UNIVERSIDADE FEDERAL DO PARANÁ
DEPARTAMENTO DE ESTATÍSTICA

ANÁLISE DE REGRESSÃO LINEAR

Suely Ruiz Giolo

C U R I T I B A
Estado do Paraná – Brasil
2007

1. Introdução

A análise de regressão linear múltipla pode ser vista como uma extensão da análise de regressão linear simples, que envolve somente uma variável independente, para a situação em que se tem um conjunto de variáveis independentes.

Tratar com diversas variáveis independentes simultaneamente em uma análise de regressão é, em geral, mais complexo do que com uma única variável independente por algumas das seguintes razões:

- é mais difícil escolher um *bom* modelo;
- é impossível visualizar o modelo ajustado na presença de mais do que duas variáveis independentes;
- é algumas vezes difícil interpretar as estimativas dos parâmetros do modelo de regressão escolhido.

Em geral, o objetivo ao se ajustar um modelo de regressão linear múltipla é o de predizer a variável resposta por meio das variáveis independentes. Em diversas situações, se concluirá que as predições da variável resposta realizadas por meio de uma única variável independente são muito imprecisas e, sendo assim, modelos com mais de uma variável independente serão analisados para melhorar tais predições. Dentre um conjunto possível de variáveis independentes, a intenção é escolher um subconjunto que produza um *bom* modelo, isto é, um modelo parcimonioso que forneça estimativas precisas da variável resposta e que faça sentido prático, uma vez que nem sempre o modelo escolhido, em termos estatísticos, é aplicável ou faz sentido na prática. O conhecimento do problema sob análise e a interação com o pesquisador são imprescindíveis para a escolha do modelo final.

Formalmente, para o ajuste de um modelo de regressão linear múltipla, a variável resposta, bem como as variáveis independentes, deveriam ser contínuas. Na prática, contudo, as variáveis independentes (regressoras) podem ser de qualquer outra natureza. As categóricas, por exemplo, são incorporadas ao modelo por meio de variáveis indicadoras *(dummy)*.

1.1 Modelo de regressão linear múltipla (MRLM)

Considerando duas variáveis independentes (regressoras) X_1 e X_2, o modelo de regressão linear múltipla é expresso por $Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ e é chamado *modelo de primeira ordem* por ser linear nos parâmetros e nas regressoras.

Assumindo que $E(\varepsilon) = 0$ tem-se, para valores fixos $\mathbf{x} = (x_1, x_2)$ de $\mathbf{X} = (X_1, X_2)$, que $E(Y \mid \mathbf{x}) = \beta_o + \beta_1 x_1 + \beta_2 x_2$, o qual geometricamente descreve um *plano* (superfície de resposta). A cada ponto nesse *plano* tem-se a esperança de Y, isto é $E(Y \mid \mathbf{x})$, em uma dada combinação dos níveis de X_1 e X_2.

Um outro exemplo de modelo de regressão linear múltipla é dado por qualquer polinomial de ordem ≥ 2. Logo, em um MRLM, o termo *linear* refere-se à linearidade dos parâmetros e não das regressoras.

1.1.1 Interpretação dos parâmetros na ausência de interações

Considere o modelo $E(Y|\ \mathbf{x}) = \beta_o + \beta_1 x_1 + \beta_2 x_2$.

- O parâmetro β_o é o intercepto do plano da regressão. Se a extensão do modelo incluir o ponto $\mathbf{x} = (x_1, x_2) = (0, 0)$, o parâmetro β_o fornece a resposta esperada nesse ponto. Caso contrário, não possui qualquer significado como um termo isolado no modelo de regressão.
- O parâmetro β_1 indica a mudança ocorrida na esperança de Y a cada unidade de mudança em X_1 *quando X_2 é mantida fixa*.
- Similarmente β_2 indica a mudança ocorrida na esperança de Y a cada unidade de mudança em X_2 *quando X_1 é mantida fixa*.

Exemplo: Considere o modelo $E(Y|\ \mathbf{x}) = 20 + 0{,}95x_1 - 0{,}5x_2$ e suponha que X_2 é mantida fixa em $x_2 = 20$ de modo que $E(Y \mid x_2 = 20) = 10 + 0{,}95x_1$. Então, $\beta_1 = 0{,}95$ indica, a cada acréscimo de uma unidade em X_1, um acréscimo esperado em Y de 0,95 unidades, desde que o valor de X_2 seja mantido fixo em 20. O mesmo é verdadeiro para qualquer outro valor fixo de X_2. Similarmente, $\beta_2 = -0{,}5$ indica que o decréscimo esperado em Y é de 0,5 unidades a cada acréscimo de uma unidade em X_2, desde que o valor de X_1 permaneça fixo.

Os parâmetros β_1 e β_2 são usualmente denominados *coeficientes de regressão parciais* por refletirem o efeito parcial de uma variável independente quando a outra variável é incluída no modelo e mantida fixa.

De modo geral, a resposta Y pode estar associada a *k* regressoras $X_1, ..., X_k$ e, sendo assim, tem-se, para $\mathbf{X} = \mathbf{x}$ que:

$$\begin{aligned} Y &= \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_k + \varepsilon \\ &= \beta_o + \sum_{j=1}^{k} \beta_j x_j + \varepsilon. \end{aligned}$$

Assumindo-se que $E(\varepsilon) = 0$, segue que: $\boxed{E(Y|\ \mathbf{x}) = \beta_o + \sum_{j=1}^{k} \beta_j x_j}$,

que descreve um *hiperplano* (nem sempre possível de ser visualizado) no espaço *k*-dimensional das variáveis regressoras X_j ($j = 1, 2, ..., k$).

A interpretação dos parâmetros é análoga ao caso de duas regressoras, ou seja, o parâmetro β_j ($j = 1, 2, ..., k$) indica a mudança esperada em Y a cada acréscimo de uma unidade em X_j, mantendo fixas as demais regressoras.

Obs: Os modelos de regressão de primeira ordem apresentados mostram regressoras cujos efeitos na esperança de Y são aditivos e, portanto, não interagem. Considerando $k = 2$, isto significa que o efeito de X_1 na esperança de Y não depende dos níveis de X_2 e, analogamente, o efeito de X_2 não depende dos níveis de X_1. As duas regressoras, nesse caso, são ditas apresentarem *efeitos aditivos* ou *não interagirem*.

1.2 Efeito da interação de regressoras

Considere, agora, o modelo de regressão linear com duas regressoras X_1 e X_2 dado por $Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$ ou $E(Y|\ \mathbf{x}) = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$, se for assumido que $E(\varepsilon) = 0$. No modelo citado, $x_1 x_2$ representa a interação entre as regressoras X_1 e X_2.

Se a interação for significativa, então o efeito de X_1 na esperança de Y depende do nível de X_2 e, analogamente, o efeito de X_2 na esperança de Y depende do nível de X_1. Assim:

- quando X_2 for mantida fixa tem-se, a cada unidade de mudança em X_1, que a mudança ocorrida na esperança de Y é de $\beta_1 + \beta_3 x_2$ unidades e, similarmente,
- quando X_1 for mantida fixa tem-se, a cada unidade de mudança em X_2, uma mudança na esperança de Y de $\beta_2 + \beta_3 x_1$ unidades.

1.3 Representação matricial do modelo de regressão linear geral

Para uma amostra de tamanho *n*, em que para valores prefixados de **X** tem-se observações independentes de **Y**, a representação matricial do modelo de regressão linear geral é dada por:

$$\boxed{\mathbf{Y}_{n\,x\,1} = \mathbf{X}_{n\,x\,p}\ \boldsymbol{\beta}_{p\,x\,1} + \boldsymbol{\varepsilon}_{n\,x\,1}}$$

em que: $\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{bmatrix}$ $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}$ $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_k \end{bmatrix}$ $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{bmatrix}$,

com k o número de regressoras, p = k + 1 o número de parâmetros, **Y** o vetor associado à variável resposta, **X** uma matriz de constantes, $\boldsymbol{\beta}$ o vetor de p parâmetros desconhecidos e $\boldsymbol{\varepsilon}$ o vetor de erros tal que $\boldsymbol{\varepsilon} \sim$ Normal com $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ e matriz de variância-covariância $\boldsymbol{\Sigma}(\boldsymbol{\varepsilon}) = \sigma^2\,\mathbf{I}$.

De acordo com o modelo e suposições apresentadas, segue que $\mathbf{Y} \sim$ Normal $(\mathbf{X}\boldsymbol{\beta};\ \sigma^2\mathbf{I})$, visto que $E(\mathbf{Y} \mid \mathbf{x}) = \mathbf{X}\boldsymbol{\beta}$ e $\boldsymbol{\Sigma}(\mathbf{Y}) = \sigma^2\,\mathbf{I}$.

1.4 Estimação dos parâmetros por Mínimos Quadrados Ordinários

Para a obtenção dos parâmetros por MQO faz-se necessário minimizar a soma de quadrados dos erros, isto é, minimizar:

$$\text{SQerros} = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \varepsilon_1^{\,2} + \varepsilon_2^{\,2} + ... + \varepsilon_n^{\,2} = [\, \varepsilon_1 \;\; \varepsilon_2 \;\; ..... \;\; \varepsilon_n \,] \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{bmatrix} = \boldsymbol{\varepsilon' \varepsilon}.$$

Como $\mathbf{Y = X\beta + \varepsilon}$, segue que $\boldsymbol{\varepsilon = Y - X\beta}$ e, desse modo, tem-se:

$$\text{SQerros} = \boldsymbol{\varepsilon' \varepsilon = (Y - X\beta)'(Y - X\beta) = Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta}$$
$$\boldsymbol{= Y'Y - 2\,\beta'X'Y + \beta'X'X\beta},$$

pois $\mathbf{Y'X\beta}$ é um escalar, bem como $\boldsymbol{\beta'X'Y = (Y'X\beta)'}$ também o é. Assim, $\mathbf{Y'X\beta = \beta'X'Y}$. Derivando, então, a SQerros em relação a $\boldsymbol{\beta}$ obtém-se:

$$\partial \text{SQerros} / \partial \boldsymbol{\beta} = -\,\mathbf{2\,X'Y + 2\,X'X\,\beta}.$$

Avaliando a expressão resultante em $\hat{\boldsymbol{\beta}}$, o vetor que anula a derivada, segue que:

$$-\,\mathbf{2\,X'Y + 2\,X'X\hat{\beta} = 0}$$
$$\mathbf{X'X\hat{\beta} = X'Y}$$
$$\boldsymbol{\hat{\beta} = (X'X)^{-1}X'Y},$$

desde que **(X'X)** seja inversível, o que ocorre quando as regressoras forem linearmente independentes.

Para o modelo $\mathbf{Y = X\beta + \varepsilon}$, o estimador de MQO $\hat{\boldsymbol{\beta}}$, também é estimador de máxima verossimilhança.

1.5 Valores ajustados ou preditos e resíduos

O vetor de valores ajustados $\hat{Y}_i$ será denotado por $\hat{\boldsymbol{Y}}$ e o vetor dos termos residuais $e_i = Y_i - \hat{Y}_i$ por **e**, de modo que:

$$\hat{\boldsymbol{Y}}_{n\,x\,1} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \cdots \\ \hat{Y}_n \end{bmatrix} \quad \text{e} \quad \mathbf{e}_{n\,x\,1} = \begin{bmatrix} e_1 \\ e_2 \\ .. \\ e_n \end{bmatrix}.$$

Em termos matriciais, o vetor $\hat{\boldsymbol{Y}}$ pode então ser representado por:

$\hat{\boldsymbol{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ (como $\hat{\boldsymbol{\beta}} = \mathbf{(X'X)^{-1}\,X'Y}$)

$\hat{\boldsymbol{Y}} = \mathbf{X\,(X'X)^{-1}\,X'Y}$ (fazendo $\mathbf{H = X\,(X'X)^{-1}\,X'}$)

$\hat{\boldsymbol{Y}} = \mathbf{H\,Y}$

↘ **conhecida por matriz "chapéu"**

e, os resíduos, por sua vez, representados por:

$\mathbf{e} = \mathbf{Y} - \hat{\boldsymbol{Y}}$ (como $\hat{\boldsymbol{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$)

$\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ (como $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\,\mathbf{Y}$)

$\mathbf{e} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$ (fazendo $\mathbf{I} - \mathbf{H} = \mathbf{M}$)

$\mathbf{e} = \mathbf{M}\,\mathbf{Y}$. → **matriz de projeção**

1.6 Propriedades dos estimadores de MQO

O estimador $\hat{\boldsymbol{\beta}}$ apresenta todas as propriedades do caso linear simples, ou seja:

- $\hat{\boldsymbol{\beta}}$ é não-viciado, isto é, $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
- $\hat{\boldsymbol{\beta}}$ é não-viciado de mínima variância e sua matriz de variância-covariância é dada por $\text{Var-Cov}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) = \sigma^2\,(\mathbf{X'X})^{-1}$

em que, considerando $\mathbf{C} = (\mathbf{X'X})^{-1} = \begin{bmatrix} C_{11}\ C_{12} \ldots C_{1,k+1} \\ C_{12}\ C_{22} \ldots C_{2,k+1} \\ \ldots \quad \ldots \quad \ldots \ldots \\ C_{k1}\ C_{k2} \ldots C_{k+1,k+1} \end{bmatrix}$,

tem-se: $\text{Var}(\hat{\beta}_j) = \sigma^2 C_{j+1,\,j+1}$ $\quad j = 0, 1, \ldots, k$

$\text{Cov}(\hat{\beta}_m;\ \hat{\beta}_j) = \sigma^2 C_{m+1,\,j+1}$ $\quad m, j = 0, 1, \ldots, k$ e $m \neq j$.

- Ainda, assumindo os erros ε_i ($i = 1, .., n$) normalmente distribuídos, segue que $\hat{\boldsymbol{\beta}}$ é também estimador de máxima verossimilhança (EMV) de $\boldsymbol{\beta}$ e, sendo assim, $\hat{\boldsymbol{\beta}}$ é não-viciado, de mínima variância, consistente e suficiente.

1.7 Estimação de σ^2

Assim como em regressão linear simples, é possível obter um estimador para σ^2 utilizando-se a soma de quadrados residual, SQres, dada por:

$$\text{SQres} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e'e} =$$

$$= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\,(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{Y'Y} - 2\,\hat{\boldsymbol{\beta}}\,'\mathbf{X'Y} + \hat{\boldsymbol{\beta}}\,'\mathbf{X'X}\hat{\boldsymbol{\beta}}\,.$$

De fato, de $\mathbf{X'X}\hat{\boldsymbol{\beta}} = \mathbf{X'X}\,(\mathbf{X'X})^{-1}\,\mathbf{X'Y} = \mathbf{X'Y}$ segue que:

$$SQres = \mathbf{Y'Y} - \hat{\boldsymbol{\beta}}\,\mathbf{'X'Y},$$

que possui (n – p) graus de liberdade associados, visto que p parâmetros são estimados no modelo de regressão. Tem-se, assim, o quadrado médio residual definido por:

$$QMres = SQres / (n - p),$$

em que é possível mostrar que $E(QMres) = \sigma^2$. Portanto, um estimador não-viciado de σ^2 é dado por:

$$\hat{\sigma}^2 = QMres.$$

1.8 Análise de Variância (ANOVA)

Em termos matriciais tem-se:

$$SQres = \mathbf{Y'Y} - \hat{\boldsymbol{\beta}}\,\mathbf{'X'Y}$$

$$SQreg = \hat{\boldsymbol{\beta}}\,\mathbf{'X'Y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

$$SQtotal = \mathbf{Y'Y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}.$$

- De fato,

$$SQtotal = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(Y_i^2 - 2Y_i\overline{Y} + \overline{Y}^2\right) = \sum_{i=1}^{n} Y_i^2 - 2\overline{Y}\sum_{i=1}^{n} Y_i + \sum_{i=1}^{n}\overline{Y}^2$$

$$= \sum_{i=1}^{n} Y_i^2 - 2n\overline{Y}\frac{\sum_{i=1}^{n} Y_i}{n} + \sum_{i=1}^{n}\overline{Y}^2 = \mathbf{Y'Y} - n\,\overline{Y}^2 = \mathbf{Y'Y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}.$$

$$SQreg = SQtotal - SQres = (\mathbf{Y'Y} - n\,\overline{Y}^2) - (\mathbf{Y'Y} - \hat{\boldsymbol{\beta}}\,\mathbf{'X'Y}) = \hat{\boldsymbol{\beta}}\,\mathbf{'X'Y} - n\,\overline{Y}^2$$

$$= \hat{\boldsymbol{\beta}}\,\mathbf{'X'Y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}.$$

Tem-se, assim, a tabela de análise de variância apresentada a seguir:

Tabela da análise de variância (ANOVA) do modelo de regressão.

F.V.	S.Q.	g.l.	Q.M.	F	p-valor
Regressão	$\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$	p – 1	SQreg/(p – 1)	QMreg/QMres	depende de F
Resíduos	$\mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$	n – p	SQres/(n – p)	---	---
Total	$\mathbf{Y}'\mathbf{Y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$	n – 1	----		

n = tamanho amostral e p = número de parâmetros.

O teste F, mostrado na tabela, testa a existência de regressão linear entre a variável resposta Y e o conjunto de regressoras X_1, X_2, ..., X_k. Formalmente as hipóteses sob teste são:

H_o: $\beta_1 = \beta_2 = ... = \beta_k = 0$
H_a: nem todos os β_j (j = 1, 2, ..., *k*) são iguais a zero.

Obs: Somente a existência de uma relação de regressão não assegura que predições possam ser feitas usando tal relação.

Outra medida possível de ser definida é o **coeficiente de determinação múltiplo**, denotado po R^2, o qual é dado por.

$$R^2 = \frac{\boldsymbol{SQreg}}{\boldsymbol{SQtotal}} = 1 - \frac{\boldsymbol{SQres}}{\boldsymbol{SQtotal}} \qquad (0 \leq R^2 \leq 1).$$

Este coeficiente mede a redução proporcional da variação total de Y associada ao uso do conjunto de variáveis X_1, X_2, ..., X_k. R^2 assume o valor *zero* quando $\beta_j = 0$ (j = 1, ..., k) e assume o valor *um* quando todas as observações cairem diretamente na superfície de resposta, isto é, quando $Y_i = \hat{Y}_i$ para todo *i*.

Comentários

- Um valor de R^2 grande não implica necessariamente que o modelo ajustado seja útil. Outros aspectos precisam ser avaliados (suposições, parcimonia e sentido prático do modelo, dentre outros).

- Adicionar mais variáveis independentes ao modelo pode somente aumentar R^2 e nunca reduzí-lo, pois a SQres não pode tornar-se maior com mais variáveis independentes e a SQtotal é sempre a mesma para um certo conjunto de dados. Como R^2 pode tornar-se grande pela inclusão de um grande número de variáveis independentes, é sugerido que se faça uso de uma medida modificada, o **coeficiente de determinação múltiplo ajustado**, denotado por R^2_a, que ajusta R^2 dividindo cada soma de quadrados por seus graus de liberdade associados. Tem-se então:

$$R^2_a = 1 - \frac{SQres/(n-p)}{SQtotal/(n-1)} = 1 - \frac{(n-1)SQres}{(n-p)SQtotal}.$$

Note, que o coeficiente R^2_a pode tornar-se menor quando uma variável independente for adicionada ao modelo, pois o decréscimo na SQres pode ser compensado pela perda de graus de liberdade do denominador (n – p). Se R^2 e R^2_a diferirem dramaticamente um do outro, então existe grande chance de que o modelo tenha sido superespecificado, isto é, termos que contribuem não significativamente para o ajuste devem ter sido incluídos desnecessariamente. Avaliar R^2 e R^2_a é, desse modo, de grande utilidade no processo de seleção de variáveis, especialmente nos casos em que há um número grande de regressoras disponíveis.

Em um MRLM, é possível também obter o **coeficiente de correlação múltipla** entre Y e o conjunto de regressoras $X_1, X_2, ...X_k$, isto é:

$$r = + \sqrt{R^2}$$

sendo R^2 o coeficiente de determinação múltiplo apresentado anteriormente.

Esse coeficiente é uma generalização do coeficiente de correlação linear simples entre duas variáveis fornecendo, desse modo, a correlação linear entre Y e o conjunto de variáveis $X_1, X_2, ...X_k$.

1.9 Diagrama de Dispersão

Em regressão linear simples, o diagrama de dispersão é certamente uma ferramenta importante para analisar a relação entre Y e X. Poder-se-ia, então, pensar que esta ferramenta também seria igualmente útil em regressão linear múltipla, de modo que a análise visual dos gráficos de Y *versus* X_1, Y *versus* X_2, ..,Y *versus* X_k pudesse ajudar a acessar as relações entre Y e cada variável independente. Infelizmente isto, em geral, não é verdadeiro. Para discutir esse fato, considere os dados a seguir que foram gerados a partir da equação $Y = 8 - 5X_1 + 12\ X_2$.

Y	X_1	X_2
10	2	1
17	3	2
48	4	5
27	1	2
55	5	6
26	6	4
9	7	3
16	8	4

Para esses dados, observe, a partir da Figura 1, que o diagrama de dispersão de Y *versus* X_1 não exibe uma relação aparente entre essas duas variáveis. Já o diagrama de Y *versus* X_2, indica uma relação linear positiva com inclinação de aproximadamente 8. Ambos os diagramas conduzem, portanto, a informações errôneas a respeito da relação existente entre Y e X_1, bem como entre Y e X_2.

