmetros da dependência linear mas de avaliar a existência, ou não, da dependência linear. Então, supondo a matriz inversa $C = (X^T X)^{-1}$ é possível generalizar o caso univariado para o modelo de regressão múltipla definindo $C_{jj} = \frac{1}{1 - R_j^2}$ onde R_j^2 é o coeficiente de determinação em um modelo de regressão onde a j -ésima variável regressora é definida como a variável dependente e as demais variáveis regressoras representam as variáveis independentes. Define-se então o Fator de Inflação da Variância (Variance inflation factor - VIF) por:

$$VIF_j = \frac{1}{1 - R_j^2} \quad (3.24)$$

lembrando que $Var(\hat{\beta}_j) = \sigma^2 \times VIF_j$ se as variáveis forem padronizadas. Ou seja, a variância do estimador pode ser inflada pela existência de correlações lineares entre variáveis independentes. O código a seguir mostra como estimar o VIF para o modelo de regressão múltiplo.

```
> require(car)
> modelo <- lm(mpg ~ disp + hp + wt + drat, data=mtcars)
> vif(modelo)
      disp      hp      wt      drat 
 8.209402 2.894373 5.096601 2.279547
```

Um código didático (alternativo) é apresentado a seguir.

```
X <- model.matrix(modelo)
X <- X[, -1] ## Retira o intercepto

for(i in 1:dim(X)[2]){
  media.X <- mean(X[, i])
  ss.X <- sum((X[, i] - media.X)^2)
  X[, i] <- (X[, i] - media.X)/sqrt(ss.X)
}

VIF <- diag(solve(t(X) %*% X))
```

Na prática, quando $VIF(\hat{\beta}_j) > 10$ diz-se que a multicolinearidade é elevada, indicando a necessidade de ajuste das variáveis independentes. Sem a padronização, é possível mostrar que $Var(\hat{\beta}_j) \propto \frac{1}{1 - R_j^2}$.

Uma interpretação alternativa para o VIF pode ser construído a partir da equação 3.24. Colocando em evidência o coeficiente de determinação da j -ésima variável preditora, obtém-se o seguinte resultado:

$$R_j^2 = 1 - \frac{1}{VIF_j} \quad (3.25)$$

Ou seja, supondo o limite para o qual $VIF = 10$, este valor representa um coeficiente de determinação de $R_j^2 = 0.90$ (90%). Portanto, uma forte multicolinearidade representa um modelo de regressão onde a j -ésima variável regressora pode ser predita pelas demais variáveis regressoras em um modelo de regressão linear para o qual o coeficiente de determinação é maior ou igual a 0.90.

3.12 DECOMPOSIÇÃO PARCIAL DA SOMA DOS QUADRADOS DOS ERROS

A estatística F também pode ser utilizada para avaliar a contribuição de uma variável, ou grupo de variáveis na soma dos quadrados de regressão. Considere o modelo de regressão com k variáveis independentes $\mathbf{Y} = \mathbf{X}\beta + \epsilon$. Queremos investigar se um conjunto de variáveis $r < k$ contribui significantemente para o modelo de regressão. Seja então o vetor de coeficientes definido como $\beta^T = [\beta_1 \ \beta_2]$ onde β_1 é um vetor de dimensão $(p - r) \times 1$ e β_2 possui dimensão $r \times 1$. Queremos testar a hipótese:

$$H_0 : \beta_2 = 0$$

O modelo de regressão pode ser escrito como: $\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon$. Para avaliar a contribuição de β_2 , inicialmente ajustamos o modelo assumindo que a hipótese nula é verdadeira, ou seja $\mathbf{Y} = \mathbf{X}_1\beta_1 + \epsilon$ e definimos a soma dos quadrados de regressão deste modelo por $SQ_{Reg}(\beta_1)$. A soma dos quadrados atribuída ao vetor β_2 considerando o vetor β_1 já no modelo é definido como:

$$SQ_{Reg}(\beta_2|\beta_1) = SQ_{Reg}(\beta) - SQ_{Reg}(\beta_1)$$

com r graus de liberdade. Esta soma de quadrados é chamada de *soma extra de quadrados atribuída a β_2* e representa o aumento na soma de quadrados de regressão ao se adicionar os parâmetros definidos em β_2 . A estatística F de teste é

$$F|H_0 = \frac{SQ_{Reg}(\beta_2|\beta_1)/r}{QM_{Res}} \sim F_{(r,n-p)}$$

O exemplo a seguir ilustra o uso da decomposição da soma dos quadrados na análise de uma nova variável. São ajustados dois modelos, o primeiro contém todas as variáveis de interesse e o segundo modelo não contém uma das variáveis (x_4). Ambos o ajuste do modelo completo e a análise da decomposição da soma parcial dos quadrados dos erros são apresentados. É possível perceber que o P-valor associado à variável x_4 é idêntico em ambos os casos.

```
> model01 <- lm(y ~ x1 + x2 + x3 + x4, data=dados)
> model02 <- lm(y ~ x1 + x2 + x3, data=dados)
> summary(model01)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = dados)

Residuals:
Min      1Q  Median      3Q     Max 
-1.48849 -0.42409 -0.00011  0.40310  1.80726 

Coefficients:
Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.73077   0.87703   0.833   0.4069    
x1          0.69854   0.06754  10.343 <2e-16 ***
x2          0.45769   0.17617   2.598   0.0109 *  
x3         -0.01371   0.01137  -1.206   0.2309    
x4          0.08404   0.06080   1.382   0.1702    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.748 on 92 degrees of freedom
```

```
Multiple R-squared:  0.5976,   Adjusted R-squared:  0.5801
F-statistic: 34.15 on 4 and 92 DF,  p-value: < 2.2e-16
```

```
> anova(model01, model02)
Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3 + x4
Model 2: y ~ x1 + x2 + x3
Res.Df   RSS Df Sum of Sq    F Pr(>F)
1     92 51.477
2     93 52.546 -1   -1.069 1.9106 0.1702
```

Esta abordagem também é conhecida como análise da decomposição parcial da soma dos quadrados dos erros. O objetivo é avaliar, estatisticamente, a contribuição na redução da soma dos quadrados dos erros de uma nova variável, ou de um conjunto de novas variáveis, dado um modelo pré-conhecido.

3.12.1 CASO PARTICULAR:

Seja um modelo de regressão contendo 3 variáveis independentes, na forma:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Neste caso, pode-se representar a soma dos quadrados totais na forma:

$$\begin{aligned} SQ_T &= SQ_{Reg}(\beta_1, \beta_2, \beta_3 | \beta_0) + SQ_{res} \\ &= SQ_{Reg}(\beta_1 | \beta_0) + SQ_{Reg}(\beta_2 | \beta_1, \beta_0) + SQ_{Reg}(\beta_3 | \beta_1, \beta_2, \beta_0) + SQ_{res} \end{aligned} \quad (3.26)$$

ou seja:

$$\begin{aligned} SQ_{Reg} &= SQ_{Reg}(\beta_1, \beta_2, \beta_3 | \beta_0) \\ &= SQ_{Reg}(\beta_1 | \beta_0) + SQ_{Reg}(\beta_2 | \beta_1, \beta_0) + SQ_{Reg}(\beta_3 | \beta_1, \beta_2, \beta_0) \end{aligned} \quad (3.27)$$

É importante destacar que, segundo a equação 3.27, a análise da decomposição da soma dos quadrados dos erros pressupõe uma ordem pré estabelecida de comparação dos modelos. Ou seja, a equação 3.27 assume a seguinte sequência: (a) $y = \beta_0$, (b) $y = \beta_0 + \beta_1 x_1$, (c) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ e (d) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$. Neste sentido, é mais provável de ocorrer diferenças entre os P-valores estimados pelo modelo de regressão contendo todas as variáveis e os P-valores obtidos pela análise da decomposição parcial da soma dos quadrados dos erros. Este fato é ilustrado no exemplo a seguir.

```
> modelo <- lm(y ~ x1 + x2 + x3, data=dados)
> summary(modelo)

Call:
lm(formula = y ~ x1 + x2 + x3, data = dados)

Residuals:
Min      1Q  Median      3Q      Max 
-1.60891 -0.44895 -0.02807  0.45602  1.91755 

Coefficients:
Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.146941  0.772372   0.190  0.84953  
x1          0.687819  0.067418  10.202 < 2e-16 *** 
x2          0.549937  0.163838   3.357  0.00114 **  
x3          -0.009486  0.011003  -0.862  0.39081  
---

```

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7517 on 93 degrees of freedom
Multiple R-squared: 0.5892, Adjusted R-squared: 0.576
F-statistic: 44.47 on 3 and 93 DF, p-value: < 2.2e-16

> anova(modelo)
Analysis of Variance Table

Response: y
Df Sum Sq Mean Sq F value    Pr(>F)
x1      1 69.003 69.003 122.1258 < 2.2e-16 ***
x2      1  5.948  5.948 10.5279  0.001634 **
x3      1  0.420  0.420  0.7433  0.390815
Residuals 93 52.546   0.565
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Ou seja, no primeiro caso o P-valor é calculado a partir da inversa da matriz de regressão $(X^T X)^{-1}$ que está sujeita ao efeito de multicolinearidade. No segundo caso, avalia-se a contribuição de cada variável na redução da soma dos quadrados dos erros, assumindo uma ordem hierárquica entre as variáveis preditoras. Embora estejam próximos, os P-valores utilizando a decomposição parcial, são influenciados pela ordem das variáveis conforme demonstra o exemplo a seguir:

```

> modelo <- lm(y ~ x3 + x2 + x1, data=dados)
> anova(modelo)
Analysis of Variance Table

Response: y
Df Sum Sq Mean Sq F value    Pr(>F)
x3      1 3.679   3.679  6.5116  0.01235 *
x2      1 12.882  12.882 22.7999 6.675e-06 ***
x1      1 58.810  58.810 104.0856 < 2.2e-16 ***
Residuals 93 52.546   0.565
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Neste caso, a interpretação é a seguinte: a variável x_3 resultou em uma significativa redução da SQE em relação ao modelo que continha apenas o intercepto (β_0); a variável x_2 resultou em uma significativa redução da SQE em relação ao modelo com intercepto e x_3 ; finalmente, a variável x_1 resultou em uma significativa redução da SQE em relação ao modelo com as variáveis x_3 e x_2 . É conhecido que os P-valores obtidos a partir da matriz $(X^T X)^{-1}$ estão sujeitos à multicolinearidade. Por outro lado, os P-valores utilizando a decomposição parcial da soma dos quadrados dos erros são sensíveis à ordem das variáveis no modelo. Uma maneira de tornar a análise da decomposição da SQE mais robusta consiste em garantir que a ordem de entrada das variáveis no modelo seja em ordem decrescente da contribuição parcial na redução da SQE. Por exemplo, inicialmente a variável que apresenta a maior redução da SQE é inserida no modelo. Na sequência, a segunda variável que apresenta a maior redução da SQE parcial é inserida no modelo, e assim por diante. Utilizando este procedimento, procura-se minimizar o efeito da multicolinearidade realizando uma análise hierárquica da decomposição parcial da soma dos quadrados dos erros.

3.13 ANÁLISE DOS RESÍDUOS

Ao realizar a análise dos resíduos do modelo, os mesmos são comumente transformados para que sejam comparados com os quartis de uma normal padronizada, ou com uma distribuição t -Student com $n - p$ graus de liberdade. Esta análise seria correta, caso os resíduos fossem variáveis aleatórias independentes. Entretanto, embora os resíduos sejam variáveis aleatórias com distribuição normal, os mesmos não são independentes. Vale recordar que $r \sim Normal(0, \sigma^2[I - H])$. Portanto, uma análise mais coerente seria realizar a padronização dos resíduos considerando os

elementos da diagonal da matriz de covariância:

$$r_i^* = \frac{y_i - \hat{\mu}_i}{\sigma \sqrt{(1 - h_{ii})}} \quad (3.28)$$

Neste caso, é possível afirmar que $r_i^* \sim \text{Normal}(0, 1)$. Para o exemplo proposto, a comparação dos resíduos padronizados utilizando o desvio estimado a partir da soma dos quadrados dos resíduos e os elementos da diagonal da matriz H é apresentada na figura 3.8.

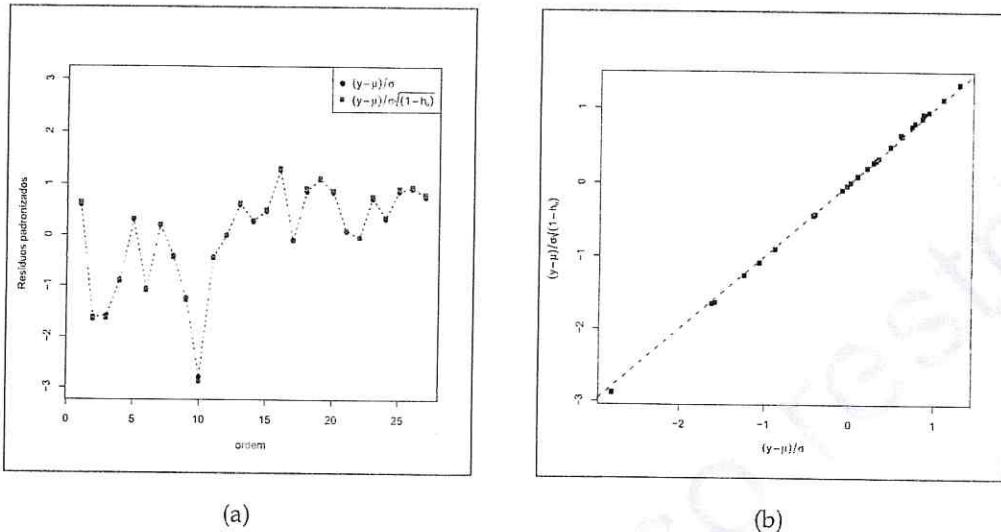


Figura 3.8: Comparação dos resíduos padronizados ajustados utilizando a estimativa do desvio padrão do erro e utilizando os elementos da diagonal da matriz H .

3.14 RESÍDUO DELETADO

O resíduo *deletado* é a diferença entre o i -ésimo valor observado, y_i , e a resposta de um modelo estimado sem a i -ésima observação, ou : $r_{(i)} = y_i - \hat{\mu}_{(i)}$. Ou seja $\hat{\mu}_{(i)}$ é a resposta de um modelo ajustado com $n - 1$ observações. É possível mostrar analiticamente que: $r_{(i)} = \frac{r_i}{1 - h_{ii}}$.

O resíduo deletado é também conhecido como do tipo *leave-one-out*.

Demonstração. Sejam definidos os seguintes elementos:

$$\mathbf{X} = \begin{bmatrix} a & d \\ b & e \\ c & f \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} b & e \end{bmatrix} \quad \mathbf{x}^T \mathbf{x} = \begin{bmatrix} b \\ e \end{bmatrix} \begin{bmatrix} b & e \end{bmatrix} = \begin{bmatrix} b^2 & be \\ be & e^2 \end{bmatrix}$$

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} a^2 + b^2 + c^2 & ad + be + cf \\ ad + be + cf & d^2 + e^2 + f^2 \end{bmatrix} \quad (\mathbf{X}^T \mathbf{X} - \mathbf{x}^T \mathbf{x}) = \begin{bmatrix} a^2 + c^2 & ad + cf \\ ad + cf & d^2 + f^2 \end{bmatrix}$$

Ou seja, $(\mathbf{X}^T \mathbf{X} - \mathbf{x}^T \mathbf{x})$ é a matriz $\mathbf{X}^T \mathbf{X}$ com a i -ésima linha removida.

Propriedade:

$$(\mathbf{X}^T \mathbf{X} - \mathbf{x}^T \mathbf{x})^{-1} = (\mathbf{X}^T \mathbf{X})^{-1} + \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T \mathbf{x} (\mathbf{X}^T \mathbf{X})^{-1}}{1 - \mathbf{x} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T}$$

(Demonstração)

$$\mathbf{I} = (\mathbf{X}^T \mathbf{X} - \mathbf{x}^T \mathbf{x})^{-1} (\mathbf{X}^T \mathbf{X} - \mathbf{x}^T \mathbf{x})$$

vamos assumir que a propriedade mostrada é verdadeira. Então:

$$\begin{aligned} & \left[(\mathbf{X}^T \mathbf{X})^{-1} + \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T \mathbf{x} (\mathbf{X}^T \mathbf{X})^{-1}}{1 - \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T} \right] (\mathbf{X}^T \mathbf{X} - \mathbf{x}^T \mathbf{x}) = \\ &= \mathbf{I} - (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T \mathbf{x} + \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T \mathbf{x}}{1 - \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T} - \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T \mathbf{x} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T \mathbf{x}}{1 - \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T} = \\ &= \mathbf{I} + \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T \mathbf{x} - (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T \mathbf{x} [1 - \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T] - (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T [\mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T] \mathbf{x}}{1 - \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T} = \\ &= \mathbf{I} + \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T \mathbf{x} [\mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T] - (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T \mathbf{x} [\mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T]}{1 - \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T} = \mathbf{I} \end{aligned}$$

Ou seja, a afirmação é verdadeira. Neste caso, é possível mostrar que: $x_i (\mathbf{X}^T \mathbf{X}) x_i^T = h_{ii}$. Então, segue que:

$$(\mathbf{X}^T \mathbf{X} - x_i^T x_i)^{-1} = (\mathbf{X}_{(i)}^T \mathbf{X}_{(i)})^{-1} = (\mathbf{X}^T \mathbf{X})^{-1} + \frac{(\mathbf{X}^T \mathbf{X})^{-1} x_i^T x_i (\mathbf{X}^T \mathbf{X})^{-1}}{1 - h_{ii}}$$

onde $\mathbf{H} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$.

Vamos considerar agora o vetor de parâmetros, β , estimado sem a i-ésima observação:

$$\hat{\beta}_{(i)} = (\mathbf{X}_{(i)}^T \mathbf{X}_{(i)})^{-1} \mathbf{X}_{(i)}^T y_{(i)}$$

o resíduo deletado é então definido como:

$$\begin{aligned} r_{(i)} &= y_i - \hat{y}_{(i)} \\ &= y_i - x_i \hat{\beta}_{(i)} \\ &= y_i - x_i (\mathbf{X}_{(i)}^T \mathbf{X}_{(i)})^{-1} \mathbf{X}_{(i)}^T y_{(i)} \\ &= y_i - x_i \left[(\mathbf{X}^T \mathbf{X})^{-1} + \frac{(\mathbf{X}^T \mathbf{X})^{-1} x_i^T x_i (\mathbf{X}^T \mathbf{X})^{-1}}{1 - h_{ii}} \right] \mathbf{X}_{(i)}^T y_{(i)} \\ &= y_i - x_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_{(i)}^T y_{(i)} - \frac{x_i (\mathbf{X}^T \mathbf{X})^{-1} x_i^T x_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_{(i)}^T y_{(i)}}{1 - h_{ii}} \\ &= y_i - x_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_{(i)}^T y_{(i)} \left[1 + \frac{h_{ii}}{1 - h_{ii}} \right] \\ &= y_i - x_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_{(i)}^T y_{(i)} \left[\frac{1}{1 - h_{ii}} \right] \\ &= \frac{(1 - h_{ii}) y_i - x_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_{(i)}^T y_{(i)}}{1 - h_{ii}} \end{aligned}$$

como: $\mathbf{X}^T y = \mathbf{X}_{(i)}^T y_{(i)} + x_i^T y_i \Rightarrow \mathbf{X}_{(i)}^T y_{(i)} = \mathbf{X}^T y - x_i^T y_i$. Então:

$$\begin{aligned} r_{(i)} &= \frac{(1-h_{ii})y_i - x_i(\mathbf{X}^T \mathbf{X})^{-1}(\mathbf{X}^T y - x_i^T y_i)}{1-h_{ii}} \\ &= \frac{(1-h_{ii})y_i x_i \hat{\beta} + h_{ii}y_i}{1-h_{ii}} \\ &= \frac{y_i - x_i^T \hat{\beta}}{1-h_{ii}} \\ &= \frac{r_i}{1-h_{ii}}. \end{aligned}$$

□

A variância do resíduo padronizado é dada por:

$$\begin{aligned} Var(r_{(i)}) &= Var\left(\frac{r_i}{1-h_{ii}}\right) \\ &= \frac{1}{(1-h_{ii})^2} [\sigma^2(1-h_{ii})] \\ &= \frac{\sigma^2}{1-h_{ii}} \end{aligned}$$

onde $i = 1, 2, \dots, n$.

3.15 RESÍDUO DELETADO: UMA DEMONSTRAÇÃO COMPUTACIONAL

A grande vantagem do resíduo deletado é que ele representa uma análise de validação cruzada do tipo *leave-one-out* sem a necessidade do ajuste de n modelos. É necessário apenas o ajuste de um único modelo com todas as observações. Essa característica é demonstrada a seguir em um código R.

Inicialmente, vamos calcular os resíduos deletados ajustando n modelos. Para cada modelo uma amostra é retirada da base de dados.