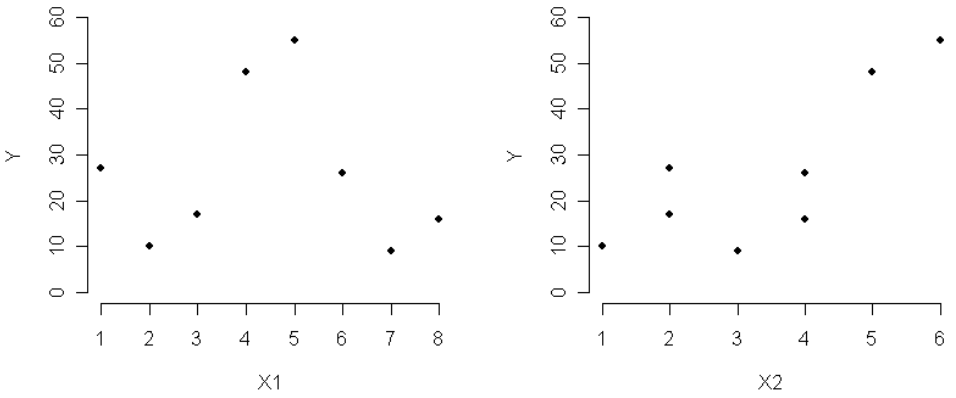


Figura 1: Diagramas de dispersão de Y *versus* X_1 e Y *versus* X_2.

Como exemplificado, os diagramas de dispersão de Y *versus* X_j (j = 1, ..., *k*) podem gerar enganos quando duas variáveis regressoras atuam em Y de modo aditivo e sem ruído (erro). Situações mais realísticas com diversas variáveis regressoras e erros nos Y_i's podem, portanto, produzir enganos ainda maiores.

Na presença de apenas uma variável regressora dominante, o correspondente diagrama de dispersão geralmente revelará isto. Contudo, quando diversas regressoras são importantes, ou quando as regressoras estiverem relacionadas entre si, esses diagramas serão praticamente inúteis.

2. Intervalos de confiança

2.1 Intervalo de confiança para os coeficientes da regressão

Na obtenção de intervalos de confiança dos coeficientes β_j (j = 0, 1,.., k), tem-se, em decorrência da suposição de que $\varepsilon_i \sim N(0; \sigma^2)$, $i = 1, ..., n$, que:

$$Y_i \sim N(\beta_o + \sum_{j=1}^{p} \beta_j x_{ij}, \sigma^2) \qquad i = 1, .., n.$$

Como $\hat{\boldsymbol{\beta}}$ é uma combinação linear dos Y_i's segue que:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (\mathbf{X'X})^{-1}).$$

Logo,

$$\hat{\beta}_j \sim N(\beta_j ; \sigma^2 C_{ii}) \qquad j = 0, 1, ..., k \quad i = j + 1,$$

em que C_{ii} é o i-ésimo elemento da diagonal da matriz $\mathbf{(X'X)}^{-1}$. Assim,

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{ii}}} \sim t_{n-p} \qquad \text{para } j = 0, 1,..., k \quad \text{e} \quad i = j + 1,$$

em que p = número de parâmetros do modelo ajustado e $\hat{\sigma}^2$ = QMres.

Portanto, um I.C.(1-α)100% para β_j ($j = 0, 1,..., k$) é dado por:

$$\hat{\beta}_j \pm t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 C_{ii}} .$$

Usualmente $\sqrt{\hat{\sigma}^2 C_{ii}}$ é chamado erro-padrão do coeficiente de regressão $\hat{\beta}_j$. Note que se o valor zero pertencer ao I.C., haverá evidências de que o parâmetro β_j não é estatísticamente significativo ao nível de significância α.

2.2 Intervalo de confiança para a resposta esperada

Para um particular ponto $\mathbf{x}_o = (1, x_{01}, x_{02}, ..., x_{0k})$ pode-se estimar a resposta esperada, bem como seu respectivo intervalo de confiança. O valor estimado em $\mathbf{x}_o$ e sua variância estimada são obtidos, respectivamente, por:

$$\hat{\boldsymbol{Y}}_o = \mathbf{x_0}\hat{\boldsymbol{\beta}} \qquad \text{e} \qquad \hat{V}(\hat{\boldsymbol{Y}}_0) = \hat{\sigma}^2 \mathbf{x_0}'(X'X)^{-1} \mathbf{x_o},$$

de modo que um I.C.(1-α)100% para a resposta esperada em $\mathbf{x}_0$ é:

$$\hat{\boldsymbol{Y}}_0 \pm t_{\alpha/2, n-p} \sqrt{\hat{V}(\hat{\boldsymbol{Y}}_0)} .$$

3 Testes de hipóteses

3.1 Teste para a significância da regressão

Para testar a significância da regressão, isto é, testar a existência da relação linear entre Y e pelo menos uma variável regressora X_1, X_2, ..., X_k, pode ser utilizado o teste F apresentado anteriormente na tabela da ANOVA, cuja estatística de teste é:

$$F = QMreg / QMres$$

e que, sob Ho, tem distribuição $F_{p-1;\, n-p}$. Se Ho for rejeitada, haverá evidências de que pelo menos um β_j difere de zero.

3.2 Testes para os coeficientes individuais da regressão

O teste F discutido anteriormente, testa o efeito conjunto, e não individual, das regressoras. Como o interesse está em manter no modelo somente as regressoras que são de real importância na explicação da resposta, é de interesse testar a significância de cada coeficiente individual, β_j, o que equivale a testar as hipóteses:

H_0: $\beta_j = 0$
H_a: $\beta_j \neq 0$.

A estatística de teste usada, em geral, para testar as hipóteses apresentadas é dada por:

$$t^* = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{ii}}} = \frac{\hat{\beta}_j}{e.p.(\hat{\beta}_j)} \overset{\text{sob } H_0}{\underset{\sim}{\uparrow}} t_{n-p} \quad (j = 0, 1, ..., k \quad \text{e} \quad i = j + 1),$$

em que C_{ii} é o i-ésimo elemento da diagonal da matriz $\mathbf{(X'X)^{-1}}$ e $\hat{\sigma}^2 = QMres$. Se H_o não for rejeitada, haverá evidências de que a contribuição da regressora X_j para a explicação de Y não é significativa e, desse modo, X_j pode ser excluída do modelo. Caso contário, a regressora deve ser mantida no modelo.

Este teste é chamado ***teste t parcial ou marginal***, pelo fato do coeficiente $\hat{\beta}_j$ depender de todas as outras regressoras X_i ($i \neq j$) presentes no modelo. É um teste, portanto, da contribuição de X_j ***na presença*** das outras regressoras no modelo.

3.2.1 Método da SQextra para testar os coeficientes da regressão

Para determinar a contribuição da regressora X_j para a SQreg, na presença das demais regressoras X_i ($i \neq j$) no modelo, pode-se também usar o, assim denominado, método da SQextra, método este também possível de ser usado para determinar a contribuição de um subconjunto de variáveis regressoras para o modelo.

A SQextra mede o **acréscimo** marginal na SQreg, quando uma ou diversas regressoras são adicionadas ao modelo de regressão ou, equivalentemente, a **redução** marginal na SQres, quando uma ou mais regressoras são adicionadas ao modelo.

Para ilustrar o método, considere para os dados de um estudo em que se tenha n = 20, Y = variável resposta e as regressoras contínuas X_1, X_2 e X_3, os modelos e resultados apresentados a seguir:

- Regressão de Y em X_1: $\hat{Y} = -1{,}496 + 0{,}8572x_1$.

F.V.	SQ	g.l.	QM
Reg	352,27	1	352,27
Res	143,12	18	7,95
Total	495,39	19	---

d.p. $(\hat{\beta}_1)$ = 0,1288

- Regressão de Y em X_2: $\hat{Y} = -23{,}634 + 0{,}8565x_2$.

F.V.	SQ	g.l.	QM
Reg	381,97	1	381,97
Res	113,42	18	6,30
Total	495,39	19	---

d.p. $(\hat{\beta}_2)$ = 0,11

- Regressão de Y em X_1 e X_2: $\hat{Y} = -19{,}174 + 0{,}2224x_1 + 0{,}6594x_2$.

F.V.	SQ	g.l.	QM
Reg	385,44	2	192,72
Res	109,95	17	6,47
Total	495,39	19	---

d.p. $(\hat{\beta}_1)$ = 0,3034

d.p. $(\hat{\beta}_2)$ = 0,2912

- Regressão de Y em X_1, X_2 e X_3: $\hat{Y} = 117{,}08 + 4{,}344x_1 - 2{,}857x_2 - 2{,}186x_3$.

F.V.	SQ	g.l.	QM
Reg	396,98	3	132,33
Res	98,41	16	6,15
Total	495,39	19	---

d.p. $(\hat{\beta}_1)$ = 3,016

d.p. $(\hat{\beta}_2)$ = 2,582

d.p. $(\hat{\beta}_3)$ = 1,596

Observe que:

- quando X_1 e X_2 estão no modelo tem-se SQres (X_1,X_2) = 109,95 e
- quando somente X_1 está no modelo tem-se SQres (X_1) = 143,12

e, ainda,

- quando X_1 e X_2 estão no modelo tem-se SQreg (X_1,X_2) = 385,44 e
- quando somente X_1 está no modelo tem-se SQreg (X_1) = 352,27.

A **diferença** entre as duas somas de quadrados de resíduos, ou entre as duas somas de quadrados da regressão, é chamada **Soma de Quadrados Extra** e será denotada por SQ_E $(X_2| X_1)$. Então:

$$\begin{aligned} SQ_E\,(X_2|\,X_1) &= SQres\,(X_1) - SQres(X_1, X_2) \\ &= SQreg\,(X_1,X_2) - SQreg\,(X_1) \\ &= 33{,}17. \end{aligned}$$

Esta **redução** na SQres ou **acréscimo** na SQreg é o resultado de adicionar X_2 ao modelo quando X_1 já se encontra nele. Assim, a $SQ_E(X_2 \mid X_1)$ mede o **efeito marginal** em adicionar X_2 ao modelo na presença de X_1.

Analogamente, é possível obter outras *somas de quadrados extra,* tais como:

- **efeito marginal** de adicionar X_3 ao modelo quando X_1 e X_2 estão presentes.

$SQ_E(X_3 \mid X_1, X_2) = SQres(X_1, X_2) - SQres(X_1, X_2, X_3) = 109{,}95 - 98{,}41 = 11{,}54$

ou

$SQ_E(X_3 \mid X_1, X_2) = SQreg(X_1, X_2, X_3) - SQreg(X_1, X_2) = 396{,}98 - 385{,}44 = 11{,}54.$

Tem-se, nesse caso, uma **redução** na SQres de 11,54 unidades ao quadrado ou, equivalentemente, um **acréscimo** na SQreg de 11,54 unidades ao quadrado ao se adicionar X_3 ao modelo em que X_1 e X_2 se encontram presentes.

- **efeito marginal** de adicionar X_2 e X_3 ao modelo quando X_1 está presente.

$SQ_E(X_2, X_3 \mid X_1) = SQres(X_1) - SQres(X_1, X_2, X_3) = 143{,}12 - 98{,}41 = 44{,}71$

ou

$SQ_E(X_2, X_3 \mid X_1) = SQreg(X_1, X_2, X_3) - SQreg(X_1) = 396{,}98 - 352{,}27 = 44{,}71.$

Tem-se, aqui, uma **redução** na SQres de 44,71 unidades ao quadrado ou, equivalentemente, um **acréscimo** na SQreg de 44,71 unidades ao quadrado ao serem adicionadas X_2 e X_3 ao modelo em que X_1 já se encontra presente.

O interesse, contudo, não está somente em obter tais reduções ou acréscimos, mas sim saber se a variável (ou as variáveis) X_j deve, ou não, ser incluída no modelo. Para essa finalidade, já foi visto que a estatística de teste parcial t^* é apropriada. Porém, alternativamente, pode-se usar a estatística de teste parcial F*, que usa as somas de quadrados extra.

No **exemplo** sob análise, pode ser de interesse, por exemplo, testar se a variável X_3 deve ser adicionada ao modelo contendo X_1 e X_2, o que equivale a testar as hipóteses: $H_0: \beta_3 = 0$ *versus* $H_a: \beta_3 \neq 0$.

Se H_0 não for rejeitada tem-se o **modelo reduzido** $E(Y \mid \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ e,
se H_0 for rejeitada tem-se o **modelo completo** $E(Y \mid \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.

A estatística de teste parcial F* para testar tais hipóteses é expressa por:

$$F^* = \frac{SQ_E(X_3 \mid X_1, X_2) / [(n-3)-(n-4)]}{SQres(X_1, X_2, X_3) / (n-4)} = \frac{SQ_E(X_3 \mid X_1, X_2) / 1}{QMres(X_1, X_2, X_3)} \sim F_{1;\, n-4.}$$

Para os dados do exemplo tem-se:

$F^* = 11{,}54 / 6{,}15 = 1{,}88$ (p-valor = 0,189)
$t^* = -2{,}186 / 1{,}596 = -1{,}37$ (p-valor = 0,189).

A partir dos resultados obtidos, conclui-se que X_3 não contribui significativamente e pode, portanto, ser retirada do modelo de regressão contendo X_1 e X_2.

Obs: Relembre que $F^* = (t^*)^2$.

Se for de interesse, o teste F*, pode também ser utilizado para testar se um subconjunto de regressoras pode ser retirado do modelo completo. Por exemplo:

- Testar se X_2 e X_3 podem ser retiradas do modelo completo, isto é, do modelo contendo X_1, X_2 e X_3. Nesse caso tem-se as hipóteses:

$$H_o: \beta_2 = \beta_3 = 0 \quad \textit{versus} \quad H_a: \beta_2 \neq 0 \text{ ou } \beta_3 \neq 0.$$

Se H_o não for rejeitada $\Rightarrow$ tem-se o **modelo reduzido**: $Y = \beta_o + \beta_1 x_1 + \varepsilon$ e,
se H_0 for rejeitada $\Rightarrow$ tem-se o **modelo completo**: $Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$.

Em sendo, $F^* = [(33{,}17 + 11{,}54)/2] / [98{,}41/16] = [44{,}71/2] / 6{,}15 = 3{,}63$, para o qual o p-valor associado à distribuição $F_{2;16}$ é igual a 0,05, é possível concluir pela rejeição da hipótese nula e, desse modo, opta-se pelo modelo completo.

3.2.2 - Anova contendo a decomposição da SQreg

Em regressão linear múltipla, uma variedade de decomposições ou desdobramentos da SQreg em SQ_E pode ser obtida. Por exemplo, no caso de duas variáveis independentes X_1 e X_2 pode-se obter:

$$SQreg\,(X_1, X_2) = SQ_E\,(X_1) + SQ_E\,(X_2 | X_1) \text{ ou,}$$
$$SQreg\,(X_1, X_2) = SQ_E\,(X_2) + SQ_E\,(X_1 | X_2).$$

A análise de variância (Anova), contendo uma dessas duas possíveis decomposições, é representada por:

F.V.	S.Q	g.l.	Q.M.
Regressão	**SQreg (X_1,X_2)**	**2**	**QMreg (X_1,X_2)**
X_1	$SQ_E\,(X_1)$	1	$QM_E\,(X_1)$
$X_2 \mid X_1$	$SQ_E\,(X_2 \mid X_1)$	1	$QM_E\,(X_2 \mid X_1)$
Resíduos	**SQres (X_1, X_2)**	**n – 3**	**QMres(X_1, X_2)**
Total	**Sqtotal**	**n – 1**	**--------**

Para mais de duas variáveis regressoras, a Anova com a decomposição é obtida de forma análoga. Alguns pacotes estatísticos fornecem tal decomposição, o que facilita a realização de alguns testes de interesse. A decomposição é feita, em geral, obedecendo a ordem de escolha das variáveis independentes.

Para o **exemplo** discutido na seção anterior, uma possível decomposição é apresentada na tabela a seguir.

F.V.	S.Q	g.l.	Q.M.
Regressão	**396,98**	**3**	**132,33**
X_1	352,27	1	352,27
$X_2 \mid X_1$	33,17	1	33,17
$X_3 \mid X_1, X_2$	11,54	1	11,54
Resíduos	**98,41**	**16**	**6,15**
Total	**495,39**	**19**	**----**

A partir da decomposição apresentada, é possível realizar alguns testes de interesse. Por exemplo:

(a) Teste da significância da regressão.

F* = [396,98/3] / [98,41/16] = 132,33/6,15 = 21,51 (p-valor = 7,3e-7).

(b) Teste da significância de X_3 na presença de X_1 e X_2.

F* = [11,54/1] / [98,41/16] = 11,54/6,15 = 1,88 (p-valor = 0,1892).

(c) Teste da significância de X_2 e X_3 na presença de X_1

F* = [(33,17 + 11,54)/2] / [98,41/16] = [44,71/2] / 6,15 = 3,63 (p-valor = 0,05).

3.5 - Coeficientes de determinação parcial

As SQextra não são somente úteis para testar coeficientes de um modelo de regressão linear múltipla, mas também para encontrar uma medida descritiva de relação denominada *coeficiente de determinação parcial*.

Enquanto o coeficiente de determinação múltiplo R^2 mede a proporcional redução na variabilidade de Y obtida pela introdução de um conjunto de variáveis regressoras no modelo, o coeficiente de determinação parcial mede a contribuição marginal de uma variável X_j quando outras variáveis se encontram no modelo. Logo,

- **coeficiente de determinação parcial entre Y e X_2 dado X_1 no modelo**

$$r^2_{Y2\bullet 1} = \frac{SQ_E(X_2|X_1)}{SQres(X_1)}.$$

- **coeficiente de determinação parcial entre Y e X_1 dado X_2 e X_3 no modelo**

$$r^2_{Y1\bullet 23} = \frac{SQ_E(X_1|X_2,X_3)}{SQres(X_2,X_3)}$$

e assim por diante.

Para os dados do exemplo discutido anteriormente tem-se:

(a) $r^2_{Y2\bullet1}$ = 33,17/143,12 = 0,2317 (23,17%),

(b) $r^2_{Y3\bullet12}$ = 11,54/109,95 = 0,105 (10,5%) e,

(c) $r^2_{Y1\bullet2}$ = 3,47/113,42 = 0,031 (3,1%).

Então, quando X_2 é adicionada ao modelo contendo X_1 a SQres(X_1) é reduzida em 23,17%. Ainda, a SQres(X_1, X_2) é reduzida em 10,5% quando X_3 é adicionada ao modelo e, finalmente, se o modelo contém X_2, adicionar X_1 reduz a SQres em 3,1%.

3.5.1 - Coeficientes de correlação parcial

A raiz quadrada de um coeficiente de determinação parcial é denominado *coeficiente de correlação parcial*. O sinal de cada coeficiente de correlação parcial correspondente ao do coeficiente de regressão do modelo ajustado. Assim, tem-se:

(a) $r_{Y2\bullet1} = (0,2317)^{1/2} = -0,48$ (sinal negativo pois $\hat{\beta}_2 = -2,857$),

(b) $r_{Y3\bullet12} = (0,105)^{1/2} = -0,324$ (sinal negativo pois $\hat{\beta}_3 = -2,186$) e,

(c) $r_{Y1\bullet2} = (0,031)^{1/2} = 0,176$ (sinal positivo pois $\hat{\beta}_1 = 4,344$).

Obs: os coeficientes de correlação parcial são geralmente usados nas rotinas computacionais para encontrar a próxima melhor variável independente a entrar no modelo (discutido adiante).

4. Diagnóstico do modelo de regressão linear múltipla

Foi visto que a ANOVA é útil para diagnosticar alguns aspectos do modelo de regressão linear ajustado. É necessário, contudo, verificar outros aspectos tais como: suposições dos erros, não-linearidade de algumas regressoras, multicolinearidade, bem como a existência e o efeito de pontos atípicos. Análise dos resíduos será utilizada nesse sentido.

4.1 - Análise dos resíduos

Os métodos gráficos usados em regressão linear simples são também úteis em regressão linear múltipla. Alguns gráficos adicionais podem também trazer informações importantes. Em síntese, tem-se os seguintes gráficos e suas utilidades:

(a) **Resíduos em papel de probabilidade Normal** (**e_i x F_i**)
- examinar se os erros apresentam distribuição aproximadamente Normal;
- auxiliar na detecção de pontos atípicos.

(b) **Resíduos *versus* valores ajustados (e_i x $\hat{Y}_i$)**

- verificar homogeneidade das variâncias dos erros;
- fornecer informações sobre pontos atípicos.

(c) **Resíduos *versus* seqüência de coleta (se conhecida)** ($e_{(i)}$ x i)

- informações sobre possível correlação entre os erros.

(d) **Resíduos *versus* cada X_j incluída no modelo** (e_i x X_{ij})

- informações adicionais sobre a adequacidade da função de regressão com respeito a j-ésima variável independente, ou seja, auxilia na detecção de não-linearidade na regressora X_j;
- informações sobre possível variação na magnitude da variância dos erros em relação a variável independente X_j;
- informações sobre dados atípicos.

(e) **Resíduos parciais *versus* X_{ij} para cada X_j no modelo** (e_{ij}^* x X_{ij})

- revelar mais precisamente a relação entre os resíduos e cada regressora X_j. O *i*-ésimo resíduo parcial para a regressora X_j é definido por:

$$e_{ij}^* = e_i + \hat{\beta}_j x_{ij} \qquad (i = 1, ..., n)$$

$$e_{ij}^* = (Y_i - \hat{Y}_i) + \hat{\beta}_j x_{ij} \qquad (i = 1, ..., n).$$

O gráfico dos resíduos parciais, como comumente é referenciado, é semelhante ao gráfico dos resíduos *versus* X_{ij} e permite ao experimentador avaliar falhas de linearidade, presença de *outliers* e heterogeneidade de variâncias.

Se, por exemplo, a relação entre Y e X_j não for linear, o gráfico dos resíduos parciais indicará mais precisamente do que o gráfico e_i *versus* X_j como transformar os dados para obter a linearidade. A justificativa para isto é que o gráfico dos resíduos parciais mostra a relação entre Y e X_j após o efeito das outras regressoras X_i ($i \neq j$) ter sido removido e, desse modo, este gráfico mostrará mais claramente a influência de X_j em Y na presença das outras regressoras.