```
library(lasso2)
data(Prostate)
dados <- Prostate

res01 <- c()
for(cont in 1:dim(dados)[1]){
  modelo <- lm(lpsa ~ lcavol + lweight + sv1, data=dados[-cont,])
  res01[cont] <- dados[cont,"lpsa"] - predict(modelo, newdata=dados)[cont]
}
```

Agora, vamos utilizar a matriz de projeção para calcular os resíduos deletados utilizando somente o modelo ajustado com **todos** os dados.

```
modelo <- lm(lpsa ~ lcavol + lweight + sv1, data=dados)
X <- model.matrix(modelo)
H <- tcrossprod(X %*% solve(crossprod(X)), X)
res02 <- residuals(modelo) / (1-diag(H))
```

Uma comparação dos resíduos é mostrada na figura 3.9. O eixo x apresenta os resíduos calculados a partir do ajuste de n modelos e no eixo y apresenta os resíduos calculados a partir da matriz de projeção. A linha azul representa a reta da igualdade $y = x$, que demonstra que os resíduos são semelhantes, ou seja, iguais.

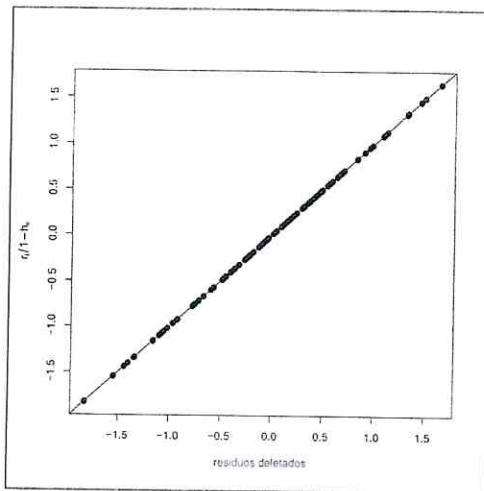


Figura 3.9: Comparação dos resíduos deletados utilizando os elementos da diagonal da matriz H a ajustando n modelos de regressão linear, para cada modelo uma observação é retirada.

3.16 A ESTATÍSTICA PRESS

A estatística PRESS (*Predicted Residuals Sum of Squares*) é definida como a soma dos quadrados dos resíduos preditivos ou $\text{PRESS} = \sum_i r_{(i)}^2$. Esta estatística representa a capacidade preditiva do modelo de regressão utilizando a metodologia de validação cruzada *leave-one-out*. Teoricamente a metodologia *leave-one-out* consiste em ajustar n modelos de regressão onde para cada modelo, a i -ésima observação é retirada da amostra e os resíduos são calculados a partir das diferenças entre o valor observado e a resposta do modelo ajustado sem a observação, ou seja, o resíduo preditivo. Como demonstrado anteriormente, esta abordagem não necessita do ajuste de n modelos e sim do ajuste de um modelo com todas as observações e então, utilizando os elementos da diagonal da matriz de projeção, os resíduos preditivos são calculados.

$$\begin{aligned}\text{PRESS} &= \sum_{i=1}^n [y_i - \hat{y}_{(i)}]^2 \\ &= \sum_{i=1}^n \left(\frac{r_i}{1 - h_{ii}} \right)^2\end{aligned}\tag{3.29}$$

A partir da estatística PRESS é possível calcular um coeficiente de determinação preditivo:

$$R_{\text{prediction}}^2 = 1 - \frac{\text{PRESS}}{SQT}$$

3.17 O ESTIMADOR RIDGE REGRESSION E A ESTATÍSTICA PRESS

Nem sempre o modelo *estatisticamente correto* representa o modelo com a melhor capacidade preditiva. Caso o analista deseje investigar ou melhorar a capacidade preditiva de um modelo ou mesmo manter várias variáveis em um modelo, mesmo não sendo significativas, é possível selecionar estimadores que agregar a característica preditiva ao ajuste. Uma alternativa é utilizar um estimador do tipo *Ridge Regression* descrito a partir do seguinte problema de minimização:

$$\tilde{\beta} = \arg \min_{\beta} \{(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)\} \quad (3.30)$$

$$\text{sujeito a: } \sum_{i=1}^p \beta_i^2 \leq c \text{ ou } \beta^T \beta \leq c \quad (3.31)$$

A solução é obtida utilizando multiplicadores de Lagrange e, então, definindo a seguinte função objetivo:

$$J(\beta, \lambda) = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda(\beta^T \beta - c)$$

onde $\lambda > 0$. O estimador é definido por:

$$\begin{aligned} \frac{\partial J(\beta, \lambda)}{\partial \beta} &= -2\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\beta) + 2\lambda\beta = 0 \\ 0 &= -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\beta + 2\lambda\beta \\ 0 &= -\mathbf{X}^T\mathbf{Y} + (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\beta \\ \tilde{\beta} &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \end{aligned} \quad (3.32)$$

Como pode ser visto na equação 3.32, o estimador depende do parâmetro λ que também precisa ser estimado.

3.17.1 PROPRIEDADES

1. $E(\tilde{\beta}) = E[(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}] = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\beta$,
ou seja, é um estimador viesado. Se $\lambda \rightarrow \infty : (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1} \rightarrow (\lambda\mathbf{I})^{-1} \rightarrow 0$ ou seja,
 $E(\tilde{\beta}) \underset{\lambda \rightarrow +\infty}{\approx} 0$.
2. $Var(\tilde{\beta}) \underset{\lambda \rightarrow +\infty}{\approx} 0$

$$\begin{aligned} Var(\tilde{\beta}) &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^TVar(\mathbf{Y})\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1} \end{aligned}$$
3. Matriz de projeção: $\mathbf{H}_{(\lambda)} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T$
4. Caso particular: Na prática, a restrição da soma dos quadrados dos parâmetros não deve ser aplicada ao intercepto do modelo, $\beta_0: \sum_{i=1}^p \beta_i^2 \leq c$. Esta restrição pode ser aplicada diretamente ao vetor β na forma: $\beta^T \mathbf{I}^* \beta \leq c$, onde \mathbf{I}^* é uma matriz identidade de dimensão $p \times p$ cujo primeiro elemento da diagonal é nulo: $[\mathbf{I}^*]_{1,1} = 0$. O estimador é então definido por

$$\tilde{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}^*)^{-1}\mathbf{X}^T\mathbf{Y}$$

Esta restrição se faz necessária pois garante que $\hat{y}_{(\lambda \rightarrow +\infty)} = \bar{y}$. Pois, quando $\lambda \rightarrow +\infty: \beta_i \rightarrow 0 \forall i \neq 0$. Nesta condição, tornar o intercepto *livre* da restrição faz com que o seu estimador seja a média amostral, $\tilde{\beta}_0 \underset{\lambda \rightarrow +\infty}{\approx} \bar{y}$.