Observe que e_{ij}* *versus* X_j deve ser **linear** com inclinação próxima a $\hat{\beta}_j$ se a relação entre Y e X_j for linear.

(f) **Resíduos *versus* X_k omitidas do modelo**

- ajuda a revelar a dependência da resposta Y com uma ou mais das regressoras não presentes no modelo. Qualquer estrutura (padrão sistemático), que não o aleatório, indicarão que a inclusão daquela variável pode melhorar o modelo.

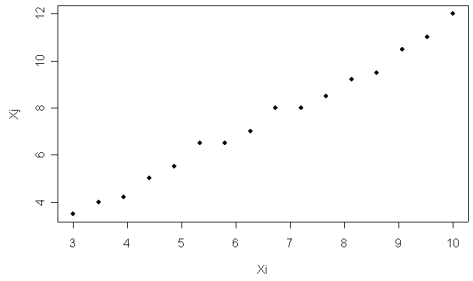
(g) **Resíduos *versus* interações não incluídas no modelo**

- úteis para examinar se alguma, algumas ou todas as interações são requeridas no modelo. Um padrão sistemático nestes gráficos, que não o aleatório, sugere que o efeito da interação pode estar presente.

(h) **Gráfico da regressora X_i *versus* regressora X_j ($i \neq j$)**

- útil para estudar a relação entre as variáveis regressoras e a disposição dos dados no espaço **X**;
- encontrar pontos atípicos.

Considere, como exemplo, o gráfico a seguir.



Este gráfico mostra que as regressoras X_1 e X_2 são altamente correlacionadas e conseqüentemente, pode não ser necessário incluir ambas no modelo.

Quando duas ou mais variáveis regressoras forem altamente corelacionadas, *multicolinearidade* está presente nos dados. Logo, o problema de multicolinearidade existe quando há uma dependência quase-linear entre as regressoras. A presença de multicolinearidade pode afetar seriamente o ajuste por MQO e, em algumas situações, produzir modelos quase inúteis. A matriz de correlação $\mathbf{r}_{XX}$ das regressoras é uma ferramenta útil na detecção de multicolinearidade.

$$\mathbf{r}_{XX} = \begin{bmatrix} 1 & r_{12} & .. & r_{ik} \\ r_{21} & 1 & & r_{2k} \\ .. & .. & .. & .. \\ r_{k1} & r_{k2} & .. & 1 \end{bmatrix}.$$

A matriz $\mathbf{r}_{XX}$ é simétrica, isto é, $r_{ij} = r_{ji}$ e se r_{ij} for próximo de zero, então X_i e X_j não são altamente correlacionadas. Por outro lado, se r_{ji} for próximo de |1|, então X_i e X_j são altamente correlacionadas.

4.2 Propriedades dos resíduos

Foi visto anteriormente que $\mathbf{e = (I - H)Y}$, então segue que:

$$\mathbf{e = (I - H)\,(X\beta + \varepsilon) = X\beta - HX\beta + (I - H)\,\varepsilon}$$
$$\mathbf{= X\beta - X(X'X)^{-1}X'X\,\beta + (I - H)\,\varepsilon = (I - H)\,\varepsilon}.$$

Logo, $\mathbf{E(e) = 0}$

e, $\mathbf{V(e) = V[(I - H)\,\varepsilon] = (I - H)\,V(\varepsilon)\,(I - H)' = (I - H)\,\sigma^2\,I\,(I - H)' = \sigma^2(I - H)}$

pois, $\mathbf{(I - H)}$ é simétrica ($\mathbf{(I - H) = (I - H)'}$) e idempotente ($\mathbf{(I - H)(I - H)) = (I - H)}$).

Portanto, $E(e_i) = 0$, $Var(e_i) = \sigma^2(1 - h_{ii})$, bem como pode ser mostrado que $Cov(e_i, e_j) = -\sigma^2 h_{ij}$. Ainda, a distribuição dos resíduos é também normal, pois estes são combinações lineares dos Y_i's, que têm distribuição normal. Em síntese:

$$e_i \sim N(0, \sigma^2(1 - h_{ii})) \qquad i = 1, ..., n$$
$$Cov(e_i, e_j) = -\sigma^2(h_{ij}) \qquad i, j = 1, ..., n \quad (i \neq j).$$

Os resíduos não são, portanto, independentes e possuem variâncias diferentes que dependem do valor de X correspondente a x_i.

Com o intuito de melhor analisar os resíduos, levando em conta sua variabilidade, algumas transformações foram propostas na literatura. Algumas delas são apresentadas no quadro a seguir.

1) Resíduos *standardized*	$d_i = \dfrac{e_i}{\sqrt{QMres}}$
2) Resíduos *studentized*	$r_i = \dfrac{e_i}{\sqrt{QMres(1 - h_{ii})}}$
3) Resíduos PRESS	$e_{(i)} = \dfrac{e_i}{1 - h_{ii}}$
4) Resíduos *studendized* externamente (R-Student)	$t_i = \dfrac{e_i}{\sqrt{S^2_{(i)}\ (1 - h_{ii})}}$ sendo $S^2_{(i)} = \dfrac{(n - p - 1)QMres - e_i^2(1 - h_{ii})}{n - p}$

h_{ii} corresponde ao *i*-ésimo componente da diagonal da matriz $\mathbf{H = X(X'X)^{-1}X'}$.

Diversos autores recomendam o uso dos resíduos *studentized* em vez de e_i ou d_i. A justificativa é que h_{ii} é uma medida da localização do *i*-ésimo ponto no espaço **X** e a variância de e_i depende de onde o ponto $\mathbf{x}_i$ cai. Então, como $0 \leq h_{ii} \leq 1$, usar o QMres para estimar a variância dos resíduos faz com que ela seja superestimada.

- Pontos com grande resíduo e grande h_{ii} são observações, possivelmente, altamente influentes no ajuste por MQO.
- Resíduos associados a pontos os quais h_{ii} é grande terão resíduos PRESS grandes. Esses pontos geralmente serão altamente influentes.

5. Multicolinearidade

Adicional às analises dos gráficos de X_i *versus* X_j ($i \neq j$) e da matriz de correlação $\mathbf{r}_{XX}$, é possível utilizar outros recursos para diagnosticar a presença de colinearidade ou multicolinearidade.

5.1 Fatores de inflação da variância (VIF)

O VIF para o j-ésimo coeficiente de regressão pode ser escrito por:

$$VIF_j = \frac{1}{1 - R^2_j},$$

em que R^2_j é o coeficiente de determinação múltiplo obtido da regressão de X_j com as demais variáveis regressoras. Claramente, se X_j for quase linearmente dependente com alguma das outras regressoras, então R^2_j será próximo de 1 e VIF_j será grande. Experiências práticas indicam que VIF maiores que 10 implicam que os coeficientes de regressão associados estão sendo pobremente estimados devido a multicolineridade.

5.2 Análise dos autovalores da matriz $\mathbf{r}_{XX}$

As raízes características, ou autovalores de $\mathbf{r}_{XX}$, denotadas por $\lambda_1, \lambda_2, ..., \lambda_k$, podem ser usados para medir a extensão da multicolinearidade nos dados. Se existirem uma, ou mais, dependência linear nos dados, então uma, ou mais, das raízes características serão pequenas. Auto valores de $\mathbf{r}_{XX}$ são as raízes características da equação $| \mathbf{r}_{XX} - \lambda \mathbf{I} | = 0$.

Alguns analistas preferem, no entanto, examinar o número de condição da matriz $\mathbf{r}_{XX}$ dado por:

$$k = \frac{\lambda_{max}}{\lambda_{min}}.$$

Geralmente, se $k < 100 \Rightarrow$ não existem sérios problemas de multicolinearidade,
se $100 < k < 1000 \Rightarrow$ moderada a forte multicolinearidade e,
se $k > 1000 \Rightarrow$ severa multicolinearidade.

Os índices de condição da matriz $\mathbf{r}_{XX}$ são dados por: $k_j = \frac{\lambda_{max}}{\lambda_j}$.

Exemplo: Suponha Y = variável resposta e X_1, ..., X_9 as regressoras, de modo que os autovalores obtidos sejam:

$\lambda_1 = 4{,}2048$	$\lambda_4 = 1{,}0413$	$\lambda_7 = 0{,}0136$
$\lambda_2 = 2{,}1626$	$\lambda_5 = 0{,}3845$	$\lambda_8 = 0{,}0051$
$\lambda_3 = 1{,}1384$	$\lambda_6 = 0{,}0495$	$\lambda_9 = 0{,}0001.$

Assim, $k = 42048$, o que implica em severa multicolinearidade. Ainda,

$k_1 = 1{,}0$	$k_4 = 4{,}04$	$k_7 = 309{,}18$
$k_2 = 1{,}94$	$k_5 = 10{,}94$	$k_8 = 824{,}47$
$k_3 = 3{,}69$	$k_6 = 84{,}96$	$k_9 = 42048,$

e como k_7 e $k_8 > 100$ e $k_9 > 1000$, há indícios de multicolinearidade envolvendo as variáveis X_7, X_8 e X_9.

5.3 Determinante da matriz $\mathbf{r}_{xx}$

O determinante da matriz $\mathbf{r}_{XX}$ pode ser usado como um indicador de existência de multicolineridade. Os valores possíveis deste determinante são $0 \leq \det(\mathbf{r}_{XX}) \leq 1$. Se $\det(\mathbf{r}_{XX}) = 1$, as regressoras são ortogonais, enquanto $\det(\mathbf{r}_{XX}) = 0$ implica em dependência linear exata entre as regressoras. O grau de multicolinearidade torna-se mais severo quando o determinante aproxima-se de zero.

6. Diagnóstico de influência

Não é incomum em uma análise de dados encontrar um subconjunto de observações que exerça uma influência desproporcional no modelo de regressão ajustado. Localizar essas observações e acessar seu impacto no modelo é, desse modo, de interesse. A seguir, diversas medidas de influência são apresentadas.

6.1 Pontos de Alavancagem

A disposição dos pontos no espaço **X** é importante para a determinação das propriedades do modelo. Em particular, observações potencialmente remotas têm desproporcional alavancagem nos parâmetros estimados, bem como nos valores preditos e nas usuais estatísticas sumárias. Para localizar esses pontos remotos no espaço **X**, Hoaglin e Welsh (1978) sugeriram o uso da matriz *chapéu***,** a qual é obtida por $\mathbf{H = X(X'X)^{-1}X'}$. De acordo com esses autores, a inspeção dos elementos da matriz **H** pode revelar pontos que são potencialmente influentes em virtude de sua localização no espaço **X**. Atenção é usualmente centrada nos elementos da diagonal da matriz **H**, ou seja, nos h_{ii}. Como,

$$\sum_{i=1}^{n} h_{ii} = \text{rank}(\mathbf{H}) = \text{rank}(\mathbf{X}) = \text{p},$$

em que p é o número de parâmetros do modelo, tem-se que o tamanho médio de um elemento da diagonal da matriz **H** é $\bar{h}$ = p/n e, assim, como uma regra um tanto grosseira, tem-se que:

> se $h_{ii} > 2(p/n) \Rightarrow$ a observação *i* é um possível ponto de alta alavancagem.

6.2 Influência nos coeficientes da regressão

Se for desejado, contudo, considerar a localização do ponto, bem como a variável resposta, Cook (1979) sugeriu o uso de uma medida que considera o quadrado da distância entre as estimativas $\hat{\boldsymbol{\beta}}$ obtidas com todas as *n* observações (pontos) e as estimativas obtidas retirando-se a *i*-ésima observação (ponto), denotada por $\hat{\boldsymbol{\beta}}_{(i)}$. Essa medida é expressa por:

$$D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{p \text{ QMres}} \qquad i = 1, ..., n.$$

Pontos com grandes valores de D_i têm considerável influência nas estimativas $\hat{\boldsymbol{\beta}}$. Os valores de D_i são comparados com a distribuição $F_{\alpha,\, p,\, n-p}$. Se $D_i \approx F_{\alpha,\, p,\, n\,-\,p,}$ então retirar o ponto *i* deve deslocar $\hat{\boldsymbol{\beta}}$ para o limite de uma região de confiança de 50% de $\boldsymbol{\beta}$ baseado nos dados completos. Isto é uma grande discordância e indica que as estimativas obtidas por MQO são sensíveis ao *i*-ésimo ponto. Como $F_{0.5;\, n;\, n\,-\,p} \approx 1$, usualmente consideram-se os pontos em que $D_i > 1$ como sendo possivelmente influentes. Idealmente, seria desejado que cada estimativa $\hat{\boldsymbol{\beta}}_{(i)}$ permanecesse dentro dos limites de uma região de confiança de 10 ou 20%.

Belsley, Kuh e Welsch (1980) sugeriram, ainda, uma estatística que indica o quanto cada coeficiente de regressão $\hat{\beta}_j$ muda, em unidades de desvio-padrão, se a *i*-ésima observação for removida. Esta estatística é dada, para j = 0, 1, ..., k, por:

$$\text{DFBeta}_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{(S_{(i)})^2 C_{j+1,j+1}}} \qquad i = 1, .., n,$$

sendo $C_{j+1, j+1}$ o (j +1)-ésimo elemento da diagonal da matriz $\mathbf{C} = (\mathbf{X'X})^{-1}$.

Um valor grande de $\text{DFBeta}_{j,i}$ indica que a observação *i* tem considerável influência no j-ésimo coeficiente de regressão. O ponto de corte $2/\sqrt{n}$ é, em geral, usado para comparar os $\text{DFBetas}_{j,i.}$ Para amostras grandes, observações as quais $|\text{DFBeta}_{j,i}| > 2/\sqrt{n}$ merecem atenção. Para amostras pequenas ou moderadas, as observações que merecem atenção são aquelas em que $|\text{DFBeta}_{j,i}| > 1$.

6.3 Influência nos valores ajustados

É possível, também, investigar a influência da *i*-ésima observação nos valores ajustados (preditos). Uma medida razoável é:

$$\text{DFFit}_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{(S_{(i)})^2 h_{ii}}} \qquad i = 1, ..., n,$$

sendo $\hat{y}_{(i)}$ o valor predito de $\hat{y}_i$ sem o uso da *i*-ésima observação. O denominador é somente uma padronização. Assim, DFFit_i é o quanto o valor ajustado muda, em unidades de desvio-padrão, se a *i*-ésima observação for removida.

Geralmente, observações em que $|\text{DFFit}_i| > 1$, para amostras pequenas ou moderadas, e $|\text{DFFit}_i| > 2\sqrt{p/n}$, para amostras grandes, merecem atenção.

6.4 Influência na precisão da estimação

As medidas D_i, $\text{DFBeta}_{j,i}$ e DFFit_i fornecem uma visão do efeito de cada observação nos coeficientes estimados e nos valores ajustados. Elas não fornecem, contudo, qualquer informação sobre a precisão geral da estimação. Para expressar o papel da *i*-ésima observação na precisão da estimação pode ser definido a medida a seguir.

$$\text{Covratio}_i = \frac{|(\text{X}'_{(i)} X_{(i)})^{-1} (S_{(i)})^2|}{|(\text{X'X})^{-1} \text{QMres}|} \qquad i = 1, ..., n.$$

Pontos de corte para Covratio_i não são fáceis de serem obtidos. Belsley et al.(1980) sugeriram que se $\text{Covratio}_i > 1 + (3p/n)$ ou se $\text{Covratio}_i < 1 - (3p/n)$, então, o *i*-ésimo ponto deve ser um possível ponto influente. O limite inferior é somente válido quando $n > 3p$. Em geral, esses pontos de corte são mais apropriados para amostras grandes.

EXEMPLO: Um engarrafador de bebidas está analisando os serviços de rotina realizados no sistema de distribuição de máquinas acionadas por moedas. Ele está interessado em predizer o tempo requerido para esses serviços de rotina que incluem a estocagem da máquina com bebidas e pequenas manutenções. O engenheiro industrial responsável sugeriu duas variáveis como sendo as que mais afetam o tempo requerido por esses serviços: a quantidade de bebida estocada e a distância percorrida pelo profissional responsável pelos serviços. Os dados estão na Tabela 1.

Tabela 1: Dados observados no estudo de máquinas acionadas por moedas.

Tempo requerido (em minutos)	Quantidade estocada (em unidades)	Distância percorrida (em pés)
16.68	7	560
11.50	3	220
12.03	3	340
14.88	4	80
13.75	6	150
18.11	7	330
8.00	2	110
17.83	7	210
79.24	30	1460
21.50	5	605
40.33	16	688
21.00	10	215
13.50	4	255
19.75	6	462
24.00	9	448
29.00	10	776
15.35	6	200
19.00	7	132
9.50	3	36
35.10	17	770
17.90	10	140
52.32	26	810
18.75	9	450
19.83	8	635
10.75	4	150

Fonte: Montgomery and Peck (1992).

Para predizer o tempo requerido para os serviços de rotina, utilizando como regressoras a quantidade de bebida estocada e a distância percorrida pelo profissional responsável pelos serviços, existem vários modelos de regressão a serem investigados e, dentre eles: a) regressão de Y em X_1 e X_2, b) regressão de Y em X_1 e c) regressão de Y em X_2

Observando, inicialmente, a matriz de correlação $\mathbf{r}_{XX,}$ a seguir,

$$r_{xx} = \begin{bmatrix} 1 & 0{,}8242 \\ 0{,}8242 & 1 \end{bmatrix},$$

bem como os fatores de inflação da variância (VIF_j):

$$VIF_1 = 1/[1-(0{,}8242)^2] = 3{,}1185 < 10$$

e

$$VIF_2 = 1/[1-(0{,}8242)^2] = 3{,}1185 < 10,$$

é possível observar que, embora X_1 e X_2 sejam multicolineares, tem-se ambos os VIF's menores que 10, o que indica que os coeficientes da regressão de Y em X_1 e X_2 não estarão sendo tão pobremente estimados por causa dessa multicolinearidade.

Ajustando, então, os três modelos mencionados foram obtidos os resultados apresentados a seguir.

Modelos	$\hat{\beta}_o$	$\hat{\beta}_1$	$\hat{\beta}_2$	QMres	R^2	R_a^2
Y em X_1 e X_2	2,341	1,615	0,014	10,6	0,9596	0,9559
Y em X_1	3,321	2,176	-	17,5	0,9305	0,9275
Y em X_2	4,961	-	0,042	51,5	0,7951	0,7862

Dos resultados apresentados, pode-se observar que, dentre os modelos considerados, os dois primeiros parecem ser bons candidatos ao melhor modelo.

A partir dos testes *t* associados aos parâmetros, mostrados na Tabela 2, há evidências para a rejeição das hipóteses nulas Ho: $\beta_1 = 0$ e Ho: $\beta_2 = 0$, pois os p-valores são menores que 0,001. Logo, há evidências de que ambas as regressoras, X_1 e X_2, são importantes na predição de Y. Caso não haja problemas relativos aos pressupostos assumidos para esse MRLM, tem-se, de acordo com o coeficiente de determinação, R^2, que as regressoras X_1 e X_2 estariam, conjuntamente, explicando em torno de 96% da variação total de Y.

Observe, ainda, que o modelo Y em X_1 parece ser também um bom candidato, visto ter uma quantidade pequena de parâmetros em que, a regressora X_1, sozinha, estaria explicando em torno de 93% da variação total de Y. Escolhendo o modelo Y em X_1 e X_2 tem-se as estimativas dos parâmetros apresentadas na Tabela 2.

Tabela 2. Estimativas dos parâmetros do modelo de Y em X_1 e X_2.

	$\hat{\beta}_j$	e.p($\hat{\beta}_j$)	*t*	*p*-valor
Intercepto	2,341	1,096	2,135	0,044
X_1	1,615	0,170	9,464	3,25e-9
X_2	0,014	0,003	3,981	0,0006

A Anova, com a decomposição da SQreg, apresentada no Quadro I, mostra que o modelo com somente X_1 reduz a SQtotal em 5382,4 unidades ao quadrado e que, ao adicionarmos X_2 ao modelo contendo X_1, há uma redução na SQres de 168,4 unidades ao quadrado.

Quadro I - Análise de variância do MRLM Y em X_1 e X_2.

F.V.	**G.L.**	**SQ**	**QM**	**F**	**Pr(>F)**
Regressão	2	5550,8	2775,4	261,20	4,687e-16
X1	1	5382,4	5382,4	506,62	< 2,2e-16
X2	1	168,4	168,4	15,85	0,0006312
Resíduos	22	233,7	10,6		
Total	24	5784,5			

A análise de resíduos evidenciou, como pode ser observado nos gráficos dos resíduos apresentados na Figura 2, que a observação 9 causa alguns problemas ao modelo ajustado.

Do diagnóstico de influência, apresentado na Tabela 3, pode-se observar:

1) os elementos h_{ii} da diagonal da matriz **H** mostram que as observações 9 e 22 são maiores que $2\,\bar{h} = 2(p/n) = 6/25 = 0,24$, o que evidencia que tais pontos devem ser investigados, pois são potenciais pontos influentes;

2) o maior valor da distância de Cook é $D_9 = 3,42$, indicando que as estimativas dos parâmetros obtidas por MQO são sensíveis a essa observação;

3) inspeção dos DFFit's revela que as observações 9 e 22 excedem o valor 1 (ponto de corte para pequenas amostras);

4) inspeção dos DFBeta's mostram, considerando o valor 1 como ponto de corte, que as observações 9 e 22 apresentam grande efeito em pelo menos uma das três estimativas dos parâmetros;

5) os pontos de corte para Covratio, nesse estudo, são 0,64 e 1,36. Assim, as observações 9 e 22 e, também, a observação 16, apresentam-se como possíveis pontos influentes.