Uma vez que o estimados de *ridge regression* é viciado, inferência estatística não é aconselhada. Entretanto, como o estimador passa a depender de um único parâmetro λ , significa que todos os coeficientes passam a ser *controlados* pelo λ . Na prática, caso o interesse no ajuste do modelo seja obter uma solução com a máxima capacidade preditiva, independente da quantidade de variáveis, pode-se então estimar o valor de λ que minimiza a estatística *PRESS*, por exemplo.

$$\hat{\lambda} = \arg \min_{\lambda} \sum_{i=1}^n \left(\frac{r_{i(\lambda)}}{1 - h_{ii(\lambda)}} \right)^2 \quad (3.33)$$

3.17.2 EXEMPLO DE APLICAÇÃO DO ESTIMADOR ridge NA SUAVIZAÇÃO DE MODELOS POLINOMIAIS

Em casos onde o número de parâmetros (ou variáveis) é elevado e a seleção das variáveis não é prioritária, é possível ajustar os estimadores de forma a maximizar a resposta preditiva do modelo. A figura 3.10 apresenta uma base de dados para a qual foi ajustado um modelo polinomial super-dimensional. O polinômio ajustado possui grau 40, $y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_{40} x^{40}$. Em virtude do super-dimensionamento do modelo linear a sua resposta apresenta um super-ajuste (*overfitting*) aos dados. O uso do estimador *ridge* permite a suavização da resposta do polinômio sem a redução do número de parâmetros lineares do mesmo.

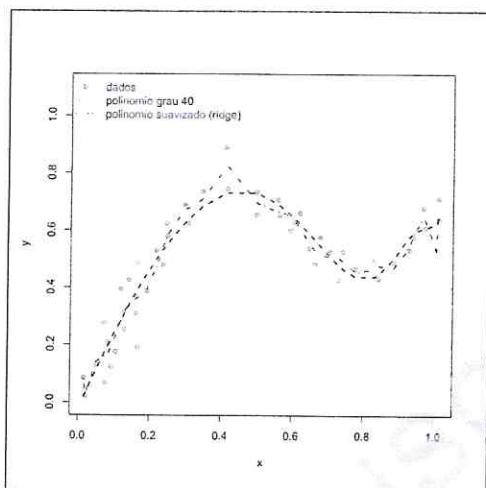


Figura 3.10: Exemplo do uso do estimador *ridge* na suavização de modelos polinomiais.

Portanto, o uso do estimador *ridge* é aconselhável quando não é prioritária a seleção de variáveis, mas é desejável um modelo com máxima capacidade preditiva. Nesses casos, o valor de λ pode ser estimado de forma a maximizar a estatística PRESS.

3.18 O ESTIMADOR LASSO E A ESTATÍSTICA PRESS

O estimador LASSO é definido a partir da minimização da soma dos quadrados dos resíduos sujeito à restrição de que a soma dos valores absolutos dos pesos seja limitado por um valor pré-fixado c .

$$\tilde{\beta} = \arg \min_{\beta} \{ (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta) \} \\ \text{sujeito a: } \sum_{i=1}^p |\beta_i| \leq c \quad (3.34)$$

Apesar da soma dos valores absolutos dos coeficientes representar uma restrição não-linear, a solução pode ser encontrada a partir de um sistema de restrições lineares na forma:

$$\tilde{\beta} = \arg \min_{\beta} \{ (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta) \} \\ \text{sujeito a: } \mathbf{C}\beta \leq \mathbf{D} \quad (3.35)$$

onde \mathbf{C} é uma matriz e \mathbf{D} é um vetor coluna. Lawson e Hanson (1995) sugerem modificar sequencialmente as matrizes \mathbf{C} e \mathbf{D} , iniciando com $\mathbf{D} = c$ e $\mathbf{C}^T = \text{sign}(\hat{\beta}_0)$, onde $\text{sign}(.)$ é a

função sinal e $\hat{\beta}_0$ é o estimador de mínimos quadrados sem restrição. Uma vez obtido o primeiro estimador $\tilde{\beta}$ com a restrição linear inicial, se a condição $\sum_{j=1} |\tilde{\beta}_j| > c$ for verificada então uma nova linha (i) deverá ser incluída nas matrizes C e D: $C_i^T = sign(\tilde{\beta})$ e $D = c\mathbf{1}$, $\mathbf{1} = (1, \dots, 1)^T$. Na sequência, uma nova solução é gerada, $\tilde{\beta}_{k+1}$. Este procedimento é repetido até que seja constatado que $\sum_{i=1}^p |\beta_i| - c \leq \xi$, onde ξ é um parâmetro de tolerância, ou o processo é interrompido após um número máximo de iterações (M).

O estimador associado à equação 3.35 é obtido da seguinte forma, utilizando multiplicadores de lagrange no problema de otimização com restrição linear, define-se a seguinte função objetivo,

$$J(\beta, \lambda) = (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta) + \lambda^T (\mathbf{C}\beta - \mathbf{D}) \quad (3.36)$$

cuja solução é:

$$\begin{aligned} \frac{\partial J}{\partial \beta} &= -\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\tilde{\beta}) + \mathbf{C}^T\lambda \\ 0 &= -\mathbf{X}^T\mathbf{Y} + \mathbf{X}^T\mathbf{X}\tilde{\beta} + \mathbf{C}^T\lambda \\ \tilde{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1} [\mathbf{X}^T\mathbf{Y} - \mathbf{C}^T\lambda] \end{aligned} \quad (3.37)$$

$$\begin{aligned} \frac{\partial J}{\partial \lambda} &= \mathbf{C}\tilde{\beta} - \mathbf{D} \\ 0 &= \mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1} [\mathbf{X}^T\mathbf{Y} - \mathbf{C}^T\lambda] - \mathbf{D} \\ \mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\lambda &= \mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} - \mathbf{D} \\ \lambda &= [\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1} [\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} - \mathbf{D}] \end{aligned} \quad (3.38)$$

uma vez que o estimador para λ também possui forma fechada e considerando $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, o estimador $\tilde{\beta}$ pode ser escrito como:

$$\begin{aligned} \tilde{\beta} &= \hat{\beta} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T [\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1} [\mathbf{C}\hat{\beta} - \mathbf{D}] \\ &= \hat{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T [\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1} [\mathbf{D} - \mathbf{C}\hat{\beta}] \\ &= (\mathbf{X}^T\mathbf{X})^{-1} [\mathbf{X}^T - \mathbf{C}^T [\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1} \mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T] \mathbf{Y} + \\ &\quad + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T [\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1} \mathbf{D} \end{aligned} \quad (3.39)$$