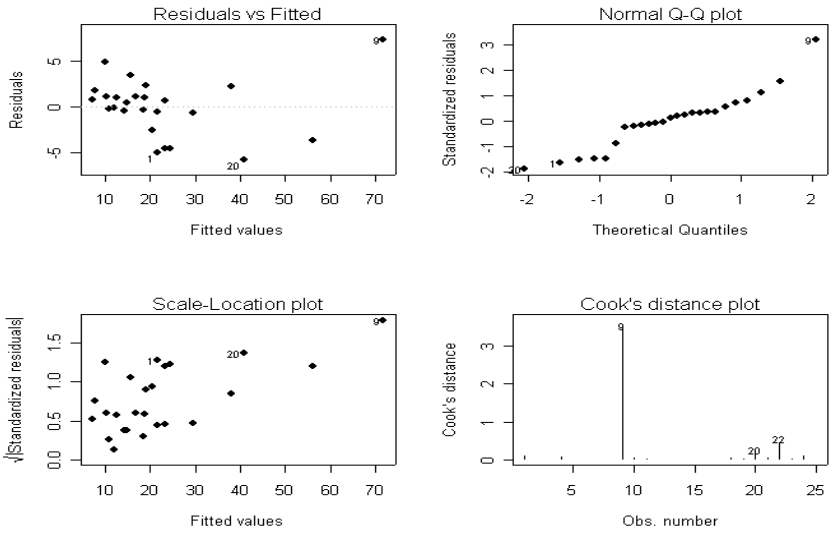


Figura 2. Análise gráfica dos resíduos do modelo de regressão Y em X_1 e X_2.

Tabela 3 – Resultados das estatísticas para detecção de pontos influentes.

	dfb.1.	dfb.X1	dfb.X2	dffit	cov.r	cook.d	hat(h_{ii})	
1	-0.18727	0.41131	-0.43486	-0.5709	0.871	1.00e-01	0.1018	
2	0.08979	-0.04776	0.01441	0.0986	1.215	3.38e-03	0.0707	
3	-0.00352	0.00395	-0.00285	-0.0052	1.276	9.46e-06	0.0987	
4	0.45196	0.08828	-0.27337	0.5008	0.876	7.76e-02	0.0854	
5	-0.03167	-0.01330	0.02424	-0.0395	1.240	5.43e-04	0.0750	
6	-0.01468	0.00179	0.00108	-0.0188	1.200	1.23e-04	0.0429	
7	0.07807	-0.02228	-0.01102	0.0790	1.240	2.17e-03	0.0818	
8	0.07120	0.03338	-0.05382	0.0938	1.206	3.05e-03	0.0637	
9	-2.57574	0.92874	1.50755	4.2961	0.342	3.42e+00	0.4983	*
10	0.10792	-0.33816	0.34133	0.3987	1.305	5.38e-02	0.1963	
11	-0.03427	0.09253	-0.00269	0.2180	1.172	1.62e-02	0.0861	
12	-0.03027	-0.04867	0.05397	-0.0677	1.291	1.60e-03	0.1137	
13	0.07237	-0.03562	0.01134	0.0813	1.207	2.29e-03	0.0611	
14	0.04952	-0.06709	0.06182	0.0974	1.228	3.29e-03	0.0782	
15	0.02228	-0.00479	0.00684	0.0426	1.192	6.32e-04	0.0411	
16	-0.00269	0.06442	-0.08419	-0.0972	1.369	3.29e-03	0.1659	
17	0.02886	0.00649	-0.01570	0.0339	1.219	4.01e-04	0.0594	
18	0.24856	0.18973	-0.27243	0.3653	1.069	4.40e-02	0.0963	
19	0.17256	0.02357	-0.09897	0.1862	1.215	1.19e-02	0.0964	
20	0.16804	-0.21500	-0.09292	-0.6718	0.760	1.32e-01	0.1017	
21	-0.16193	-0.29718	0.33641	-0.3885	1.238	5.09e-02	0.1653	
22	0.39857	-1.02541	0.57314	-1.1950	1.398	4.51e-01	0.3916	*
23	-0.15985	0.03729	-0.05265	-0.3075	0.890	2.99e-02	0.0413	
24	-0.11972	0.40462	-0.46545	-0.5711	0.948	1.02e-01	0.1206	
25	-0.01682	0.00085	0.00559	-0.0176	1.231	1.08e-04	0.0666	

Claramente, as observações 9 e 22 são as que merecem maior atenção em nossa análise. Para investigar o efeito dessas observações no modelo de regressão Y em X_1 e X_2, observe os resultados apresentados a seguir.

Modelo	$\hat{\beta}_o$	$\hat{\beta}_1$	$\hat{\beta}_2$	QMres	R^2
com obs 9 e 22	2,341	1,616	0,014	10,62	0,9596
sem obs 9	4,447	1,498	0,010	5,90	0,9487
sem obs 22	1,916	1,786	0,012	10,06	0,9564
sem obs 9 e 22	4,643	1,456	0,011	6,16	0,9072

Note, que retirar a observação 9 produz mudanças de 90% em β_0, 7,3% em β_1 e 28% em β_2. Portanto, esta observação exerce razoável influência nos coeficientes. Por outro lado, a retirada da observação 22 produz mudanças relativamente menores nos coeficientes da regressão. A retirada de ambas produz mudanças similares àquelas observadas quando da retirada somente da observação 9. Conclui-se, assim, que as observações 9 e 22, mais fortemente a 9, influenciam no ajuste do modelo.

Investigações subseqüentes, realizadas junto ao pesquisador, podem revelar razões para a retirada de uma, ou ambas, as observações da análise. Nesse caso, as análises devem ser refeitas.

Gráficos dos resíduos do modelo de regressão Y em X_1 e X_2 sem a observação 9 são apresentados na Figura 3. Comparando com os gráficos da Figura 2, note que os pressupostos encontram-se melhores atendidos sem a referida observação.

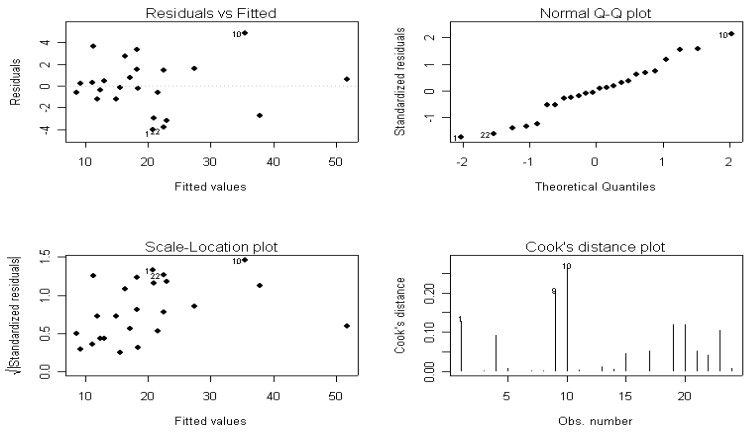


Figura 3. Gráficos dos resíduos do modelo Y em X_1 e X_2 sem a observação 9.

Relembrando que o modelo de regressão Y em X_1 mostrou ser também um bom candidato para a análise desses dados, visto ter apresentado $R^2 = 0,9305$, seria interessante comparar os resíduos desse modelo com os do modelo de regressão Y em X_1 e X_2, ambos sem a observação 9. Esses gráficos encontram-se na Figura 4 e mostram resultados muito similares aos obtidos anteriormente (Figura 3).

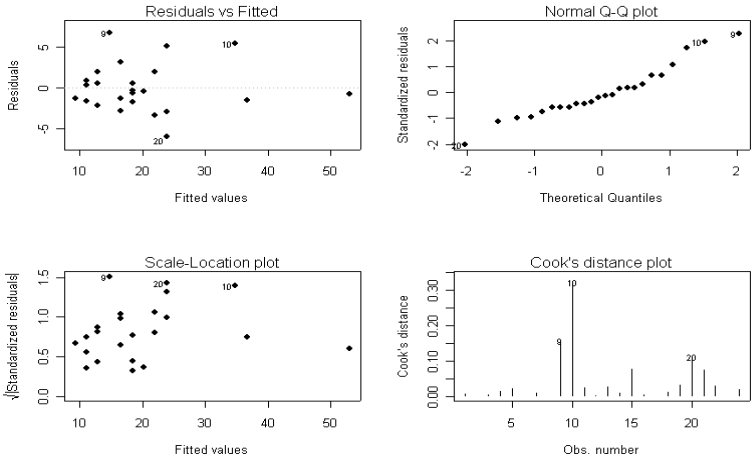


Figura 4. Gráficos dos resíduos do modelo Y em X_1 sem a observação 9.

A partir da discussão dos resultados apresentada, pode-se, juntamente com o pesquisador, proceder a escolha por um dos modelos. A escolha deve ser por aquele que se apresentar mais razoável para a predição da variável resposta Y.

Caso o modelo escolhido seja aquele com as regressoras X_1 e X_2 sem a observação 9, tem-se o modelo ajustado dado por:

$$\hat{Y} = 4{,}447 + 1{,}498\ x_1 + 0{,}010\ x_2.$$

A representação gráfica desse modelo é mostrada na Figura 5.

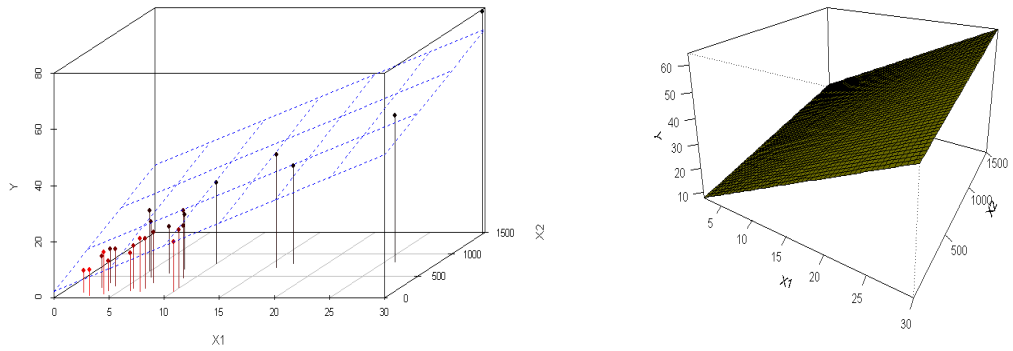


Figura 5. Plano ajustado aos dados de serviços de rotina em máquinas acionadas por moedas.

A partir das estimativas dos parâmetros, pode-se concluir que, mantendo-se a quantidade de bebida estocada constante, há um acréscimo de 0,010 minutos no tempo médio para a realização dos serviços de rotina a cada acréscimo de uma unidade na distância percorrida (X_2). De forma análoga, mantendo-se a distância constante, há um acréscimo de 1,498 minutos no tempo médio para a realização dos serviços de rotina a cada acréscimo de uma unidade em X_1 (quantidade de bebida estocada).

Um intervalo de 95% confiança para a resposta média em, por exemplo, $x_1 = 8$ e $x_2 = 275$ resulta em (17,65; 20,79) minutos, sendo $\hat{Y} = 19{,}22$ minutos.

7. Métodos para tratar com a multicolinearidade

7.1 Coleta adicional de dados

Em alguns casos, coletar dados adicionais para combinações de X_i e X_j em que se tenha poucos dados observados pode ajudar a solucionar, ou amenizar, o problema que decorre da presença de multicolinearidade.

Infelizmente, a coleta de dados adicionais nem sempre é possível devido aos custos ou mesmo a impossibilidade devido ao processo sendo estudado. Além disso, o problema pode ser devido a características estruturais da população e, portanto, coletar novos dados nesses casos tem pouco valor. Por exemplo, para um estudo da relação entre renda familiar (X_1) e tamanho da residência (X_2), é altamente provável que não se encontre algumas combinações dessas duas variáveis, tal como a combinação: renda alta e residência muito pequena.

7.2 Reespecificação do modelo

Em situações as quais se tenha duas ou mais variáveis regressoras altamente correlacionadas em um modelo de regressão, pode-se tentar fazer uso de alguma reespecificação desse modelo, a fim de contornar o problema que a multicolinearidade em geral acarreta nas estimativas dos parâmetros. Uma forma de reespecificação é *redefinir as regressoras.* Para ilustrar, suponha que em um determinado estudo as regressoras X_1, X_2 e X_3 apresentem alta correlação, isto é sejam multicolineares. Tentar redefinir essas regressoras por meio de uma função como, por exemplo:

$$X = (X_1 + X_2) / X_3 \qquad \text{ou} \qquad X = X_1 * X_2 * X_3,$$

que, por um lado, preserva a informação contida nas regressoras originais e, por outro, reduz, ou contorna, o problema da multicolinearidade (mal condicionamento da matriz **X**), é uma forma de reespecificar o modelo. Um exemplo de redefinição de variáveis que pode ser citado como sendo bastante usual em diversos estudos é o índice de massa corporal definido por $IMC = X_1/(X_2)^2$, sendo X_1 o peso e X_2 a altura.

Outra forma de reespecificação muito usada é a *eliminação de regressoras* do modelo. Assim, se, por exemplo, X_1, X_2 e X_3 forem multicolineares, eliminar X_3, ou outra, pode ser útil. A técnica de eliminação é altamente efetiva, mas pode, em alguns casos, prejudicar o poder preditivo do modelo.

7.3 Regressão Ridge

Foi visto que o estimador de MQ $\hat{\boldsymbol{\beta}}$ apresenta a propriedade de ser não-viciado. Foi visto, ainda, que em algumas situações em que as regressoras não são ortogonais, estimativas muito pobres podem ser obtidas usando-se esse estimador. Para esses casos, uma alternativa à da reespecificação, é a de se encontrar um estimador $\hat{\boldsymbol{\beta}}^*$ viciado, tal que seu vício seja pequeno, mas que sua variância seja menor do que a de $\hat{\boldsymbol{\beta}}$. O termo *regressão ridge* é usado para denominar um modelo de regressão em que esse tipo de estimador é considerado. Para mais detalhes sobre esse assunto, pode ser consultado, por exemplo, o livro de Montgomery e Peck (1992).

8. Seleção de variáveis e construção do modelo

Em diversos estudos, é comum que o pesquisador tenha uma grande quantidade de regressoras as quais ele acredita influenciarem a resposta. Dentre essas regressoras, é de interesse encontrar um subconjunto apropriado para o modelo de regressão. Selecionar esse subconjunto envolve, em geral, dois objetivos conflitantes:

1º) o modelo deveria incluir tantas quantas regressoras fossem necessárias para auxiliar na predição de Y e,

2º) o modelo deveria ser parcimonioso (conter poucas regressoras), visto que a variância da predição cresce conforme o número de regressoras cresce. Além disso, quanto mais regressoras existirem no modelo, maior o custo para coleta e manutenção do modelo.

O processo de encontrar um modelo que concilie esses objetivos é denominado *seleção da melhor equação de regressão*. Infelizmente, não existe uma definição única de *melhor*. Nessa direção, diversos procedimentos (algorítmos) foram e vendo sendo propostos para selecionar tal subconjunto de regressoras. Tais procedimentos, em geral, especificam diferentes subconjuntos de regressoras como sendo os melhores.

A seguir, são apresentados alguns critérios usados nesses procedimentos e, em seguida, os procedimentos de seleção propriamente ditos. É importante salientar que nenhum dos procedimentos de seleção fornece qualquer garantia de que a *melhor* equação de regressão tenha sido obtida. Desse modo, devem ser tratados pelo analista apenas como métodos para explorar e melhor entender a estrutura dos dados.

8.2 Critérios para avaliação dos modelos

No problema de seleção de regressoras, dois aspectos são importantes: i) encontrar um subconjunto delas para predizer Y e ii) decidir se o subconjunto escolhido é melhor do que um outro. Alguns critérios usados para essa finalidade são:

a) Coeficiente de determinação múltiplo R^2

Para cada subconjunto composto de *k* regressoras tem-se associado ao modelo de regressão correspondente, um valor para o coeficiente de determinação múltiplo, denotado por R^2_p, sendo p o número de parâmetros do modelo. O valor de R^2_p cresce quando *k* (*k* o número de regressoras) cresce e é máximo quando todas as *k* regressoras são usadas. Assim, o analista pode usar o critério de adicionar regressoras até o ponto em que a adição de uma variável não for mais útil, pois fornece um acréscimo muito pequeno em R^2_p.

b) Coeficiente de determinação múltiplo ajustado R^2_a ou QMres

Devido a algumas dificuldades em interpretar o coeficiente R^2, alguns analistas preferem usar o coeficiente de determinação múltiplo ajustado. O critério é escolher o subconjunto de regressoras que forneça o valor máximo de R^2_a, o que equivale a encontrar o subconjunto que minimize o QMres.

c) Estatística C_p de Mallows

Mallows propôs um critério que se baseia na SQres. De acordo com esse critério, deve ser calculado para cada subconjunto de *k* regressoras, a correspondente SQres e, então, obter:

$$C_p = \frac{SQres(p)}{\sigma^2} - n + 2p \quad ,$$

em que σ^2 é estimado pelo QMres do modelo com as *k* regressoras e p é o número de parâmetros em cada modelo.

Para o modelo com todas as regressoras $C_p = p$. Geralmente, valores pequenos de C_p são desejáveis. Modelos de regressão com C_p próximos da linha $C_p = p$ e abaixo dela são candidatos ao *melhor modelo*.

8.3 Técnicas computacionais para seleção de variáveis

8.3.1 Todas as regressões possíveis

Esse procedimento requer que o analista ajuste todas as equações de regressão envolvendo uma regressora, duas regressoras e, assim sucessivamente. As equações são, então, avaliadas de acordo com os critérios vistos, ou uma ponderação deles, sendo o *melhor* modelo escolhido de acordo com os resultados de tais critérios. É claro que o número de equações cresce rapidamente com o aumento do número de regressoras. Se p =10, por exemplo, tem-se $2^{10} = 1024$ regressões possíveis.

8.3.2. Pesquisa direta dos resultados dos testes *t*

A estatística de teste *t* usada para testar Ho: $\beta_j = 0$ para o modelo com todas as *k* regressoras é $t_j = \hat{\beta}_j / \text{d.p.}(\hat{\beta}_j)$. Regressoras com contribuição significativa para o modelo apresentarão *p*-valores pequenos associados aos t_j's (j = 1, ... *k*), e devem, assim, fazer parte do subconjunto das *melhores* regressoras. Assim, fixar um nível de significância α, ordenar as regressoras de acordo com a ordem crescente de seus *p*-valores associados aos t_j's e, então, incluí-las, uma a uma, no modelo enquanto *p*-valor $\leq \alpha$, deve levar ao melhor modelo ou, pelo menos, a um dos possíveis melhores modelos.

Esta estratégia de seleção é frequentemente muito efetiva quando o número de regressoras é relativamente grande ($k > 20$ ou 30).

8.3.3. Métodos Passo a Passo (*Stepwise*)

Como o procedimento que avalia todas as regressões possíveis é, em algumas situações, muito árduo, vários outros métodos têm sido propostos para a avaliação de somente um número pequeno de modelos de regressão, seja pela adição ou pela retirada de regressoras em cada passo. Esses métodos são referenciados como *procedimentos do tipo stepwise* e podem ser classificados em: i) seleção passo à frente *(forward*), ii) seleção passo atrás (*backward)* e iii) seleção passo a passo *(stepwise)*. A seguir, é descrito cada um deles em mais detalhes.

i) Seleção passo à frente *(Forward)*

Esse procedimento começa sem nenhuma regressora no modelo de regressão. Um nível de significância α é, então, estabelecido a fim de que somente regressoras que apresentem significância menor ou igual a α possam fazer parte do modelo. Tem-se, assim, os seguintes passos:

Passo 1: a *primeira* regressora a entrar no modelo é a que apresentar maior correlação simples com a resposta Y, isto é, maior r_{YXj} (j = 1, ..., *k*) ou, equivalentemente, o menor *p*-valor associado ao teste F tal que *p*-valor $\leq \alpha$ pré-estabelecido,
Passo 2: a *segunda* regressora a entrar no modelo é, agora, a com maior correlação parcial com a resposta Y, isto é, maior $r_{Yj\bullet i}$ para $j \neq i$ e i o índice da regressora escolhida no primeiro passo. Em outras palavras, a segunda regressora escolhida é aquela com maior estatística F parcial tal que *p*-valor $\leq \alpha$ pré-estabelecido,
Passos subsequentes: análogo ao segundo passo, as demais regressoras são inseridas, uma a uma, no modelo até que, em um particular passo, todas as regressoras que ainda não estevirem no modelo apresentem *p*-valores associados aos testes F parciais $> \alpha$.

Note, nesse procedimento, que uma regressora escolhida em um determinado passo para fazer parte do modelo, permanece no modelo até o passo final, não havendo a possibilidade de exclusão da mesma em qualquer outro passo subsequente.

ii) Eliminação passo atrás *(backward)*

A eliminação *backward* começa com todas as *k* regressoras no modelo. Um nível de significância α é, então, estabelecido, a fim de que somente regressoras que apresentem significância maior que α possam ser removidas do modelo. Tem-se, assim:

Passo 1: para cada regressora é calculado a estatística F parcial como se ela fosse a última regressora a entrar no modelo. A primeira regressora a ser eliminada do modelo é aquela que apresentar o menor valor para a estatística F parcial ou, equivalentemente, o maior *p*-valor associado a essa estatística tal que *p*-valor $> \alpha$ pré-estabelecido,
Passo 2 e passos subsequentes: o modelo sem a regressora escolhida no primeiro passo é ajustado e novas estatísticas F parcias são calculadas para esse novo modelo. O procedimento usado no primeiro passo é, então, repetido até que todas as regressoras que permanecerem no modelo apresentem *p*-valores associados aos testes F parciais menores ou iguais a α.