A partir da equação 3.39 o valor estimado $\hat{\mathbf{Y}}$ é definido como:

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{X}\tilde{\beta} \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} [\mathbf{X}^T - \mathbf{C}^T [\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1} \mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T] \mathbf{Y} + \\ &\quad + \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T [\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1} \mathbf{D} \end{aligned} \quad (3.40)$$

Neste caso, a matriz de projeção é: $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} [\mathbf{X}^T - \mathbf{C}^T [\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1} \mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]$. O valor estimado apresenta duas componentes: $\hat{\mathbf{Y}} = \mathbf{H}^T\mathbf{Y} + \mathbf{D}^*$, onde $\mathbf{D}^* = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T [\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1} \mathbf{D}$.

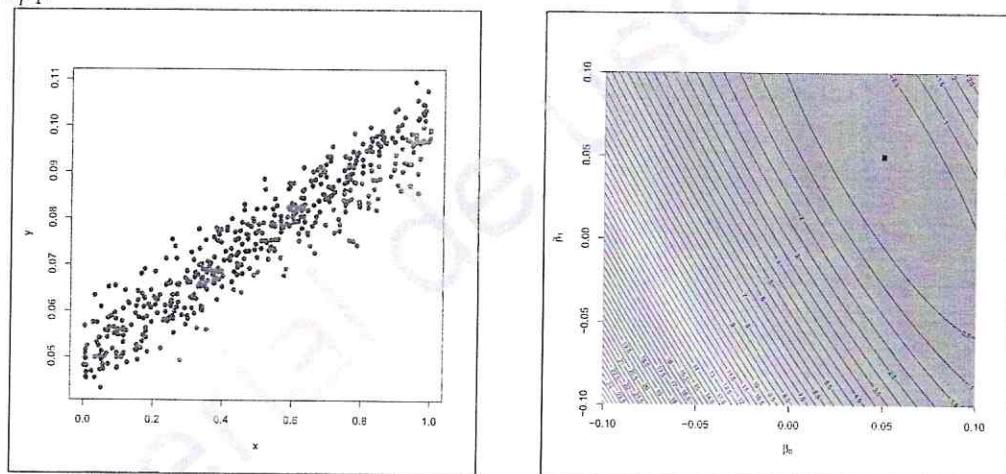
3.18.1 ESTIMANDO OS COEFICIENTES DO LASSO

Os valores possíveis para o parâmetro de restrição c estão contidos no intervalo $[0, \sum_{j \neq 0} |\hat{\beta}_j|]$, onde $\hat{\beta}$ é a solução de mínimos quadrados sem restrição. O valor ótimo para c pode ser selecionado, por exemplo, utilizando a estatística PRESS, uma vez que os resíduos deletados podem ser calculados utilizando os elementos da diagonal da matriz H .

3.19 COMPARAÇÃO DOS ESTIMADORES RIDGE E LASSO

A figura 3.11 apresenta um exemplo referente ao modelo de regressão linear simples ($y_i = \beta_0 + \beta_1 x_i + \epsilon_i$). A figura 3.11 (b) apresenta a superfície e as curvas referentes à soma dos quadrados dos erros para diferentes valores dos parâmetros β_0 e β_1 . O ponto marcado na figura 3.11 (b) representa a estimativa de mínimos quadrados. As curvas representam as isolíneas referentes à soma dos quadrados dos erros. Ou seja, todas as soluções que estejam sob a mesma linha possuem o mesmo erro. Somente o ponto possui erro mínimo.

A figura 3.12 (a) apresenta a sobreposição da superfície do erro com as isolíneas de restrição do tipo ridge. As isolíneas do método ridge são do tipo $\beta_0^2 + \beta_1^2 = c^2$. Esta equação representa um círculo de raio c . É evidente que existe um raio $c^* = \sqrt{\beta_0 + \beta_1}$ que atravessa a solução de mínimos quadrados. A solução que minimiza $\beta_0^2 + \beta_1^2$ é a origem ($\beta_0 = \beta_1 = 0$). Para valores do raio c menores que c^* a solução ridge representa a solução do círculo que intercepta a isolínea com o menor valor de erro. Nesses casos, a solução geralmente apresenta valores não nulos para β_0 e β_1 .



(a) Dados do exemplo.

(b) Soma dos quadrados dos erros para diferentes valores dos parâmetros.

Figura 3.11: Exemplo proposto para ilustrar as diferenças entre ridge e lasso.

A figura 3.12 (b) apresenta a sobreposição da superfície do erro com as isolíneas de restrição do tipo lasso. A restrição $|\beta_0| + |\beta_1| = c$ deve ser interpretada considerando as possíveis combinações dos sinais dos parâmetros β_0 e β_1 . Em geral a função valor absoluto de um número real é definido por:

$$|x| = \begin{cases} x, & \text{se } x \geq 0 \\ -x, & \text{se } x < 0 \end{cases} \quad (3.41)$$

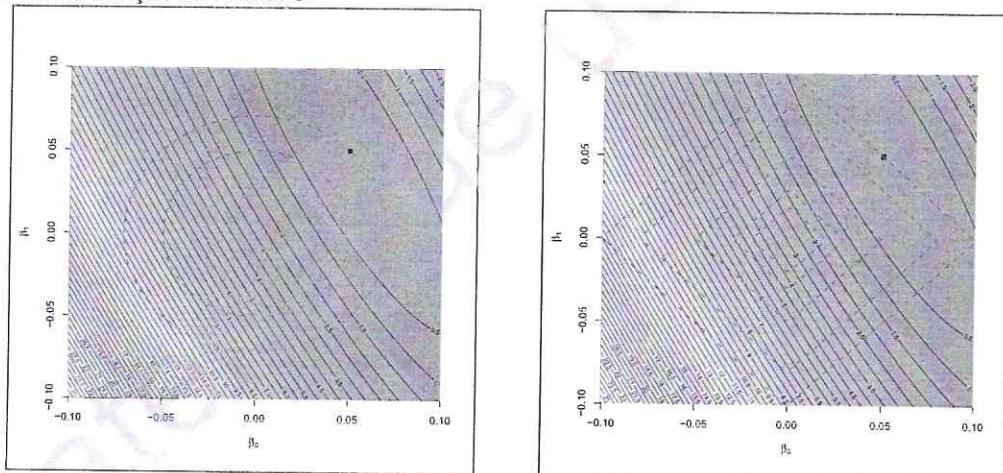
Portanto, a restrição $|\beta_0| + |\beta_1| = c$ pode ser escrita por um conjunto de restrições lineares que dependem dos possíveis cenários para os sinais de β_0 e β_1 :

1. se $\beta_0 > 0$ e $\beta_1 > 0$, então a restrição $|\beta_0| + |\beta_1| = c$ pode ser escrita como: $\beta_0 + \beta_1 = c$, ou na forma: $\beta_1 = c - \beta_0$.
2. se $\beta_0 > 0$ e $\beta_1 < 0$, então a restrição $|\beta_0| + |\beta_1| = c$ pode ser escrita como: $\beta_0 - \beta_1 = c$, ou na forma: $\beta_1 = -c + \beta_0$.
3. se $\beta_0 < 0$ e $\beta_1 > 0$, então a restrição $|\beta_0| + |\beta_1| = c$ pode ser escrita como: $-\beta_0 + \beta_1 = c$, ou na forma: $\beta_1 = c + \beta_0$.
4. se $\beta_0 < 0$ e $\beta_1 < 0$, então a restrição $|\beta_0| + |\beta_1| = c$ pode ser escrita como: $-\beta_0 - \beta_1 = c$, ou na forma: $\beta_1 = -c - \beta_0$.

Ou seja, a restrição do tipo lasso em um modelo com 2 parâmetros podem ser escrita utilizando $2^2 = 4$ restrições lineares e são formadas por equações lineares que se interceptam nos eixos ($\beta_0 = 0$ e $\beta_1 = 0$). Essas restrições lineares pode ser escritas na forma matricial, $C\beta = D$. Para o exemplo proposto: $\beta = [\beta_0, \beta_1]^T$, $D = [c, c, c, c]^T$ e

$$C = \begin{bmatrix} +1 & +1 \\ +1 & -1 \\ -1 & +1 \\ -1 & -1 \end{bmatrix}$$

Como a restrição do tipo lasso apresenta arestas é possível que para valores reduzidos da constante de restrição c a solução de mínimo erro esteja exatamente na aresta. Esta solução possui, pelo menos, um dos estimadores iguais a zero. Portanto, o método lasso é comumente preferível em relação ao método ridge pois permite, simultaneamente, o ajuste da capacidade preditiva do modelo e a seleção de variáveis.