Nesse procedimento, uma regressora escolhida em um determinado passo para ser removida do modelo, não tem a possibilidade de vir a ser incluída em qualquer outro passo subsequente.

iii) Seleção passo a passo *(stepwise)*

Assim como no procedimento de seleção *forward*, esse procedimento começa sem nenhuma regressora no modelo de regressão. Dois níveis de significância, α_1 e α_2, são, no entanto, estabelecidos. Um deles para ser usado como critério de inclusão e, o outro, como critério de exclusão de regressoras. Alguns analistas escolhem $\alpha_1 = \alpha_2$, mas isso não é necessário. Os passos usados nesse procedimento são os seguintes:

Passo 1: a *primeira* regressora a entrar no modelo é a que apresentar maior correlação simples com a resposta Y, isto é, maior r_{YXj} (j = 1, ..., *k*) ou, equivalentemente, o menor *p*-valor associado ao teste F tal que *p*-valor $\leq \alpha_1$ pré-estabelecido,
Passo 2: a *segunda* regressora a entrar no modelo é a que apresentar maior correlação parcial com a resposta Y, isto é, maior $r_{Yj \bullet i}$ para $j \neq i$ e i o índice da regressora escolhida no primeiro passo. Em outras palavras, a segunda regressora escolhida é aquela com maior estatística F parcial tal que *p*-valor $\leq \alpha_1$ pré-estabelecido,
Passo 3 e Passos subsequentes: todas as regressoras são reacessadas por meio de suas respectivas estatísticas F parciais. Em havendo regressoras no modelo, segundo α_2 estabelecido, que se mostrem redundantes na presença das demais, remove-se a que apresentar menor significância para o modelo. Caso não existam regressoras a serem removidas, analisam-se as estatísticas F parciais das regressoras que não estão no modelo. Existindo, dentre elas, regressoras com *p*-valores que sejam $\leq \alpha_1$, inclui-se a que apresentar menor *p*-valor. O procedimento pára quando não mais houver, segundo os níveis de significância α_1 e α_2 estabelecidos, regressoras que possam ser incluídas ou excluídas do modelo de regressão.

Nesse procedimento, uma regressora escolhida em um determinado passo para fazer parte do modelo, pode ser removida em um outro passo. Em cada passo pode ocorrer ou a inclusão ou a exclusão de uma determinada regressora, nunca a inclusão de uma e a exclusão de outra simultaneamente em um mesmo passo.

Comentários:

- Os procedimentos de seleção *forward*, eliminação *backward* e seleção *stepwise,* não necessariamente levam a escolha do mesmo modelo final;
- Recomenda-se que todos os procedimentos sejam aplicados na esperança de que haja alguma concordância entre eles, ou mesmo para aprender algo mais sobre a estrutura dos dados;
- O procedimento de seleção *forward* tende a concordar com o de todas as regressões possíveis para subconjuntos pequenos de regressoras, enquanto o procedimento de eliminação *backward* para subconjuntos grandes de regressoras.
- O modelo final obtido por qualquer um dos procedimentos deve ser analisado quanto ao seu sentido prático. Analistas inexperientes podem concluir por um modelo não realístico. Para o modelo escolhido sugere-se, portanto, que o analista responda as questões a seguir.

1ª) O modelo obtido é razoável? Isto é, as regressoras no modelo fazem sentido à luz do problema real?
2ª) O modelo é útil para o propósito pretendido? (custos para a coleta dos dados, regressoras observáveis na prática etc.).
3ª) Os coeficientes de regressão são razoáveis? Isto é, os sinais e magnitude dos coeficientes são realísticos e seus erros-padrão relativamente pequenos?
4ª) O modelo apresenta bom ajuste aos dados? (análise de resíduos, diagnóstico de influência etc.).

9. Extrapolações

Em regressão linear múltipla, muito cuidado deve ser tomado quanto a extrapolações. Somente interpolações são permitidas e, para fazê-las, não se deve olhar meramente para a variação de cada regressora, mas sim para a região conjunta definida por elas, como ilustrado na Figura 6.

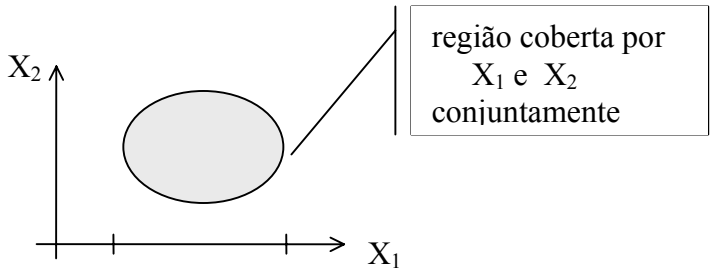


Figura 6 – Região conjunta de X1 e X2.

Para detectar se um ponto $\mathbf{x} = (1, x_{i1}, x_{i2}, ..., x_{i,k})$ pertence à região conjunta, usa-se o seguinte procedimento baseado nos elementos h_{ii} da diagonal da matriz **H**. Considere $h_{máx}$ = maior valor de h_{ii} . O conjunto de pontos **x** que satisfizerem:

$$\mathbf{x'(X'X)^{-1}x} \leq h_{máx}$$

estarão inclusos no elipsóide ou região conjunta definida pelas regressoras. Logo, se o interesse for o de predizer Y em $\mathbf{x}_o = [1 \quad x_{o1} \quad x_{o2} \quad .... \quad x_{op}]$, a localização desse ponto relativa ao elipsóide será obtida por:

$$h_{oo} = \mathbf{x'_o (X'X)^{-1} x_o}.$$

Se $h_{oo} > h_{máx}$, $\mathbf{x_o}$ estará fora do elipsóide. Caso contrário, $\mathbf{x_o}$ estará dentro ou nos limites do elipsóide.

Exemplo: Para os dados apresentados na Tabela 4 pode-se observar um conjunto de quatro regressoras, o que resulta em $2^4 = 16$ possíveis equações de regressão (todas com intercepto e sem a presença de interações).

Tabela 4: Dados observados em um estudo envolvendo quatro regressoras.

Observação i	Y_i	X_{i1}	X_{i2}	X_{i3}	X_{i4}
1	78,5	7	26	6	60
2	74,3	1	29	15	52
3	104,3	11	56	8	20
4	87,6	11	31	8	47
5	95,9	7	52	6	33
6	109,2	11	55	9	22
7	102,7	3	71	17	6
8	72,5	1	31	22	44
9	93,1	2	54	18	22
10	115,9	21	47	4	26
11	83,8	1	40	23	34
12	113,3	11	66	9	12
13	109,4	10	68	8	12

Fonte: Montgomery e Peck, 1992.

Para cada uma das 16 regressões citadas, foram obtidos: SQres, R^2, $R^2_{a,}$ QMres e C_p de Mallows. Os resultados são mostrados na Tabela 5. Tem-se, ainda, nas Tabelas 6 e 7, respectivamente, a matriz de correlações simples e os coeficientes das regressões estimados para cada uma das 16 regressões consideradas.

Tabela 5. Resumo de todas as regressões possíveis com o intercepto e sem interações.

no. de regressoras	no. de parâmetros	regressoras no modelo	SQres	R^2_p	R^2_a	QMres	C_p
Nenhuma	1	Nenhuma	2715,76	0	0	226,31	443,14
1	2	X_1	1265,68	0,5339	0,4915	115,06	202,55
1	2	X_2	906,33	0,6662	0,6359	82,39	142,49
1	2	X_3	1939,40	0,2858	0,2209	176,31	315,16
1	2	X_4	883,86	0,6745	0,6449	80,35	138,73
2	**3**	**X_1 X_2**	**57,90**	**0,9786**	**0,9744**	**5,79**	**2,68**
2	3	X_1 X_3	1227,07	0,5481	0,4578	122,70	198,10
2	3	X_1 X_4	74,76	0,9724	0,9669	7,47	5,50
2	3	X_2 X_3	415,44	0,8470	0,8164	41,54	62,44
2	3	X_2 X_4	868,88	0,6800	0,6160	86,88	138,23
2	3	X_3 X_4	175,73	0,9352	0,9223	17,57	22,37
3	4	X_1 X_2 X_3	48,11	0,9822	0,9763	5,34	3,04
3	4	X_1 X_2 X_4	47,97	0,9823	0,9764	5,33	3,02
3	4	X_1 X_3 X_4	50,83	0,9812	0,9750	5,64	3,50
3	4	X_2 X_3 X_4	73,81	0,9728	0,9637	8,20	7,34
4	5	X_1 X_2 X_3 X_4	47,86	0,9823	0,9735	5,98	5,00

Tabela 6. Matriz de correlações simples.

	X_1	X_2	X_3	X_4
X_1	1,0			
X_2	0,229	1,0		
X_3	**-0,824**	-0,139	1,0	
X_4	-0,245	**-0,973**	0,030	1,0
Y	0,731	0,816	-0,535	-0,821

Tabela 7. Estimativas por MQO para as 16 regressões consideradas.

Regressoras no modelo	$\hat{\beta}_o$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
X_1	81,479	1,869			
X_2	57,424		**0,789**		
X_3	110,203			-1,256	
X_4	117,568				-0,738
X_1 X_2	52,577	1,468	0,662		
X_1 X_3	72,349	2,312		0,494	
X_1 X_4	103,097	1,440			-0,614
X_2 X_3	72,075		0,731	-1,008	
X_2 X_4	94,160		**0,331**		-0,457
X_3 X_4	131,282			-1,200	-0,724
X_1 X_2 X_3	48,194	1,696	0,657	0,250	
X_1 X_2 X_4	71,648	1,452	0,416		-0,237
X_1 X_3 X_4	111,684	1,052		-0,410	-0,643
X_2 X_3 X_4	203,642		**-0,923**	-1,448	-1,557
X_1 X_2 X_3 X_4	62,405	1,551		0,102	-0,144

Observe, a partir da Tabela 7, que ao ser considerada somente a regressora X_2 no modelo, a estimativa de seu efeito na esperança de Y é de 0,789. Se, contudo, X_4 é adicionada a esse modelo, esse efeito é reduzido para 0,311 e, mais, se X_3 é adicionada ao modelo contendo X_2 e X_4, o efeito de X_2 muda para – 0,923. Fica, portanto, evidente que a estimativa de um coeficiente depende muito de quais outras regressoras estão presentes no modelo. As grandes mudanças observadas nos coeficientes quando regressoras são adicionadas, ou removidas, indicam a existência de substancial correlação entre as regressoras, o que pode ser confirmado por meio da matriz de correlações simples mostrada na Tabela 6. Note, a partir desta matriz, que os pares (X_1, X_3) e (X_2, X_4) encontram-se altamente correlacionados (multicolineares), visto as correlações obtidas serem próximas de –1. Essa estrutura de correlação é parcialmente responsável pelas mudanças observadas nos coeficientes estimados e, sendo assim, a multicolineridade está afetando consideravelmente a estimação dos parâmetros do modelo. Conseqüentemente, adicionar outras regressoras, seja quando X_1 e X_2, seja quando X_1 e X_4, estiverem no modelo, será de pouca utilidade, visto que a informação contida nas demais regressoras está essencialmente presente em X_1 e X_2 ou em X_1 e X_4.

Avaliando os modelos por meio dos coeficientes de determinação R^2_p, tem-se, a partir dos resultados mostrados na Tabela 5, que após a inclusão de duas regressoras no modelo, pouco se ganha em termos de R^2 com a introdução de novas regressoras. Ambos os modelos com 2 regressoras, Y em (X_1, X_2) e Y em (X_1, X_4), apresentam essencialmente o mesmo R^2. Com base nesse critério, faria pouca diferença na escolha de um ou outro.

Utilizando o critério de maximizar R^2_a, que é equivalente a minimizar o QMres, é possível observar, também a partir da Tabela 5, que o modelo com menor QMres é o que contém as regressoras X_1, X_2 e X_4 (QMres = 5,33). Observe, como esperado, que o modelo que minimiza o QMres é também o que maximiza R^2_a. Pode-se, ainda, observar, desta mesma tabela, que dois outros modelos com três regressoras (Y em X_1, X_2 e X_3 e Y em X_1, X_3 e X_4) e dois outros modelos com duas regressoras (Y em X_1 e X_2 e Y em X_1 e X_4) apresentam valores para o QMres comparáveis. Se X_1 e X_2 ou X_1 e X_4 estiverem no modelo, existe pequena redução no QMres com a adição de outras regressoras. Por esse critério e, entre os dois últimos modelos mencionados, o modelo Y em X_1 e X_2 parece ser mais apropriado por apresentar menor QMres.

Pelo critério de Mallows, observam-se quatro modelos considerados aceitáveis (modelos com $C_p < p = 5$). Se outros fatores forem levados em consideração, tais como, por exemplo, custos e dificuldades na obtenção das medidas, parece ser mais apropriado a escolha do modelo Y em X_1 e X_2 por este apresentar menor C_p.

Nos Quadros 2, 3 e 4 podem ser observados os resultados dos procedimentos de seleção *forward*, *backward* e *stepwise* aplicados, respectivamente, aos dados mostrados na Tabela 4. A partir dos resultados, tem-se:

i) modelo resultante da seleção *forward*: Y em X_4, X_1 e X_2
ii) modelo resultante da seleção *backward*: Y em X_1 e X_2
iii) modelo resultante da seleção *stepwise*: Y em X_1 e X_2.

Quadro 2. Resultados da seleção *forward* (passo à frente) obtidos no *software* R.

```
Start:  AIC= 443.14
 Y ~ 1
       Df Sum of Sq     RSS       Cp  F value     Pr(F)
+ X4    1   1831.90  883.87 138.8038  22.7985 0.0005762 ***
+ X2    1   1809.43  906.34 142.5613  21.9606 0.0006648 ***
+ X1    1   1450.08 1265.69 202.6533  12.6025 0.0045520 **
+ X3    1    776.36 1939.40 315.3145   4.4034 0.0597623 .
<none>               2715.76 443.1410

Step:  AIC= 138.8
 Y ~ X4
       Df Sum of Sq    RSS       Cp  F value     Pr(F)
+ X1    1    809.10  74.76   5.5020 108.2239 1.105e-06 ***
+ X3    1    708.13 175.74  22.3876  40.2946 8.375e-05 ***
+ X2    1     14.99 868.88 138.2977   0.1725    0.6867
<none>               883.87 138.8038

Step:  AIC= 5.5
 Y ~ X4 + X1
       Df Sum of Sq    RSS     Cp F value   Pr(F)
+ X2    1    26.789 47.973 3.0222  5.0259 0.05169 .
+ X3    1    23.926 50.836 3.5010  4.2358 0.06969 .
<none>               74.762 5.5020

Step:  AIC= 3.02
 Y ~ X4 + X1 + X2
       Df Sum of Sq    RSS     Cp F value Pr(F)
<none>               47.973 3.0222
+ X3    1     0.109 47.864 5.0040  0.0182 0.896

Call:
lm(formula = Y ~ X4 + X1 + X2)

Coefficients:
(Intercept)           X4           X1           X2
    71.6483      -0.2365       1.4519       0.4161
```

Quadro 3. Resultados da seleção *backward* (passo atrás) obtidos no *software* R.

```
Start:  AIC= 5
 Y ~ X1 + X2 + X3 + X4
       Df Sum of Sq    RSS     Cp F value   Pr(F)
- X3    1     0.109 47.973 3.0222  0.0182 0.89592
- X4    1     0.247 48.111 3.0453  0.0413 0.84407
- X2    1     2.972 50.836 3.5010  0.4968 0.50090
<none>               47.864 5.0040
- X1    1    25.951 73.815 7.3436  4.3375 0.07082 .

Step:  AIC= 3.02
 Y ~ X1 + X2 + X4
       Df Sum of Sq    RSS       Cp  F value     Pr(F)
- X4    1      9.93  57.90   2.6830   1.8633   0.20540
<none>               47.97   3.0222
- X2    1     26.79  74.76   5.5020   5.0259   0.05169 .
- X1    1    820.91 868.88 138.2977 154.0076 5.781e-07 ***

Step:  AIC= 2.68
 Y ~ X1 + X2
       Df Sum of Sq     RSS      Cp F value     Pr(F)
<none>               57.90   2.683
- X1    1    848.43  906.34 142.561 146.523 2.692e-07 ***
- X2    1   1207.78 1265.69 202.653 208.582 5.029e-08 ***

Call:
lm(formula = Y ~ X1 + X2, data = exe5)
Coefficients:
(Intercept)           X1           X2
    52.5773       1.4683       0.6623
```

Quadro 4. Resultados da seleção *stepwise* (passo a passo) obtidos no *software* R.

```
Start:  AIC= 443.14
 Y ~ 1
       Df Sum of Sq     RSS       Cp  F value     Pr(F)
+ X4    1   1831.90  883.87 138.8038  22.7985 0.0005762 ***
+ X2    1   1809.43  906.34 142.5613  21.9606 0.0006648 ***
+ X1    1   1450.08 1265.69 202.6533  12.6025 0.0045520 **
+ X3    1    776.36 1939.40 315.3145   4.4034 0.0597623 .
<none>               2715.76 443.1410

Step:  AIC= 138.8
 Y ~ X4
       Df Sum of Sq     RSS       Cp  F value     Pr(F)
+ X1    1    809.10   74.76   5.5020 108.2239 1.105e-06 ***
+ X3    1    708.13  175.74  22.3876  40.2946 8.375e-05 ***
+ X2    1     14.99  868.88 138.2977   0.1725 0.6866842
<none>               883.87 138.8038
- X4    1   1831.90 2715.76 443.1410  22.7985 0.0005762 ***

Step:  AIC= 5.5
 Y ~ X4 + X1
       Df Sum of Sq     RSS       Cp  F value     Pr(F)
+ X2    1     26.79   47.97   3.0222   5.0259   0.05169 .
+ X3    1     23.93   50.84   3.5010   4.2358   0.06969 .
<none>               74.76   5.5020
- X1    1    809.10  883.87 138.8038 108.2239 1.105e-06 ***
- X4    1   1190.92 1265.69 202.6533 159.2952 1.815e-07 ***

Step:  AIC= 3.02
 Y ~ X4 + X1 + X2
       Df Sum of Sq    RSS       Cp  F value     Pr(F)
- X4    1      9.93  57.90   2.6830   1.8633   0.20540
<none>               47.97   3.0222
+ X3    1      0.11  47.86   5.0040   0.0182   0.89592
- X2    1     26.79  74.76   5.5020   5.0259   0.05169 .
- X1    1    820.91 868.88 138.2977 154.0076 5.781e-07 ***

Step:  AIC= 2.68
 Y ~ X1 + X2
       Df Sum of Sq     RSS       Cp  F value     Pr(F)
<none>               57.90   2.6830
+ X4    1      9.93   47.97   3.0222   1.8633    0.2054
+ X3    1      9.79   48.11   3.0453   1.8321    0.2089
- X1    1    848.43  906.34 142.5613 146.5227 2.692e-07 ***
- X2    1   1207.78 1265.69 202.6533 208.5818 5.029e-08 ***
Call:
lm(formula = Y ~ X1 + X2)
Coefficients:
(Intercept)           X1           X2
    52.5773       1.4683       0.6623
```

Observe que não existe uma escolha clara da *melhor* equação de regressão. Cada procedimento sugere, em geral, equações diferentes. Todas são candidatas ao *modelo final* e devem ser analisadas quanto a sua adequacidade, pontos influentes, efeito de multicolinearidade etc.

Utilizando uma ponderação dos resultados obtidos quando da utilização dos critérios e métodos de seleção, parece razoável a indicação de dois modelos como candidatos ao *melhor* modelo. São eles, o modelo Y em X_1 e X_2 seguido do modelo Y em X_1 e X_4. Uma análise dsses dois modelos quanto a sua adequacidade (análise de resíduos, pontos influentes, necessidade de interação etc.) certamente auxiliarão na decisão de escolha do modelo final.

10. Validação dos Modelos de Regressão

Para todo modelo ajustado deve ser verificado, de algum modo, sua validade. O objetivo da validação de um modelo é o de verificar se ele funcionará na prática fornecendo, assim, uma proteção tanto para o modelo ajustado quanto para o usuário.

10.1 Técnicas de Validação

Três procedimentos são úteis para verificar a validade de um modelo:

- Análise dos coeficientes do modelo e dos valores preditos por meio de:
  - comparações com experimentos anteriores, quando existirem;
  - resultados de simulação.
- Coleta de novos dados para verificar o desempenho preditivo do modelo.
- Partição (*split*) dos dados, que consiste em deixar parte dos dados originais fora da análise para investigar o desempenho preditivo do modelo com a parte não utilizada no ajuste.

As técnicas de validação citadas são úteis seja para dados de experimentos não planejados, seja para dados de experimentos planejados. O planejamento é sempre importante por ajudar a minimizar problemas quanto, dentre outros, a:

- fatores importantes não serem deixados de lado;
- identificação apropriada da variação (*range*) dos fatores.

Em experimentos planejados é comum a inclusão de um conjunto extra de observações para a verificação do desempenho preditivo do modelo ajustado.

11. Regressão com parte das regressoras categóricas

Em uma análise de regressão, as regressoras são usualmente quantitativas. Em diversos estudos, contudo, não é incomum que algumas delas sejam qualitativas. Por exemplo: estação do ano, turno de trabalho, sexo etc. Uma regressora qualitativa não apresenta, em geral, uma escala de medida natural, mas sim níveis ou categorias. Para quantificar o efeito das categorias de uma regressora qualitativa na resposta Y é usual que regressoras dessa natureza sejam incorporadas aos modelos de regressão por meio de variáveis indicadoras, também denominadas de variáveis *dummy*.

Dentre as diversas maneiras de identificar quantitativamente as categorias de uma variável qualitativa, as variáveis indicadoras, que tomam valores 0 e 1, é uma delas. Como ilustração, considere um experimento em que se tenha a resposta Y e as regressoras X_1, de natureza contínua, e X_2, de natureza qualitativa com dois níveis ou categorias, nível 1 e nível 2. Para incorporar a regressora X_2 no modelo de regressão, de modo que o efeito na esperança de Y de cada um de seus níveis possa ser quantificado, define-se a seguinte variável indicadora ou *dummy*:

$$X_2 = \begin{cases} 0 \text{ se nível 1} \\ 1 \text{ se nível 2.} \end{cases}$$

Será visto adiante, que considerar a regressora X_2 desta forma no modelo de regressão, leva a interpretações simples dos parâmetros a ela associados. Em geral, uma regressora qualitativa com *m* níveis ou categorias é representada por *m*-1 variáveis indicadoras, cada qual tomando os valores 0 e 1.