(a) Curvas de restrição do método ridge.

(b) Curvas de restrição do método lasso.

Figura 3.12: Comparação das curvas de restrição para os métodos ridge e lasso.

Em ambos os casos, ridge e lasso, os valores ótimos dos parâmetros de restrição (λ para ridge e c para lasso) podem ser obtidos a partir da estatística PRESS.

3.20 AJUSTE DE CURVAS UTILIZANDO MÍNIMOS QUADRADOS PONDERADOS

Embora seja incomum, o modelo de regressão linear pode ser utilizado para o ajuste de modelos não lineares utilizando aproximações lineares locais. Nesses casos, é possível utilizando o estimador de mínimos quadrados ponderados. O estimador de mínimos quadrados ponderados pode ser obtido a partir da minimização da soma ponderada dos quadrados dos erros:

$$\tilde{\beta} = \arg \min_{\beta} \sum_{i=1}^n w_i (y_i - x_i \beta)^2 \quad (3.42)$$

também representada na forma matricial,

$$\tilde{\beta} = \arg \min_{\beta} (\mathbf{Y} - \mathbf{X}\beta)^T \mathbf{W} (\mathbf{Y} - \mathbf{X}\beta) \quad (3.43)$$

onde \mathbf{W} é uma matriz diagonal e $w_i = [\mathbf{W}]_{ii}$ é o peso atribuído a cada observação. A solução é obtida a partir da derivada de primeira ordem da soma ponderada dos quadrados dos erros:

$$\begin{aligned} \frac{\partial (\mathbf{Y} - \mathbf{X}\beta)^T \mathbf{W} (\mathbf{Y} - \mathbf{X}\beta)}{\partial \beta} &= 0 \\ -2\mathbf{X}^T \mathbf{W} (\mathbf{Y} - \mathbf{X}\beta) &= 0 \\ -2\mathbf{X}^T \mathbf{W} \mathbf{Y} + 2\mathbf{X}^T \mathbf{W} \mathbf{X} \tilde{\beta} &= 0 \\ \mathbf{X}^T \mathbf{W} \mathbf{X} \tilde{\beta} &= \mathbf{X}^T \mathbf{W} \mathbf{Y} \\ \tilde{\beta} &= (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y} \end{aligned}$$

3.20.1 PROPRIEDADES DO ESTIMADOR DE MÍNIMOS QUADRADOS PONDERADOS

Utilizando as propriedades de esperança e covariância, já demonstradas em seções anteriores, é possível mostrar que:

1. $E(\tilde{\beta}) = \beta$:

$$\begin{aligned} E(\tilde{\beta}) &= E((\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}) \\ &= (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} E(\mathbf{Y}) \\ &= (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{X} \beta \\ &= \beta \end{aligned} \quad (3.44)$$

2. $Var(\tilde{\beta}) = \sigma^2 (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{W}^T \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$

3. Matriz de projeção: $\mathbf{H} = \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}$

Ou seja, o estimador de mínimos quadrados ponderados é não viciado, mas apresenta uma matriz de covariância dos estimadores que depende dos pesos atribuídos às observações. Na prática, é incomum o uso de suas propriedades estatísticas para inferência nos parâmetros do modelo.

3.20.2 APLICAÇÃO DO ESTIMADOR DE MÍNIMOS QUADRADOS PONDERADOS

O estimador de mínimos quadrados ponderados apresenta grande potencial no ajuste de modelos de regressão não lineares. Considere o problema de estimação de um modelo para a base de dados mostrada na figura 3.13 (a). Claramente os dados não se comportam segundo um modelo linear. Entretanto, é possível supor que *localmente*, ou *pontualmente*, a resposta possa ser aproximada por uma reta, como é ilustrado na figura 3.13 (b).

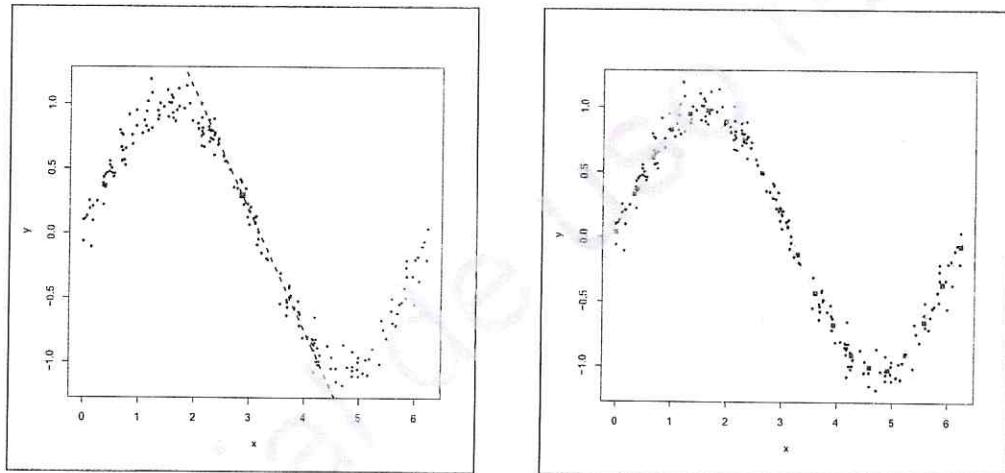
Seja x_0 um valor não-observado ao qual deseja-se estimar a resposta \hat{y} . É possível assumir que os pontos próximos de x_0 contenham informações sobre o comportamento no ponto de interesse. Um ponto importante é definir a *vizinhança* de x_0 . Utilizar o núcleo de uma distribuição normal pode ser uma alternativa para definir pesos ou o *grau* de vizinhança em relação ao ponto x_0 .

$$w_{(x_i, x_0)} = e^{-\frac{(x_i - x_0)^2}{\lambda}} \quad (3.45)$$

onde λ é um parâmetro que regular a amplitude da ponderação. Para pontos próximos de x_0 , os valores de $w_{(x_i, x_0)}$ se aproximam de um, ao passo que pontos distantes de x_0 possuem valores de $w_{(x_i, x_0)}$ próximos de zero. A velocidade de queda dos pesos é controlada pelo parâmetro λ que, neste caso, é comumente chamado de *largura de banda (bandwidth)*.

Os valores dos pesos ou ponderações podem ser utilizar para construir uma matriz diagonal de pesos, W , onde $w_{ii} = w_{(x_i, x_0)}$. O modelo de regressão ponderado é então aplicado para estimar $\hat{y}|x_0$.

Uma limitação desta abordagem é a necessidade do ajuste de um modelo para cada ponto a ser estimado. Ou seja, os valores de intercepto e inclinação do modelo passam a depender de x_0 ($\beta_0^{(x_0)}, \beta_1^{(x_0)}$). A figura 3.13 (b) ilustra o ajuste do modelo. Para cada ponto a ser estimado, indicado pelos pontos quadrados, é ajustado um modelo de regressão ponderado. Os modelos lineares ajustados são apresentados e representam as retas tangentes aos pontos.



(a) Aproximação local de um modelo de regressão linear ponderado. (b) Múltiplas soluções utilizando um modelo de regressão linear simples ponderado.

Figura 3.13: Aplicação do estimador de mínimos quadrados ponderados na aproximação de uma função não-linear utilizando o modelo de regressão linear simples.

3.20.3 EXEMPLO DE AJUSTE DE MÍNIMOS QUADRADOS PONDERADOS UTILIZANDO O R

A função de ajuste de modelo linear *lm()* do R possui um argumento para a passagem de pesos para o modelo, conforme indicado a seguir:

```
pesos <- dnorm((x0 - dados$x)/lambda)
pesos <- pesos/max(pesos)
modelo <- lm(y~x, data=dados, weights=pesos)
```

No exemplo, os pesos foram definidos utilizando a função densidade de uma normal, sendo os pesos padronizados com relação ao valor máximo. Na prática, o estimador é insensível à padronização dos pesos.

3.21 MÍNIMOS QUADRADOS NÃO LINEARES

Alguns dos princípios fundamentais de estimativa e inferência estatística para o modelo de regressão linear múltipla podem ser aplicados na estimativa de alguns modelos não-lineares. Por exemplo, seja o seguinte modelo de regressão não-linear:

$$y = \beta_0 e^{\beta_1 x_i} + \varepsilon$$

onde $\varepsilon \sim Normal(0, \sigma^2)$, ou na forma $y = f(x, \beta) + \varepsilon$. É possível perceber que o modelo é não linear com relação ao parâmetro β_1 . Queremos estimar os parâmetros do modelo a partir da minimização da soma dos quadrados dos erros, $SQE(\beta) = \sum_{i=1}^n (y_i - \beta_0 e^{\beta_1 x_i})^2$.

Derivando a função objetivo com relação aos parâmetros β_0 e β_1 , desejamos encontrar a solução para o seguinte sistema de equações não lineares:

$$\begin{aligned}\frac{\partial SQE}{\partial \beta_0} &= \sum_{i=1}^n (y_i - \hat{\beta}_0 e^{\hat{\beta}_1 x_i})(-1)(e^{\hat{\beta}_1 x_i}) = 0 \\ \frac{\partial SQE}{\partial \beta_1} &= \sum_{i=1}^n (y_i - \hat{\beta}_0 e^{\hat{\beta}_1 x_i})(-\hat{\beta}_0 e^{\hat{\beta}_1 x_i}) = 0\end{aligned}$$

A princípio, uma função não linear $f(\beta)$ pode ser representada de forma alternativa utilizando uma expansão em série de Taylor de primeira ordem, do tipo:

$$f(\beta) = f(\beta^{(0)}) + f'(\beta^{(0)})(\beta - \beta^{(0)})$$

Aplicando este mesmo princípio à nossa função não linear, dado um vetor preditor x_i , a seguinte expansão em série de Taylor de primeira ordem é utilizada:

$$f(x_i, \beta) = f(x_i, \beta^{(0)}) + \sum_{j=1}^p \left. \left(\frac{\partial f(x_i, \beta)}{\partial \beta_j} \right) \right|_{\beta=\beta^{(0)}} (\beta_j - \beta_j^{(0)}) \quad (3.46)$$

onde x_i é a i-ésima linha da matriz X . Outra representação para a Equação 3.46 é:

$$f(x_i, \beta) = f(x_i, \beta^{(0)}) + J_i^{(0)}(\beta - \beta^{(0)})$$

$$\text{onde } J_{ij} = \left. \frac{\partial f(x_i, \beta)}{\partial \beta_j} \right|_{\beta=\beta^{(0)}} (\beta_j - \beta_j^{(0)})$$

A representação matricial pode ser definida como:

$$\begin{aligned}\mathbf{f} &= \mathbf{f}(\beta^{(0)}) + \mathbf{J}^{(0)}(\beta - \beta^{(0)}) \\ &= \mathbf{f}^{(0)} + \mathbf{J}^{(0)}(\beta - \beta^{(0)})\end{aligned}$$

Como consequência, a Soma dos Quadrados dos Erros na forma matricial pode ser escrita como:

$$\begin{aligned}SQE(\beta) &= \|\mathbf{Y} - [\mathbf{f}^{(0)} + \mathbf{J}^{(0)}(\beta - \beta^{(0)})]\|^2 \\ &= \|\mathbf{Y} - \mathbf{f}^{(0)} + \mathbf{J}^{(0)}\beta^{(0)} - \mathbf{J}^{(0)}\beta\|^2\end{aligned}$$

no exemplo $\beta = [\beta_0, \beta_1]$.

A partir deste ponto, duas representações são possíveis:

1. Seja o pseudo vetor $\mathbf{Z}^{(0)}$, definido como $\mathbf{Z}^{(0)} = \mathbf{Y} - \mathbf{f}^{(0)} + \mathbf{J}^{(0)}\beta^{(0)}$, então a Soma dos Quadrados dos Erros pode ser escrita como:

$$\begin{aligned} SQE(\beta) &= \|\mathbf{Z}^{(0)} - \mathbf{J}^{(0)}\beta\|^2 \\ &= (\mathbf{Z}^{(0)} - \mathbf{J}^{(0)}\beta)^T(\mathbf{Z}^{(0)} - \mathbf{J}^{(0)}\beta) \end{aligned}$$

que representa a equação de mínimos quadrados lineares, cuja solução é:

$$\hat{\beta} = ([\mathbf{J}^{(0)}]^T[\mathbf{J}^{(0)}])^{-1}[\mathbf{J}^{(0)}]^T\mathbf{Z}^{(0)} \quad (3.47)$$

Sob certas condições, pode-se criar um processo iterativo:

$$\beta^{(k+1)} = ([\mathbf{J}^{(k)}]^T[\mathbf{J}^{(k)}])^{-1}[\mathbf{J}^{(k)}]^T\mathbf{Z}^{(k)} \quad (3.48)$$

2. De forma alternativa, a Soma de Quadrados dos Erros pode ser reagrupada como:

$$\begin{aligned} SQE(\beta) &= \|\mathbf{Y} - [\mathbf{f}^{(0)} + \mathbf{J}^{(0)}(\beta - \beta^{(0)})]\|^2 \\ &= \|(\mathbf{Y} - \mathbf{f}^{(0)}) - \mathbf{J}^{(0)}\Delta\beta\|^2 \\ &= \|\mathbf{Y}^{(0)} - \mathbf{J}^{(0)}\Delta\beta\|^2 \end{aligned} \quad (3.49)$$

onde $\Delta\beta = (\beta - \beta^{(0)})$ e $\mathbf{Y}^{(0)} = (\mathbf{Y} - \mathbf{f}^{(0)})$. A solução de mínimos quadrados lineares, neste caso, é:

$$\Delta\beta = ([\mathbf{J}^{(0)}]^T[\mathbf{J}^{(0)}])^{-1}[\mathbf{J}^{(0)}]^T\mathbf{Y}^{(0)} \quad (3.50)$$

ou na forma:

$$\beta = \beta^{(0)} + ([\mathbf{J}^{(0)}]^T[\mathbf{J}^{(0)}])^{-1}[\mathbf{J}^{(0)}]^T\mathbf{Y}^{(0)} \quad (3.51)$$

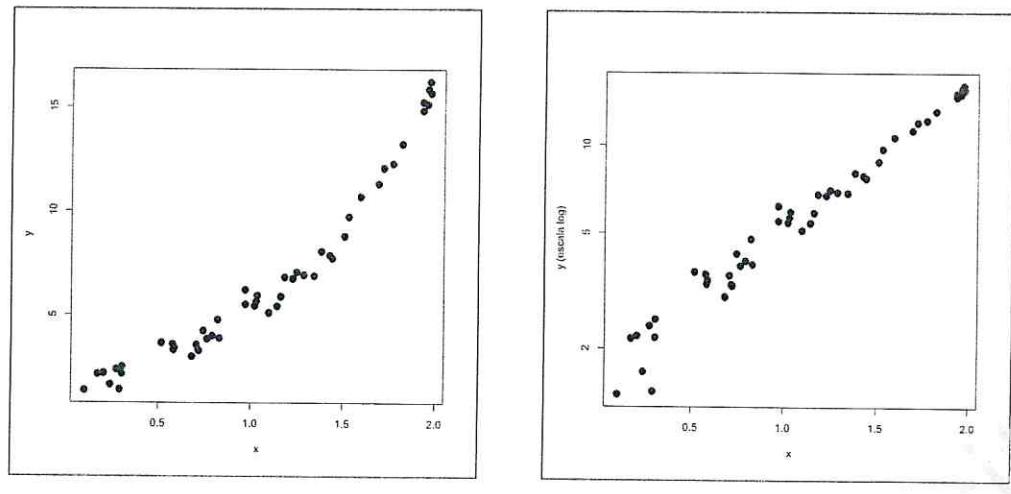
Sob certas condições, pode-se criar um processo iterativo:

$$\beta^{(k+1)} = \beta^{(k)} + ([\mathbf{J}^{(k)}]^T[\mathbf{J}^{(k)}])^{-1}[\mathbf{J}^{(k)}]^T\mathbf{Y}^{(k)} \quad (3.52)$$

O algoritmo de estimação dos parâmetros consiste em atualizar os estimadores utilizando a Equação 3.48 ou 3.52, até que seja verificada convergência. Ou seja, $\|\beta^{(k+1)} - \beta^{(k)}\| \leq \xi$, onde ξ é um nível de tolerância especificado pelo usuário. É válido lembrar que a convergência do algoritmo está associada às condições iniciais do algoritmo, ou $\beta^{(0)}$.

3.21.1 EXEMPLO DO USO DE MÍNIMOS QUADRADOS NÃO LINEARES

A figura 3.14 (a) apresenta uma base de dados na qual a função geradora é pré-conhecida como $y_i = \theta_1 e^{-\theta_2 x_i} + \epsilon_i$. A resposta pode ser linearizada utilizando uma transformação logarítmica em y , conforme apresentado na figura 3.14 (b). No entanto, na escala logarítmica os erros não são homocedásticos, como pode ser visto na figura 3.14 (b). Como consequência, qualquer inferência estatística é inconsistente. Neste caso, a solução é utilizar o ajuste de mínimos quadrados não lineares (MQNL). A abordagem MQNL requer condições iniciais para os parâmetros. Os valores iniciais podem ser gerados pelo modelo *linearizável*: $\log y_i = \log \theta_1 + \theta_2 x_i$.



(a) Base de dados original.

(b) Base de dados após transformação $\log(Y)$

Figura 3.14: Base de dados simulada.

Os resultados do ajuste utilizando o modelo linearizado e o modelo não linear são apresentados a seguir. O modelo linear estima $\log \theta_1$, o modelo não linear estima θ_1 . Com relação ao estimador $\hat{\theta}_2$, o modelo linear apresenta $\hat{\theta}_2 = 1.18688$ e o modelo não linear, $\hat{\theta}_2 = 1.18316$. Apesar de muito próximos, a estimativa do modelo não linear é robusta pois não viola as suposições de homocedasticidade dos resíduos. Mesmo utilizando o modelo de regressão não linear, é possível obter intervalos de confiança para os estimados utilizando a matriz de covariância $Cov(\hat{\beta}) = \sigma^2 ([J^{(k)}]^T [J^{(k)}])^{-1}$. Neste caso, após a convergência, aplicam-se as mesmas propriedades dos estimadores de mínimos quadrados para o modelo de regressão múltipla.