Exemplo: Suponha que um engenheiro mecânico tem por interesse relacionar a vida efetiva de uma ferramenta de corte usada em um torno mecânico com a velocidade do torno em rpm (rotações por minuto) e com o tipo de ferramenta de corte utilizada (tipo A ou B). Os dados coletados estão apresentados a seguir:

Y vida efetiva (horas)	**X_1 velocidade (rpm)**	**X_2 tipo ferramenta**	**Y vida efetiva (horas)**	**X_1 velocidade (rpm)**	**X_2 tipo ferramenta**
18.73	610	A	30.16	670	B
14.52	950	A	27.09	770	B
17.43	720	A	25.40	880	B
14.54	840	A	26.05	1000	B
13.44	980	A	33.49	760	B
24.39	530	A	35.62	590	B
13.34	680	A	26.07	910	B
22.71	540	A	36.78	650	B
12.68	890	A	34.95	810	B
19.32	730	A	43.67	500	B

Fonte: Montgomery e Peck (1992)

Note, nesse estudo, que a regressora X_2, tipo de ferramenta, é qualitativa com dois níveis, A e B. Usando uma variável indicadora *(dummy)* para incorporá-la ao modelo, a fim de quantificar o efeito de seus níveis na esperança de Y, tem-se:

$$X_2 = \begin{cases} 0 \text{ se ferramenta do tipo A} \\ 1 \text{ se ferramenta do tipo B.} \end{cases}$$

Considerando o modelo $E(Y \mid \mathbf{x}) = \beta_o + \beta_1 x_1 + \beta_2 x_2$ segue que:

$$\text{para } x_2 = 0 \Rightarrow E(Y \mid \mathbf{x}) = \beta_o + \beta_1 x_1$$
$$\text{e, para } x_2 = 1 \Rightarrow E(Y \mid \mathbf{x}) = (\beta_o + \beta_2) + \beta_1 x_1.$$

Portanto, para o tipo de ferramenta A ($x_2 = 0$), a relação entre a vida efetiva desta ferramenta e a velocidade do torno é uma reta com intercepto β_o e inclinação β_1. Analogamente, para o tipo B, uma reta com intercepto $(\beta_o + \beta_2)$ e inclinação β_1. Tem-se, assim, duas retas paralelas, isto é, duas retas com inclinação comum β_1 e interceptos diferentes. Para x_1 fixo, o parâmetro β_2 expressa a mudança na esperança do tempo de vida, resultante da mudança da ferramenta do tipo A para a do tipo B. Intervalo de confiança e teste de hipóteses para β_2 são obtidos de forma análoga aos obtidos para os parâmetros de um modelo de regressão com todas as regressoras quantitativas.

O uso de uma variável indicadora para incoporar uma regressora qualitativa com dois níveis pode ser generalizada para regressoras qualitativas com mais do que dois níveis. Considerando, por exemplo, três tipos de ferramentas (A, B e C), seriam necessárias duas variáveis indicadoras para incorporar os três níveis no modelo. Essas variáveis seriam expressas por:

$$X_{21} = \begin{cases} 1 \text{ se ferramenta tipo A} \\ 0 \text{ em caso contrário} \end{cases} \quad \text{e} \quad X_{22} = \begin{cases} 1 \text{ se ferramenta tipo B} \\ 0 \text{ em caso contrário,} \end{cases}$$

em que para as combinações possíveis de X_{21} e X_{21} tem-se:

X_{21}	X_{22}	
1	0	Ferramenta tipo A
0	1	Ferramenta tipo B
0	0	Ferramenta tipo C.

O modelo de regressão, nesse caso, fica expresso por:

$$E(Y \mid \mathbf{x}) = \beta_o + \beta_1 x_1 + \beta_2 x_{21} + \beta_3 x_{22},$$

de modo que:

X_{21}	X_{22}	
1	0	$E(Y \mid \mathbf{x}) = (\beta_o + \beta_2) + \beta_1 x_1$
0	1	$E(Y \mid \mathbf{x}) = (\beta_o + \beta_3) + \beta_1 x_1$
0	0	$E(Y \mid \mathbf{x}) = \beta_o + \beta_1 x_1$

isto é, três retas paralelas com mesma inclinação e interceptos diferentes.

Nesse modelo, para x_1 fixo, o parâmetro β_2 expressa a mudança no tempo de vida médio resultante da mudança da ferramenta do tipo C para a do tipo A e, o parâmetro β_3, a mudança no tempo de vida médio resultante da mudança da ferramenta do tipo C para a do tipo B.

Uma questão em estudos como o das ferramentas é a de porque optar em ajustar um único modelo com variáveis *dummy* e não por um modelo de regressão linear para cada tipo de ferramenta. Algumas considerações nesse sentido indicam que, em geral, é preferível ajustar um único modelo por algumas razões. O analista tem somente uma equação final, o que é mais prático. Ainda, ajustar um único modelo produz uma estimativa comum da variância dos erros, bem como se tem mais graus de liberdade do que se teria no ajuste de dois ou mais modelos de regressão lineares separados. Além disso, se as retas forem de fato paralelas (mesma inclinação), faz sentido combinar os dados para produzir uma estimativa comum desse parâmetro. Se, no entanto, as retas não forem paralelas, a inclusão da interação entre as regressoras permite que esse fato seja analisado. Modelos com variáveis *dummy* e interações que as envolvam são discutidos a seguir.

11.1 Regressoras categóricas e interações

Considere, agora, que o modelo de regressão linear que relaciona o tempo de vida das ferramentas com a velocidade do torno e o tipo de ferramenta (A e B), produza retas em que tanto as inclinações quanto os interceptos sejam diferentes. Para modelar essa situação, basta considerar no modelo a interação entre as regressoras, ou seja,

.

$$E(Y|\mathbf{x}) = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

Observe que, como X_2 assume somente os valores 0 e 1 tem-se:

$$\text{para } x_2 = 0 \ (A) \Rightarrow E(Y|\mathbf{x}) = \beta_o + \beta_1 x_1$$
$$\text{e, para } x_2 = 1 \ (B) \Rightarrow E(Y|\mathbf{x}) = (\beta_o + \beta_2) + (\beta_1 + \beta_3) x_1$$

as quais são retas concorrentes. Nesse caso, o efeito na esperança de Y resultante da mudança da ferramenta do tipo A para a do tipo B é de $\beta_2 + \beta_3 x_1$ unidades. Ou seja, o efeito do tipo de ferramenta na esperança de Y depende da velocidade do torno (X_1).

Para verificar se o efeito do tipo de ferramenta na esperança de Y realmente depende da velocidade do torno, testa-se as hipóteses H_o:$\beta_3 = 0$ *versus* Ha: $\beta_3 \neq 0$. A dependência será afirmativa se H_o for rejeitada.

Para os dados do estudo das ferramentas será visto adiante qual dos dois modelos é o mais apropriado (com ou sem a interação).

11.2 Algumas considerações sobre variáveis *dummy*

O uso de variáveis *dummy* para representar regressoras qualitativas em um modelo de regressão é, em geral, mais indicado do que proceder a alocação de códigos quaisquer às categorias dessas variáveis. A dificuldade básica com a alocação de códigos é que eles definem uma métrica para os níveis da variável qualitativa, métrica esta que pode não ser razoável.

Por exemplo, alocar arbitrariamente os códigos 1, 2 e 3 para uma regressora qualitativa com três níveis, implica em assumir que a resposta muda, em média, a mesma quantidade quando se passa de um nível para outro e, isso, pode não estar de acordo com a realidade. Portanto, alocar códigos, igualmente ou não espaçados, aos níveis de uma regressora qualitativa equivale a assumir distâncias arbitrárias, porém definidas, entre os níveis.

Variáveis indicadoras, em contraste, não impõem qualquer métrica aos níveis da variável qualitativa. Elas dependem dos dados para mostrar os efeitos diferenciais que ocorrem entre os níveis.

Exemplo: Suponha um experimento em que tenha Y, a variável resposta (quantitativa) e as regressoras X_1 (quantitativa) e X_2 (qualitativa com três níveis: A, B e C). Se for considerado os códigos 1, 2 e 3 aos níveis de X_2 tem-se que:

$$X_2 = \begin{cases} 1 \text{ se nível A} \\ 2 \text{ se nível B} \\ 3 \text{ se nível C.} \end{cases}$$

Conseqüentemente, para o modelo $E(Y \mid \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ segue que:

$$\begin{aligned} &\text{para } x_2 = 1 \ (A) \Rightarrow E(Y \mid \mathbf{x}) = (\beta_0 + \beta_2) + \beta_1 x_1, \\ &\text{para } x_2 = 2 \ (B) \Rightarrow E(Y \mid \mathbf{x}) = (\beta_0 + 2\beta_2) + \beta_1 x_1, \\ &\text{para } x_2 = 3 \ (C) \Rightarrow E(Y \mid \mathbf{x}) = (\beta_0 + 3\beta_2) + \beta_1 x_1. \end{aligned}$$

Assim, para um valor fixo x_1, quando se muda, por exemplo, do nível A para o nível B, a mudança na resposta média será de β_2 unidades. Analogamente, de B para C, também igual a β_2 unidades. Então, usar os códigos 1, 2 e 3, impõe que a resposta média muda a mesma quantidade quando se muda do nível A para B ou de B para C. Isto pode não estar de acordo com a realidade, mas é resultado dos códigos 1, 2 e 3, que assumem distâncias iguais entre os três níveis.

Por outro lado, se forem usadas variáveis *dummy*, tal que:

$$X_{21} = \begin{cases} 1 \text{ se nível A} \\ 0 \text{ caso contrário} \end{cases} \quad \text{e} \quad X_{22} = \begin{cases} 1 \text{ se nível B} \\ 0 \text{ caso contrário,} \end{cases}$$

tem-se, para o modelo $E(Y \mid \mathbf{x}) = \beta_0 + \beta_1 X_1 + \beta_2 x_{21} + \beta_3 x_{22}$ que:

$$\begin{aligned} &\text{para } x_{21} = 1 \text{ e } x_{22} = 0 \ (A) \Rightarrow E(Y \mid \mathbf{x}) = (\beta_0 + \beta_2) + \beta_1 x_1, \\ &\text{para } x_{21} = 0 \text{ e } x_{22} = 1 \ (B) \Rightarrow E(Y \mid \mathbf{x}) = (\beta_0 + \beta_3) + \beta_1 x_1, \\ &\text{para } x_{21} = 0 \text{ e } x_{22} = 0 \ (C) \Rightarrow E(Y \mid \mathbf{x}) = \beta_0 + \beta_1 x_1. \end{aligned}$$

Logo, para um valor fixo x_1, mudar do nível A para o B implica em uma mudança na esperança de Y de $(\beta_3 - \beta_2)$ unidades. Analogamente, de C para A, em uma mudança ou efeito diferencial na esperança de Y de β_2 unidades e, finalmente, de C para B, em uma mudança na esperança de Y de β_3 unidades. Note, que não existem restrições arbitrárias entre os efeitos diferenciais e, desse modo, as variáveis *dummy* são preferíveis aos códigos alocados.

11.3 Outros códigos para variáveis indicadoras

Até o momento foi usado um esquema para codificar os níveis de uma regressora qualitativa em que $m - 1$ variáveis indicadoras do tipo 0 e 1 são criadas para representar seus respectivos m níveis. Outros esquemas possíveis são apresentados a seguir.

1º. esquema alternativo: usar variáveis indicadoras do tipo 1 e –1.

Nesse caso, para uma variável qualitativa com dois níveis, A e B, tem-se:

$$X_2 = \begin{cases} 1 \text{ se nível A} \\ -1 \text{ se nível B,} \end{cases}$$

e, para uma variável qualitativa com três níveis (A, B e C):

$$X_{21} = \begin{cases} 1 \text{ se nível A} \\ -1 \text{ se nível C} \\ 0 \text{ caso contrário} \end{cases} \quad \text{e} \quad X_{22} = \begin{cases} 1 \text{ se nível B} \\ -1 \text{ se nível C} \\ 0 \text{ cao contário,} \end{cases}$$

de modo que:

Nível	X_{21}	X_{22}
A	1	0
B	0	1
C	–1	–1

2º. esquema alternativo: outra alternativa é **retirar** o termo intercepto do modelo de regressão e usar m variáveis qualitativas do tipo 0 e 1 para representar cada um dos m níveis da variável. Assim, em um modelo em que a regressora X_1 é quantitativa, e a regressora X_2 qualitativa com dois níveis, A e B, tem-se o modelo

$$E(Y|\,\mathbf{x}) = \beta_1 X_1 + \beta_2 x_{21} + \beta_3 x_{22},$$

com: $X_{21} = \begin{cases} 1 \text{ se nível A} \\ 0 \text{ c. c.} \end{cases}$ e $X_{22} = \begin{cases} 1 \text{ se nível B} \\ 0 \text{ c. c.} \end{cases}$

Importante: não se pode comparar o coeficiente de determinação obtido ao se usar o modelo de regressão sem o intercepto ($R^2_{(0)}$) com o obtido ao se usar o modelo de regressão com o intercepto (R^2). No modelo sem intercepto, a variação descrita pelo numerador e denominador de $R^2_{(0)}$ representa a dispersão em torno de zero e, no modelo com intercepto, esta variação representa a dispersão em torno da média $\bar{y}$. Existe, assim, uma forte tendência do valor $R^2_{(0)}$ ser maior que R^2. Fato análogo é observado para a qualidade do ajuste que tende a ser superior no modelo sem o intercepto. Isto resulta do fato de que somas de quadrados não corrigidas são usadas e, mesmo com performances equivalentes, $R^2_{(0)}$ pode ser consideravelmente maior que R^2. A conseqüência é uma grave confusão quando, erroneamente, modelos sem interceptos são comparados com aqueles com interceptos. Os modelos sem intercepto serão classificados, erroneamente, em um *rank* de diversos modelos sendo comparados como os melhores.

Uma alternativa para o cálculo de $R^2_{(0)}$ de modo a se ter comparações razoáveis entre os modelos com e sem intercepto é dada por:

$$R^2_{(0)} = 1 - \left[\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \ / \ \sum_{i=1}^{n} (y_i - \bar{y})^2 \right],$$

em que: $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i)^2 - \left[\left(\sum_{i=1}^{n} (y_i x_i) \right)^2 / \sum_{i=1}^{n} (x_i)^2 \right].$

Para mais detalhes ver, dentre outros, Myers (1986), Casella (1986) e Hahn (1977).

11.4 Exemplo de regressão com variável *dummy*

A seguir, é apresentada a análise e discussão dos dados do tempo de vida de ferramentas descrito na Seção 11.

1ª. análise: considerando variável *dummy* do tipo 0 e 1 para representar X_2

Considerando o modelo $E(Y|\ \mathbf{x}) = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ em que:

$$X_2 = \begin{cases} 0 \text{ se ferramenta A} \\ 1 \text{ se ferramenta B,} \end{cases}$$

foram obtidos os resultados apresentados no Quadro 5, que mostra evidências de que a interação entre as regressoras não é significativa (p-valor = 0,1955).

Quadro 5. Análise de variância do modelo com a interação entre X_1 e X_2.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	293.01	293.01	33.2545	2.889e-05
X2B	1	1125.03	1125.03	127.6847	4.891e-09
X1:X2	1	16.08	16.08	1.8248	0.1955
Residuals	16	140.98	8.81		

Excluída a interação, obtiveram-se os resultados apresentados no Quadro 6, podendo-se concluir que as regressoras X_1 e X_2 são significativas (p-valores < 0,001) e devem permanecer no modelo. As estimativas para esse modelo e alguns outros resultados encontram-se no Quadro 7. Alguns gráficos dos resíduos são mostrados na Figura 7.

Quadro 6. Análise de variância do modelo sem interação entre X_1 e X_2.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	293.01	293.01	31.716	2.990e-05
X2B	1	1125.03	1125.03	121.776	3.587e-09
Residuals	17	157.05	9.24		

Quadro 7. Estimativas dos coeficientes e outros resultados relevantes.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	36.98560	3.51038	10.536	7.16e-09
X1	-0.02661	0.00452	-5.887	1.79e-05
X2B	15.00425	1.35967	11.035	3.59e-09

Residual standard error: 3.039 on 17 degrees of freedom
Multiple R-Squared: 0.9003, Adjusted R-squared: 0.8886
F-statistic: 76.75 on 2 and 17 DF, p-value: 3.086e-09

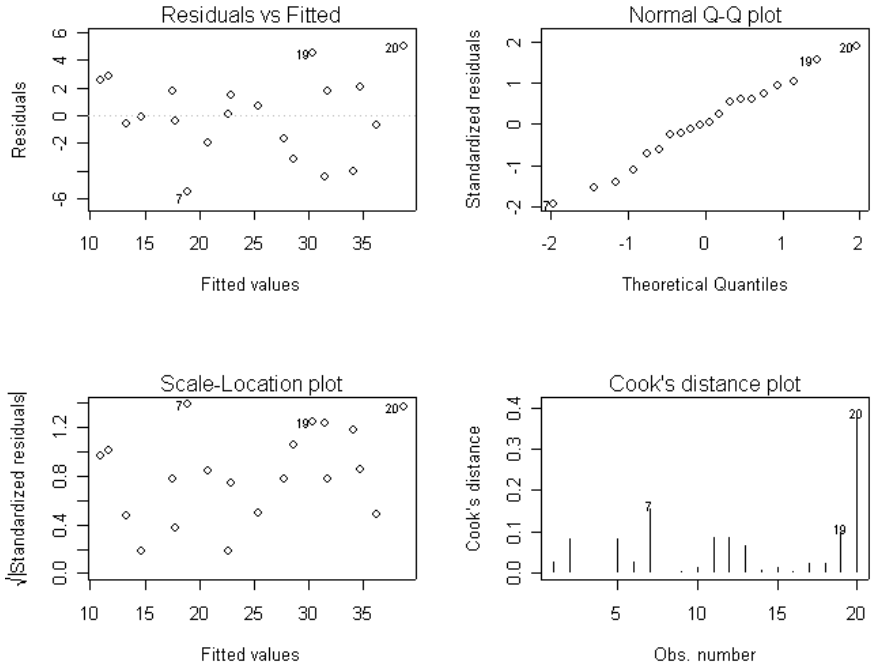


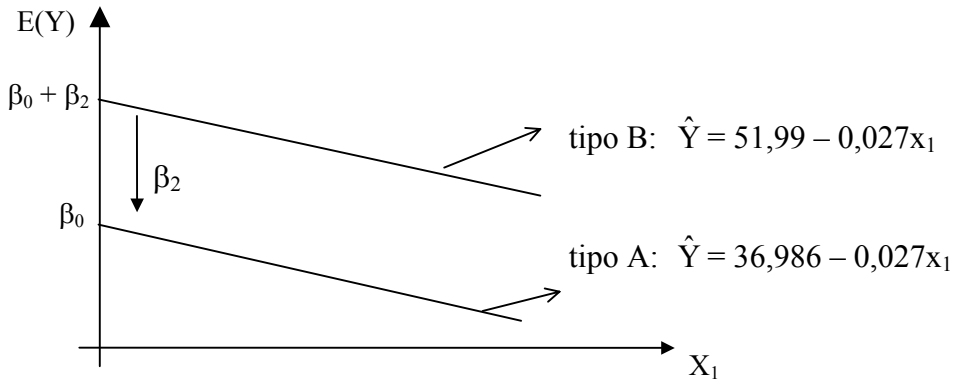
Figura 7. Análise dos resíduos do modelo Y em X_1 e X_2 (sem a interação).

Dos resultados apresentados, é possível observar que a análise de variância, bem como a análise de resíduos do modelo ajustado expresso por:

$$\hat{Y} = 36{,}986 - 0{,}027x_1 + 15{,}004x_2$$

apresentaram-se satisfatórias. O parâmetro β_2, cuja estimativa é 15,004, indica que para um valor fixo da velocidade do torno, há um aumento no tempo de vida médio de em torno 15 horas ao se trocar a ferramenta do tipo A para a do tipo B. Uma estimativa intervalar para β_2 com 95% de confiança resultou em (12,135; 17,873) horas. Espera-se, desse modo, com esse nível de confiança, que esse aumento no tempo de vida médio esteja entre 12,135 e 17,873 horas.

As retas para as ferramentas A e B apresentam, portanto, mesma inclinação e interceptos diferentes (retas paralelas), e podem ser visualizadas no gráfico a seguir.



2ª. análise: considerando variável *dummy* do tipo 1 e –1 para representar X_2, ou seja:

$$X_2 = \begin{cases} 1 \text{ se ferramenta A} \\ -1 \text{ se ferramenta B.} \end{cases}$$

Foi considerado, inicialmente, o modelo com a interação entre X_1 e X_2 mas, similar à 1ª. análise, a mesma não apresentou significância estatística. Ajustando-se então o modelo sem a interação, foram obtidos os resultados mostrados nos Quadros 8 e 9 e Figura 8.