```
> modelo <- lm(log(y) ~ x)
> summary
  Call:
  lm(formula = log(y) ~ x)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.43002    0.03629   11.85 7.33e-16 ***
x           1.18688    0.03270   36.29 < 2e-16 ***
---
Residual standard error: 0.1242 on 48 degrees of freedom
Multiple R-squared: 0.9648,      Adjusted R-squared: 0.9641
F-statistic: 1317 on 1 and 48 DF,  p-value: < 2.2e-16
```

```
> modelo <- nls(y ~ theta1*exp(theta2 * x),
+ start=list(theta1 = exp(modelo$coef[1]),
+            theta2 = modelo$coef[2] ) )
> summary(modelo)

Formula: y ~ theta1 * exp(theta2 * x)

Parameters:
            Estimate Std. Error t value Pr(>|t|)
theta1  1.54813    0.05398   28.68 <2e-16 ***
theta2  1.18316    0.02070   57.17 <2e-16 ***
---
Residual standard error: 0.4689 on 48 degrees of freedom
```

```
Number of iterations to convergence: 2
Achieved convergence tolerance: 8.354e-07
```

```
> confint(modelo)
Waiting for profiling to be done...
      2.5%    97.5%
theta1 1.441200 1.658952
theta2 1.141884 1.225427
```

3.22 EXERCÍCIOS

Exercício 3.1

Suponha o modelo de regressão $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, onde os parâmetros β estão sujeitos a restrições de igualdade do tipo $\mathbf{C}\beta = \mathbf{d}$. Mostre que a estimativa de mínimos quadrados de β é dada por:

$$\tilde{\beta} = \hat{\beta} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T \left(\mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T \right)^{-1} (\mathbf{d} - \mathbf{C}\hat{\beta})$$

- (a) Para o modelo acima, encontre: $E(\tilde{\beta})$ e $Var(\tilde{\beta})$.

Exercício 3.2

A figura 3.15 apresenta os dados referentes ao seguinte modelo teórico:

$$y_i = \theta_0 + \theta_1 \cdot e^{\theta_2 \cdot x_i} + \epsilon_i$$

onde ϵ_i é uma variável aleatória normal com média zero e variância σ^2 . Utilize o método de mínimos quadrados não-lineares para estimar todos os parâmetros do modelo. Inclua na sua análise intervalos de confiança para os parâmetros. Os dados foram gerados utilizando o seguinte código:

```
n      ← 200
x      ← runif(n)
y      ← 0.9 + 1.4*exp(-2.5*x) + rnorm(n, sd=0.01)
```

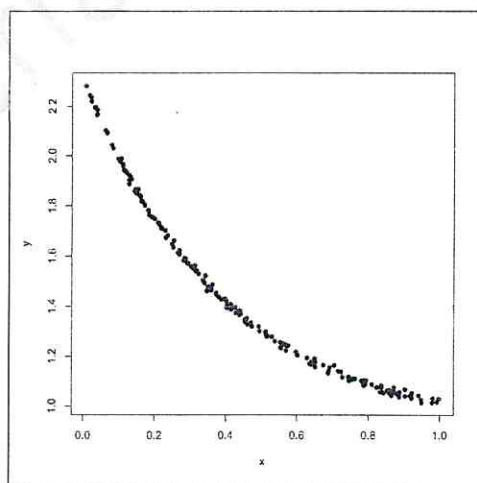


Figura 3.15: Ajuste de mínimos quadrados não lineares.