Quadro 8. Análise de variância do modelo sem a interação.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	293.01	293.01	31.716	2.990e-05
X2	1	1125.03	1125.03	121.776	3.587e-09
Residuals	17	157.05	9.24		

Quadro 9. Estimativas dos coeficientes e outros resultados relevantes.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	44.48773	3.45947	12.860	3.47e-10
X1	-0.02661	0.00452	-5.887	1.79e-05
X2	-7.50213	0.67983	-11.035	3.59e-09

```
Residual standard error: 3.039 on 17 degrees of freedom
Multiple R-Squared: 0.9003,     Adjusted R-squared: 0.8886
F-statistic: 76.75 on 2 and 17 DF,  p-value: 3.086e-09
```

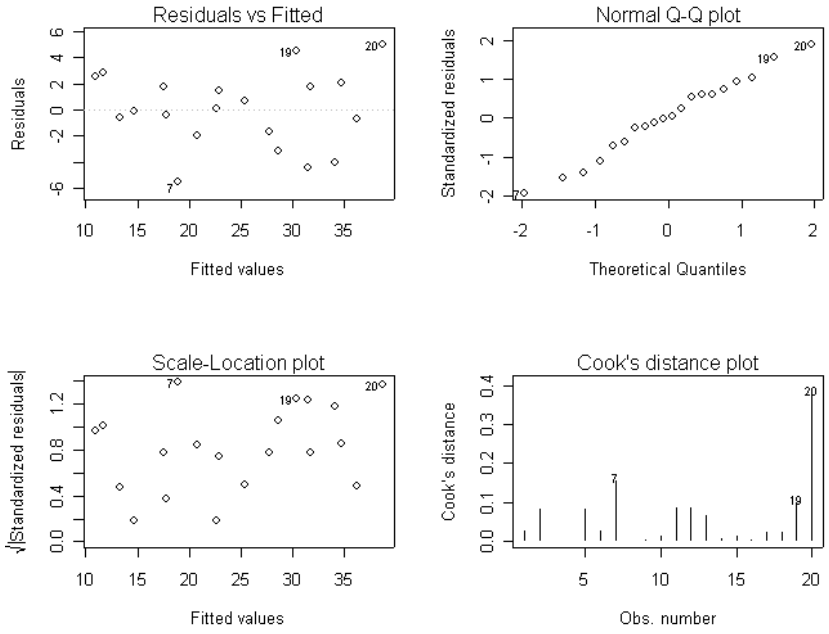
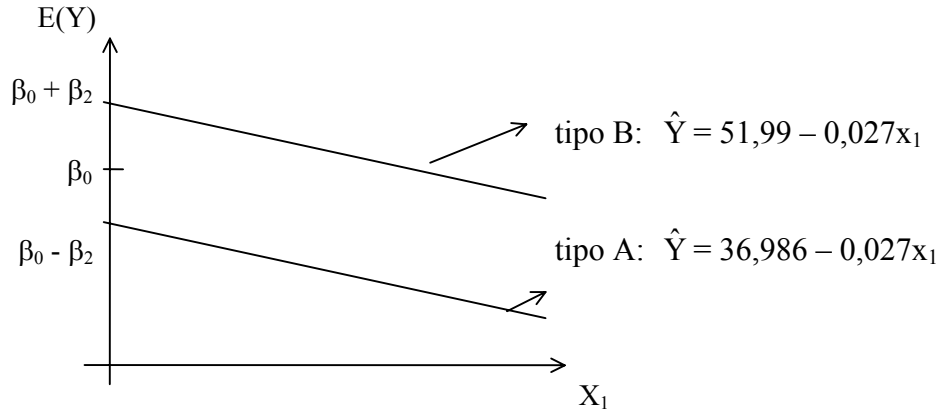


Figura 8. Análise dos resíduos do modelo sem interação.

Observe, dos resultados apresentados, que a análise de variância do modelo, bem como a análise de resíduos proporcionam resultados iguais aos da 1ª. análise e, desse modo, são satisfatórios. O modelo ajustado é, nesse caso, expressso por:

$$\hat{Y} = 44{,}488 - 0{,}027x_1 - 7{,}502x_2,$$

em que o parâmetro β_0 (intercepto) é, nesse modelo, a média dos interceptos das duas linhas de regressão na qual, para um valor fixo x_1, a ferramenta A e B diferem por β_2 unidades em direções opostas. Os modelos para as ferramentas A e B são mostrados no gráfico a seguir.



Note, que os modelos para as ferramentas A e B são exatamente os mesmos encontrados na 1ª análise. A única diferença é que na 1ª análise um dos níveis da variável qualitativa é tido como referência e, portanto, todos os demais níveis são comparados com ele. Já na 2ª análise, a referência é a média dos níveis da variável qualitativa e, sendo assim, as comparações são feitas em relação ao tempo de vida médio das ferramentas A e B e não em relação ao tempo de vida médio da ferramenta A, como na 1ª análise.

3ª. análise: considerando variáveis dummy do tipo 0 e 1 e modelo sem intercepto.

Nesse caso, tem-se:

$$X_{21} = \begin{cases} 1 & \text{se ferramenta A} \\ 0 & \text{c. c.} \end{cases} \quad \text{e} \quad X_{22} = \begin{cases} 1 & \text{se ferramenta B} \\ 0 & \text{c. c.,} \end{cases}$$

de modo que os resultados do modelo sem a interação, também não significativa, são mostrados nos Quadros 10 e 11 e Figura 9.

Quadro 10. Análise de variância do modelo sem interação e sem intercepto.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	10847.3	10847.3	1174.142	< 2.2e-16
X21	1	602.6	602.6	65.228	3.206e-07
X22	1	1991.7	1991.7	215.590	4.343e-11
Residuals	17	157.1	9.2		

Quadro 11. Estimativas dos coeficientes e outros resultados relevantes.

```
Coefficients:
    Estimate Std. Error t value Pr(>|t|)
X1  -0.02661    0.00452  -5.887 1.79e-05
X21 36.98560    3.51038  10.536 7.16e-09
X22 51.98985    3.54082  14.683 4.34e-11

Residual standard error: 3.039 on 17 degrees of freedom
Multiple R-Squared: 0.9885,     Adjusted R-squared: 0.9864
F-statistic:   485 on 3 and 17 DF,  p-value: < 2.2e-16
```

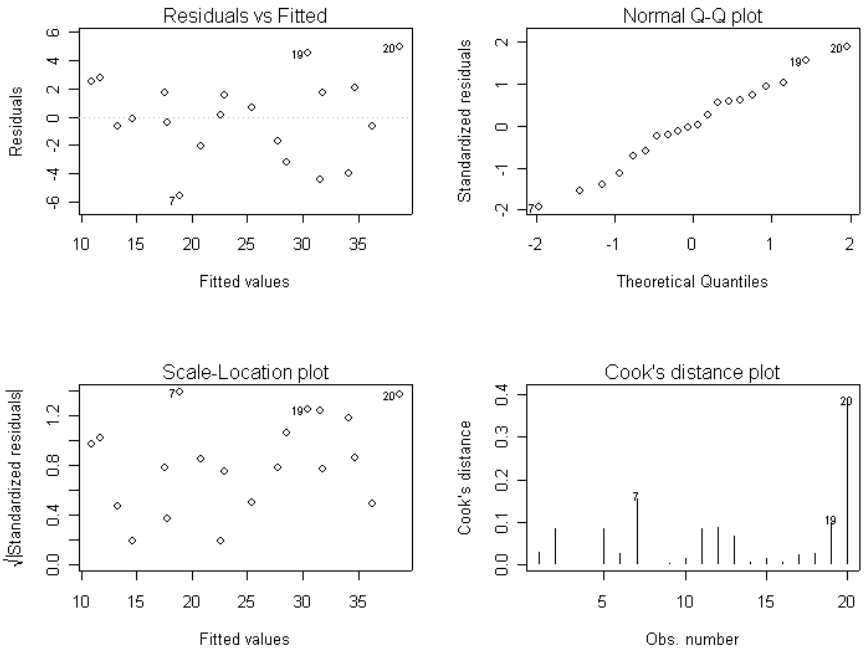


Figura 9. Análise dos resíduos do modelo sem interação e sem intercepto.

Observe, dos resultados, e como chamado a atenção anteriormente, que a análise de variância e, conseqüentemente, o valor de R^2 não são iguais aos obtidos nas análises anteriores. Para que o R^2 desse modelo possa ser comparado aos das duas análises anteriores deve-se calcular $R^2_{(o)}$ pois, em caso contrário, se terá a falsa impressão de que este modelo é melhor do que os demais, quando na realidade é eqüivalente a eles e produzem as mesmas conclusões.

O modelo ajustado correspondente a análise realizada é expresso por:

$$\hat{Y} = -0.027x_1 + 36{,}985\ x_{21} + 51{,}989\ x_{22}$$

e resulta em duas retas de regressão iguais às encontradas anteriormente. Note, para uma velocidade fixa x_1 do torno, que a diferença (51,989 – 36,985) = 15,004, representa, como nas análises anteriores, o quanto a esperança de Y muda ao se mudar da ferramenta A para a B.

12. Regressão Polinomial

O modelo de regressão polinomial é um caso especial do modelo de regressão linear geral $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, em que uma, ou mais regressoras, podem estar presentes no modelo em diversas potencias. É usual, nesses modelos, que as regressoras sejam expressas como o desvio de suas respectivas médias, isto porque nos modelos polinomiais termos como X e X^2, apresentam naturalmente alta correlacão. Expressar as regressoras como o desvio de sua média pode, em muitos casos, amenizar os problemas decorrentes da presença de multicolineridade nos modelos polinomiais. Em outros casos, mesmo centrando as regressoras na média, pode-se continuar tendo termos altamente correlacionados. Nesses casos, os *polinômios ortogonais*, descritos a seguir, podem ser úteis.

12.1 Polinômios ortogonais com uma regressora

Considere um modelo polinomial com uma única regressora X expresso por $E(\mathbf{Y}|\ \mathbf{x}) = \beta_o + \beta_1 x + \beta_2 x^2 + .. + \beta_r x^r$. Para esse modelo, as colunas da matriz $\mathbf{X}$ são, em geral, não ortogonais. Ainda, aumentar a ordem do polinômio por adicionar um termo $\beta_{r+1}x^{r+1}$, implica em recalcular a inversa $(\mathbf{X}'\mathbf{X})^{-1}$ obtendo-se estimativas dos parâmetros de ordem menor diferentes das obtidas no modelo sem esse termo. Se, no entanto, for ajustado o modelo:

$$E(Y|\ x_i) = \alpha_0 P_0(x_i) + \alpha_1 P_1(x_i) + \alpha_2 P_2(x_i) + ... + \alpha_r P(x_i), \qquad i = 1, ...n,$$

sendo $P_u(x_i)$ a u-ésima ordem do polinômio ortogonal definido de tal modo que,

$$\begin{cases} \sum_{i=1}^{n} P_q(x_i) P_s(x_i) = 0 & (q \neq s;\ q, s = 0, 1, .., r) \\ P_o(x_i) = 1, \end{cases}$$

o modelo passa a ser representado por $E(\mathbf{Y}|\ \mathbf{x}) = \mathbf{X}\boldsymbol{\alpha}$, em que a matriz $\mathbf{X}$ é composta de colunas ortogonais, dada por:

$$\mathbf{X} = \begin{bmatrix} P_o(x_1) & P_1(x_1) & \cdots P_r(x_1) \\ P_o(x_2) & P_1(x_2) & \cdots P_r(x_2) \\ \vdots & \vdots & \vdots \quad \vdots \\ P_o(x_n) & P_1(x_n) & \cdots P_r(x_n) \end{bmatrix}$$

e, portanto, tem-se:

$$\mathbf{X'X} = \begin{bmatrix} \sum_{i=1}^{n} P_o^2(x_i) & 0 & \cdots & 0 \\ 0 & \sum_{i=1}^{n} P_1^2(x_i) & \cdots & 0 \\ \vdots & \vdots & \vdots & 0 \\ 0 & 0 & \cdots & \sum_{i=1}^{n} P_r^2(x_i) \end{bmatrix}.$$

Os cinco primeiros polinomiais para o caso dos níveis de X serem igualmente espaçados são dados por:

$$P_o(x_i) = 1$$

$$P_1(x_i) = \lambda_1 \left[\frac{x_i - \bar{x}}{d} \right]$$

$$P_2(x_i) = \lambda_2 \left[\left(\frac{x_i - \bar{x}}{d} \right)^2 - \left(\frac{n^2 - 1}{12} \right) \right]$$

$$P_3(x_i) = \lambda_3 \left[\left(\frac{x_i - \bar{x}}{d} \right)^3 - \left(\frac{x_i - \bar{x}}{d} \right) \left(\frac{3n^2 - 7}{20} \right) \right]$$

$$P_4(x_i) = \lambda_4 \left[\left(\frac{x_i - \bar{x}}{d} \right)^4 - \left(\frac{x_i - \bar{x}}{d} \right)^2 \left(\frac{3n^2 - 13}{14} \right) + \frac{3(n^2 - 1)(n^2 - 9)}{560} \right]$$

sendo *d* o espaçamento entre os níveis de X e λ_1, ..., λ_4 constantes escolhidas de modo aos polinomiais apresentarem valores inteiros.

Os estimadores de MQO de α são encontrados por:

$$\hat{\boldsymbol{\alpha}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \text{ e, portanto, } \hat{\alpha}_j = \frac{\sum_{i=1}^{n} P_j(x_i) y_i}{\sum_{i=1}^{n} P_j^2(x_i)} \text{ para } j = 0, 1, .., r.$$

Como $P_o(x_i) = 1$, segue que $\hat{\alpha}_0 = \bar{y}$. Tem-se, ainda:

$$SQres(r) = \mathbf{Y'Y} - \sum_{j=1}^{r} \hat{\alpha}_j \left[\sum_{i=1}^{n} P_j(x_i) y_i \right],$$

e a soma de quadrados da regressão para qualquer parâmetro do modelo, a qual não depende dos outros parâmetros no modelo, dada por:

$$SQreg(\alpha_j) = \hat{\alpha}_j \sum_{i=1}^{n} P_j(x_i) y_i .$$

Para testar a significância do termo de ordem mais alta, isto é, testar H_o: $\alpha_r = 0$, usa-se a estatística de teste F dada por:

$$F_0 = \frac{SQreg(\alpha_r)}{SQres(r)/(n - r - 1)} .$$

Note, que se a ordem do polinômio é alterada para $r + q$, somente os q novos coeficientes precisam ser calculados, uma vez que os r coeficientes no modelo não mudam devido a propriedade de ortogonalidade dos polinomiais. Assim, um ajuste sequencial do modelo é computacionalmente fácil.

Polinômios ortogonais também podem ser construídos e usados nos casos em que os valores de X não forem igualmente espaçados. A função *poly* disponível no *software* R, por exemplo, pode ser usada para obtenção de polinômios ortogonais em que se tenha valores de X igualmente espaçados, ou não. Em tal função, os polinomiais $P_j(x_i)$ são obtidos pelo procedimento de ortogonalização de Gram-Schmidt.

Exemplo: Considere os dados a seguir em que se tem uma variável resposta Y e uma única regressora X, ambas contínuas.

Observação	Y	X	$X_b = (X - \overline{X})$	X_b^2
1	6.3	1.0	-6.263	39.228
2	11.1	1.5	-5.763	33.214
3	20.0	2.0	-5.263	27.701
4	24.0	3.0	-4.263	18.175
5	26.1	4.0	-3.263	10.648
6	30.0	4.5	-2.763	7.635
7	33.8	5.0	-2.263	5.122
8	34.0	5.5	-1.763	3.109
9	38.1	6.0	-1.263	1.596
10	39.9	6.5	-0.763	0.582
11	42.0	7.0	-0.263	0.069
12	46.1	8.0	0.737	0.543
13	53.1	9.0	1.737	3.016
14	52.0	10.0	2.737	7.490
15	52.5	11.0	3.737	13.964
16	48.0	12.0	4.737	22.437
17	42.8	13.0	5.737	32.911
18	27.8	14.0	6.737	45.384
19	21.9	15.0	7.737	59.858

Observe, a partir do diagrama de dispersão mostrado na Figura 10, que uma relação quadrática entre Y e X parece ser bastante apropriada. Este fato sugere o ajuste de um modelo polinomial de 2ª ordem em X.

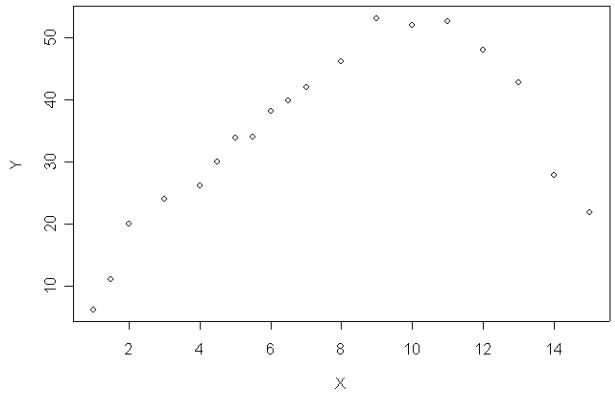


Figura 10. Diagrama de dispersão de Y versus X

Aos dados desse estudo, observa-se, contudo, uma alta correlação entre X e X^2 (r = 0,9703), o que é muito comum nesses casos. O VIF também resultou em 17,12 evidenciando que os parâmetros podem ser pobremente estimados se X e X^2 forem usadas conjuntamente no modelo. Se, contudo, for considerada a variável X centrada em sua média, isto é, $X_b = (X - \overline{X})$, tem-se uma correlação entre X_b e $(X_b)^2$ de 0,2974, bem como VIF = 1,09. Sendo assim, é recomendável o uso de X_b e $(X_b)^2$ no modelo, a fim de amenizar os efeitos causados pela colinearidade observada entre X e X^2.

Observe, a partir da Figura 11, que centrar a variável X em sua média, não altera a relação quadrática entre as variáveis Y e X. Há apenas um deslocamento dos dados no eixo x.

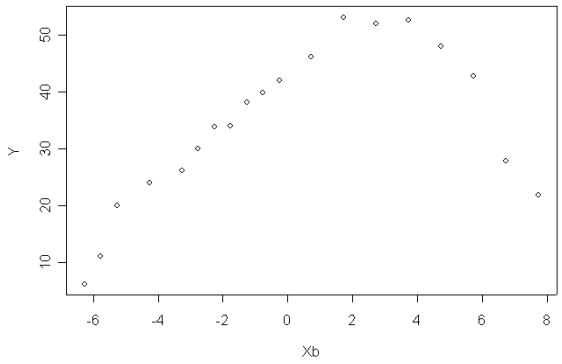


Figura 11. Diagrama de dispersão de Y versus X_b.

Considerando, então, o modelo de regressão polinomial Y em X_b e X_b^2 tem-se os resultados apresentados nos Quadros 12 e 13 e Figura 12.

Quadro 12: Análise de variância do modelo de regressão polinomial.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Xb	1	1043.43	1043.43	53.399	1.759e-06
Xb2	1	2060.81	2060.81	105.466	1.895e-08
Residuals	16	312.64	19.54		

Quadro 13: Estimativas dos coeficientes e outros resultados relevantes.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	45.29449	1.48287	30.55	1.29e-15
Xb	2.54629	0.25384	10.03	2.63e-08
Xb2	-0.63455	0.06179	-10.27	1.89e-08

```
Residual standard error: 4.42 on 16 degrees of freedom
Multiple R-Squared: 0.9085,     Adjusted R-squared: 0.8971
F-statistic: 79.43 on 2 and 16 DF,  p-value: 4.913e-09
```

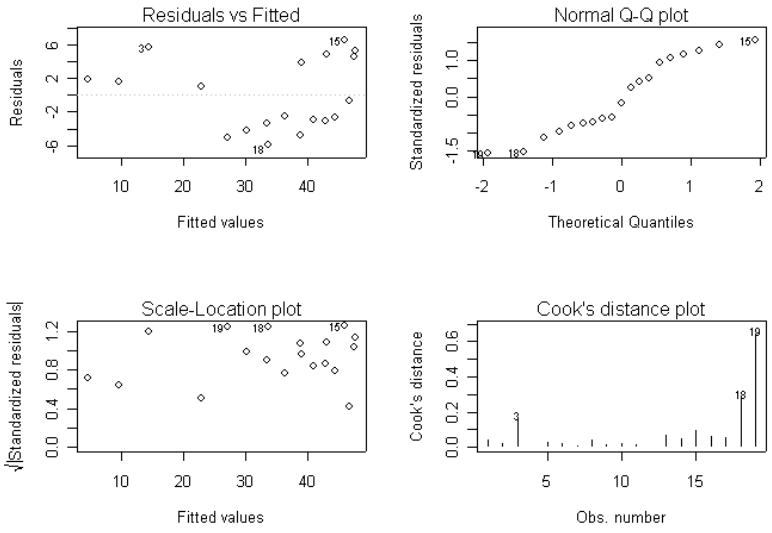


Figura 12. Análise de resíduos do modelo polinomial de 2ª ordem.

A partir dos resultados apresentados, é possível observar que o coeficiente de determinação mostrou-se satisfatório (R^2 = 0,9085). A análise de resíduos também mostrou-se razoável, tendo em vista o tamanho amostral ser relativamente pequeno. Os resultados do diagnóstico de influência (não mostrado) indicou as observações 1, 2 e 19 como merecedoras de serem investigadas junto ao pesquisador.

O modelo ajustado aos dados desse estudo, expresso por:

$$\hat{Y} = 45{,}295 + 2{,}546\,(x - \overline{x}) - 0{,}635\,(x - \overline{x})^2$$

encontra-se representado, juntamente com os valores observados, na Figura 13.

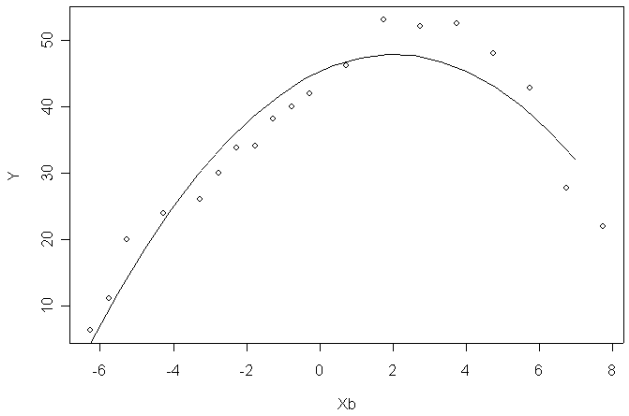


Figura 13. Dados observados e modelo polinomial ajustado.

Exemplo: Considere, agora, os dados a seguir em que Y é o custo anual médio de manutenção de um equipamento e X a produção do equipamento.

Y = custo anual médio (em dólares)	X = produção (em unidades)	$P_1(x_i)$	$P_2(x_i)$
335	50	-9	6
326	75	-7	2
316	100	-5	-1
313	125	-3	-3
311	150	-1	-4
314	175	1	-4
318	200	3	-3
328	225	5	-1
337	250	7	2
345	275	9	6

Fonte: Montgomery e Peck (1992)

A partir do diagrama de dipersão, mostrado na Figura 14, pode-se observar uma relação quadrática entre Y e X, o que sugere o ajuste de um modelo polinomial de 2ª ordem. Tem-se, ainda, uma alta correlação entre que X e X^2 (r = 0,9815) e VIF igual a 27,406, o que indica possíveis problemas na estimação dos parâmetros devido a presença de colinearidade entre essas variáveis. Por esse fato, bem como por se ter os níveis de X igualmente espaçados, será feito uso dos polinômios ortogonais $P_1(x_i)$ e $P_2(x_i)$ apresentados na Seção 12.1. Os valores desses polinomias encontram-se na tabela acima em que foram usados $\lambda_1 = 2$, $\lambda_2 = ½$, d = 25 e $\overline{x}$=162,5. Observe que a correlação entre $P_1(x_i)$ e $P_2(x_i)$ é igual a zero, uma vez que os mesmos foram obtidos de modo a serem ortogonais.

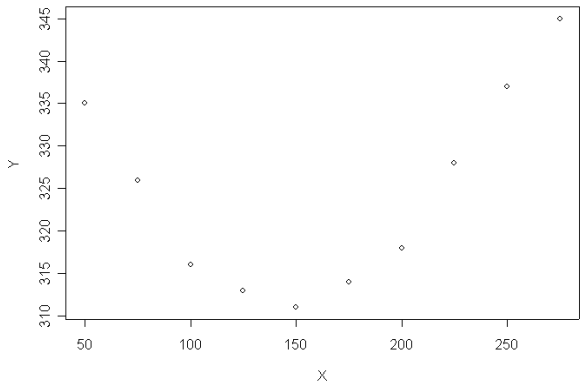


Figura 14. Diagrama de dispersão de Y versus X

Usando-se, então, os polinômios ortogonais $P_1(x_i)$ e $P_2(x_i)$ obtiveram-se os resultados mostrados nos Quadros 14 e 15 e Figura 15.

Quadro 14. Análise de variância do modelo de regressão polinomial.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
P1	1	181.89	181.89	47.717	0.0002297	***
P2	1	1031.52	1031.52	270.606	7.483e-07	***
Residuals	7	26.68	3.81			

Quadro 15. Estimativas dos coeficientes e outros resultados relevantes.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	324.3000	0.6174	525.262	< 2e-16	***
P1	0.7424	0.1075	6.908	0.000230	***
P2	2.7955	0.1699	16.450	7.48e-07	***

Residual standard error: 1.952 on 7 degrees of freedom
Multiple R-Squared: 0.9785, Adjusted R-squared: 0.9723
F-statistic: 159.2 on 2 and 7 DF, p-value: 1.461e-06

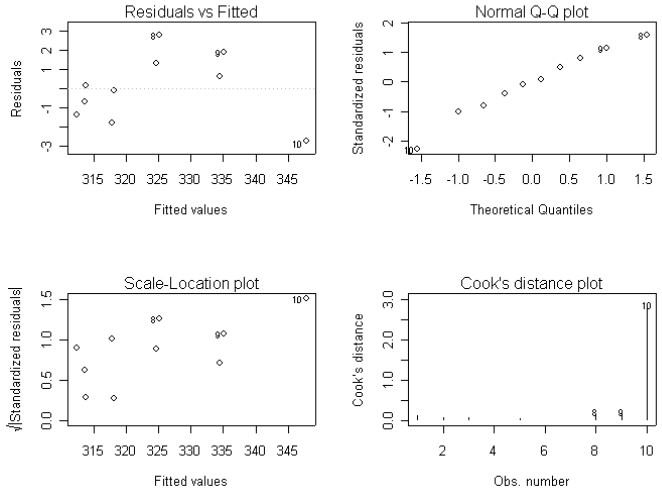


Figura 15. Análise de resíduos do modelo polinomial de 2ª ordem.

A análise de variância apresentada no Quadro 14 mostra que os termos linear e quadrático contribuem significativamente para o modelo de regressão. Os mesmos, conjuntamente, explicam 97,85% da variabilidade total de Y.

A análise de resíduos (Figura 15) apresentou-se razoável, levando-se em conta o tamanho amostral (n = 10) nesse estudo. O diagnóstico de pontos influentes mostrou que a observação 10 merece investigação adicional.

O modelo ajustado em termos dos polinôminos ortogonais, bem como em termos da regressora original, são expressos, respectivamente, por:

$$\hat{Y} = 324{,}30 + 0{,}7424\ P_1(x) + 2{,}7955\ P_2(x)$$

e

$$\hat{Y} = 324{,}30 + 0{,}7424\ (2)\left(\frac{x - 162{,}50}{25}\right) + 2{,}7955\ (1/2)\left[\left(\frac{x - 162{,}50}{25}\right) - \left(\frac{10^2 - 1}{12}\right)\right]$$

$$= 312{,}7686 + 0{,}0595(x - 162{,}50) + 0{,}0022\ (x - 162{,}50)^2.$$

Os valores observados e o modelo ajustado para os dados desse exemplo podem ser visualizados na Figura 16. Estes mostram que o modelo apresenta um bom ajuste aos dados.

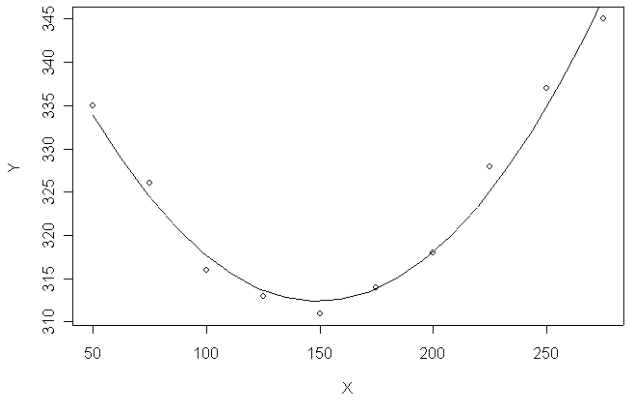


Figura 16. Dados observados e polinomial ajustado.

12.2. Regressão polinomial com mais de uma regressora

Os dados do exemplo analisado a seguir referem-se ao diâmetro (*girth*), altura (*height*) e volume de madeira, em pés cubicos (*cubic ft)*, de 31 árvores *black cherry* que foram derrubadas. Os diâmetros das árvores foram medidos entre 4 e 6 pés (*ft*) do chão e são fornecidos em polegadas (*inch*). Os dados são apresentados a seguir.

	Girth	Height	Volume		Girth	Height	Volume
1	8.3	70	10.3	17	12.9	85	33.8
2	8.6	65	10.3	18	13.3	86	27.4
3	8.8	63	10.2	19	13.7	71	25.7
4	10.5	72	16.4	20	13.8	64	24.9
5	10.7	81	18.8	21	14.0	78	34.5
6	10.8	83	19.7	22	14.2	80	31.7
7	11.0	66	15.6	23	14.5	74	36.3
8	11.0	75	18.2	24	16.0	72	38.3
9	11.1	80	22.6	25	16.3	77	42.6
10	11.2	75	19.9	26	17.3	81	55.4
11	11.3	79	24.2	27	17.5	82	55.7
12	11.4	76	21.0	28	17.9	80	58.3
13	11.4	76	21.4	29	18.0	80	51.5
14	11.7	69	21.3	30	18.0	80	51.0
15	12.0	75	19.1	31	20.6	87	77.0
16	12.9	74	22.2				

Para seis modelos ajustados a esses dados obtiveram-se, no *R*, os resultados:

```
Modelo 1  lm(formula = Volume ~ Height )

Residuals:
    Min      1Q  Median      3Q     Max
-21.274  -9.894  -2.894  12.067  29.852

            Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.1236    29.2731  -2.976 0.005835
Height        1.5433     0.3839   4.021 0.000378

Residual standard error: 13.4 on 29 degrees of freedom
Multiple R-Squared: 0.3579,      Adjusted R-squared: 0.3358
F-statistic: 16.16 on 1 and 29 DF,  p-value: 0.0003784

Modelo 2. lm(formula = Volume ~ Girth)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0654 -3.1067  0.1520  3.4948  9.5868

            Estimate Std. Error t value Pr(>|t|)
(Intercept) -36.9435     3.3651  -10.98 7.62e-12
Girth         5.0659     0.2474   20.48  < 2e-16

Residual standard error: 4.252 on 29 degrees of freedom
Multiple R-Squared: 0.9353,      Adjusted R-squared: 0.9331
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16

Modelo 3.  lm(formula = Volume ~ Height + Girth)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Height        0.3393     0.1302   2.607   0.0145 *
Girth         4.7082     0.2643  17.816  < 2e-16 ***

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-Squared: 0.948, Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16

Modelo 4.  lm(formula = Volume ~ Height + Girth + I(Girth^2))

Residuals:
    Min      1Q  Median      3Q     Max
-4.2928 -1.6693 -0.1018  1.7851  4.3489

            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.92041   10.07911  -0.984 0.333729
Height       0.37639    0.08823   4.266 0.000218
Girth       -2.88508    1.30985  -2.203 0.036343
I(Girth^2)   0.26862    0.04590   5.852 3.13e-06

Residual standard error: 2.625 on 27 degrees of freedom
Multiple R-Squared: 0.9771,      Adjusted R-squared: 0.9745
F-statistic: 383.2 on 3 and 27 DF,  p-value: < 2.2e-16

Modelo 5.   lm(formula = Volume ~ Girth + I(Girth^2) + I(Girth^3))

Residuals:
    Min      1Q  Median      3Q     Max
-5.4462 -2.3220 -0.4896  2.0225  7.4458

             Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.51838   39.47162  -0.596    0.556
Girth         5.78006    8.83752   0.654    0.519
I(Girth^2)   -0.32563    0.64248  -0.507    0.616
I(Girth^3)    0.01374    0.01515   0.907    0.373

Residual standard error: 3.345 on 27 degrees of freedom
Multiple R-Squared: 0.9627,      Adjusted R-squared: 0.9586
F-statistic: 232.4 on 3 and 27 DF,  p-value: < 2.2e-16
```

```
Modelo 6. lm(formula = Volume ~ Height+I(Height^2)+Girth+I(Girth^2))
Residuals:
    Min      1Q  Median      3Q     Max
-4.3679 -1.6698 -0.1580  1.7915  4.3581

             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.955101  63.013630  -0.015    0.988
Height       0.119372   1.784588   0.067    0.947
I(Height^2)  0.001717   0.011905   0.144    0.886
Girth       -2.796569   1.468677  -1.904    0.068
I(Girth^2)   0.265446   0.051689   5.135 2.35e-05

Residual standard error: 2.674 on 26 degrees of freedom
Multiple R-Squared: 0.9771,      Adjusted R-squared: 0.9735
F-statistic:   277 on 4 and 26 DF,  p-value: < 2.2e-16
```

Analisando os resultados obtidos para os seis modelos ajustados, tem-se o modelo 4 como sendo aparentemente o mais adequado. Análise dos VIF's associados a esse modelo revelam, contudo, valores maiores que 10. Quando a variável *Girth* é, contudo, centrada em sua média (G_b = *Girth* - $\overline{G}$), todos os VIF's, como pode ser observado a seguir, fornecem valores menores que 10.

Regressora	Height	Girth	$(Girth)^2$
VIF	1.3763	73.5750	72.4690

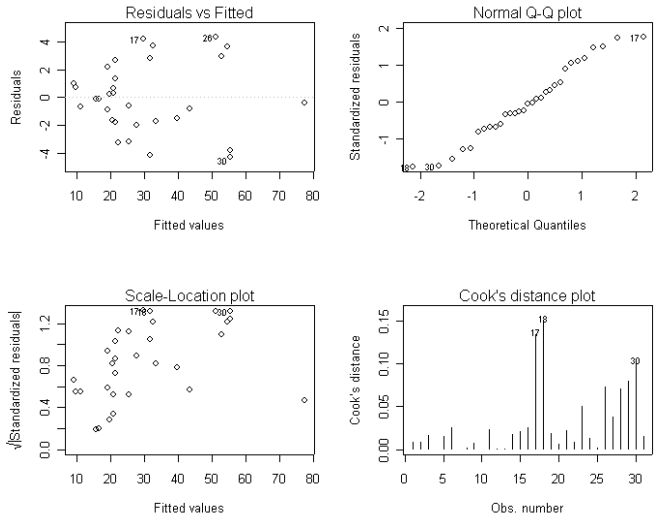
Regressora	Height	G_b	$(G_b)^2$
VIF	1.3763	1.6525	1.2438

Os principais resultados para o modelo 4 com a variável *Girth* centrada em sua média, são apresentados a seguir e mostram que este modelo apresenta, em geral, um ajuste satisfatório as dados desse estudo.

```
Modelo Final. lm(Volume~Height+Gb+I(Gb^2),data=trees)
Residuals:
    Min      1Q  Median      3Q     Max
-4.2928 -1.6693 -0.1018  1.7851  4.3489

            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.99450    6.76765  -0.147 0.884264
Height       0.37639    0.08823   4.266 0.000218 ***
Gb           4.23255    0.19630  21.561  < 2e-16 ***
I(Gb^2)      0.26862    0.04590   5.852 3.13e-06 ***

Residual standard error: 2.625 on 27 degrees of freedom
Multiple R-Squared: 0.9771,     Adjusted R-squared: 0.9745
F-statistic: 383.2 on 3 and 27 DF,  p-value: < 2.2e-16
```



BIBLIOGRAFIA SUGERIDA

[1] NETER, J.; WASSERMAN, W.; KUTNER, M.H**.** *Applied Linear Statistical Models*. 3ª. ed**.** Illinois: Richard D. Irwin, Inc., 1990.

[2] MONTGOMERY, D. C.; PECK, E. A. *Introduction to linear Regression Analysis*. New York: John Wiley & Sons, 2ª. ed., 1992.

[3] BELSLEY, D.A.; KUH, E.; WELSH, R.E. *Regression Diagnostics***.** New York: John Wiley & Sons, 1980.

[4] DRAPER, N.R.; SMITH, H. *Applied Regression Analysis*. New York: John Wiley & Sons, 2ª. ed., 1981.

[5] WONNACOTT, T.H.; WONNACOTT, R.J. *Introductory Statistics for Business and Economics.* New York: John Wiley & Sons, 1972.

[6] HOFFMANN, R.; VIEIRA, S. *Análise de Regressão*. São Paulo: Atual, 2ª. ed., 1977.

[7] FREIRE, C. L.; CHARNET, E.M.R.; BONVINO, H.; CHARNET, R. *Análise de Regressão Linear com Aplicações.* Campinas: Unicamp, 1999.

[8] BUSSAB, W.O. *Análise de variância e de regressão.* São Paulo: Atual, 1988.

[9] MYERS, R.H. *Classical and Modern Regression with Aplications*. Massachusetts: PWS Publishers, 1986.

[10] CASELLA,G. Leverage and regression through the origin. *American Statistician*, v.37, n.2, p.147, 1983.

[11] HAHN, G.J. Fitting regression models with no intercept term. *Journal of Quality Technology*, p.9-56, 1977.

ANEXO

EXEMPLOS REGRESSÃO LINEAR

Suely Ruiz Giolo

2007

EXEMPLO 1: Um experimento foi realizado para verificar o efeito da temperatura (X_1) e da concentração (X_2) na produção (Y) de um certo processo químico. Os dados obtidos foram:

Y	X_1	X_2
189	80	10
203	100	10
222	120	10
234	140	10
261	160	10
204	80	15
212	100	15
223	120	15
246	140	15
273	160	15
220	80	20
228	100	20
252	120	20
263	140	20
291	160	20
226	80	25
232	100	25
259	120	25
268	140	25
294	160	25

Fonte: Freire et al, 1999.

(a) Ajuste os modelos:

1) Y em X_1
2) Y em X_2
3) Y em X_1 e X_2
4) Y em X_2 e X_1
5) Y em X_1, X_2 e X_1X_2
6) Y em X_2, X_1 e X_1X_2

e, para cada um desses modelos anote: os coeficientes de regressão estimados, a SQres, o QMres e os coeficientes de determinação R^2 e R^2_a.

Quadro 1. Resultados dos modelos de regressão ajustados.

Modelos ajustados	**Estimativas dos parâmetros** β_o	β_1	β_2	β_3	**QMres**	**R^2**	**R_a^2**
1. Y em X_1	135,60	0,870	-	-	238,5	0,7383	0,7237
2. Y em X_2	197,58	-	2,424	-	707,3	0,2239	0,1808
3. Y em X_1 e X_2	93,18	0,870	2,424	-	36,5	0,9621	0,9577
4. Y em X_2 e X_1	93,18	0,870	2,424	-	36,5	0,9621	0,9577
5. Y em X_1, X_2 e X_1X_2	93,34	0,877	2,472	-0,0004	38,8	0,9621	0,9550
6. Y em X_2, X_1 e X_1X_2	93,34	0,877	2,472	-0,0004	38,8	0,9621	0,9550

```
================================================================================
ex1<-read.table("../~giolo/CE071/Exemplos/exemplo1.txt",h=T)
attach(ex1)

mod1<-lm(Y~X1)
mod1
summary(mod1)
anova(mod1)

mod2<-lm(Y~X2)
mod2
summary(mod2)
anova(mod2)
```

```
mod3<-lm(Y~X1+X2)
mod3
summary(mod3)
anova(mod3)

mod4<-lm(Y~X2+X1)
mod4
summary(mod4)
anova(mod4)

X3<-X1*X2
mod5<-lm(Y~X1+X2+X3)
mod5
summary(mod5)
anova(mod5)

mod6<-lm(Y~X2+X1+X3)
mod6
summary(mod6)
anova(mod6)
```

==========

(b) Apenas observando os resultados obtidos, qual modelo indicaria como sendo o *melhor* para explicar a produção (Y)?

Em uma análise preliminar, e dos resultados apresentados no Quadro I, é possível observar que os modelos 5 e 6 (com interação entre X_1 e X_2) apresentam QMres e R^2 muito similares aos modelos 3 e 4. Isso indica que a interação não está trazendo uma contribuição significativa e, portanto, não deve fazer sentido mantê-la no modelo. Os modelos 3 e 4 parecem ser as melhores opções dentre os modelos analisados, pois para ambos tem-se QMres pequeno e, ainda, $R^2 = 0,9621$, o que significa que 96,21% da variabilidade total de Y estaria sendo explicada pelas regressoras X_1 e X_2.. ***Modelo indicado****: Y em X_1 e X_2.*

Observe que as estimativas dos parâmetros associados às regressoras X_1 e X_2 não se diferenciam quando comparadas entre os modelo 3 e 4. Isso ocorreu porque a correlação entre X_1 e X_2 é nula e, portanto, os vetores X_1 e X_2 são ortogonais.

(c) Represente matricialmente o modelo indicado em (b)

$$\mathbf{Y = X\,\beta + \varepsilon} \rightarrow \begin{bmatrix} 189 \\ 203 \\ \vdots \\ 294 \end{bmatrix} = \begin{bmatrix} 1 & 80 & 10 \\ 1 & 100 & 10 \\ \vdots & \vdots & \vdots \\ 1 & 160 & 25 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{20} \end{bmatrix}.$$

(d) i) Fixe $x_2 = 10$ e obtenha $E(Y \mid x_1, x_2 = 10)$.

$E(Y \mid x_1, x_2 = 10) = 93,18 + 0,87x_1 + 2,424*10 = 117,42 + 0,87x_1$

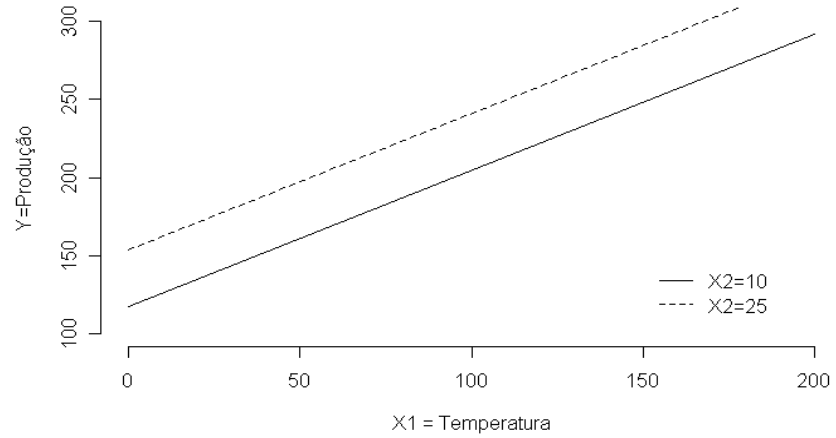
ii) Fixe $x_2 = 25$ e obtenha $E(Y \mid x_1, x_2 = 25)$.

$E(Y \mid x, x_2 = 25) = 93,18 + 0,87x_1 + 2,424*25 = 153,78 + 0,87x_1$

iii) Represente graficamente os resultados obtidos em i) e ii).

```
x1<-c(0,200)
y<-c(117.42, 291.42)
plot(x1,y,type="l",lty=1,ylim=range(c(100,300)),ylab="Y=Produção",
      xlab = "X1 = Temperatura", bty="n")
```

```
x1<-c(0,200)
y<-c(153.78, 327.78)
lines(x1,y, lty=2)
legend(150,150,lty=c(1,2),c("X2=10","X2=25"), lwd=1, bty="n")
```



(d) i) Fixe $x_1 = 100$ e obtenha $E(Y \mid x_1 = 100, x_2)$.

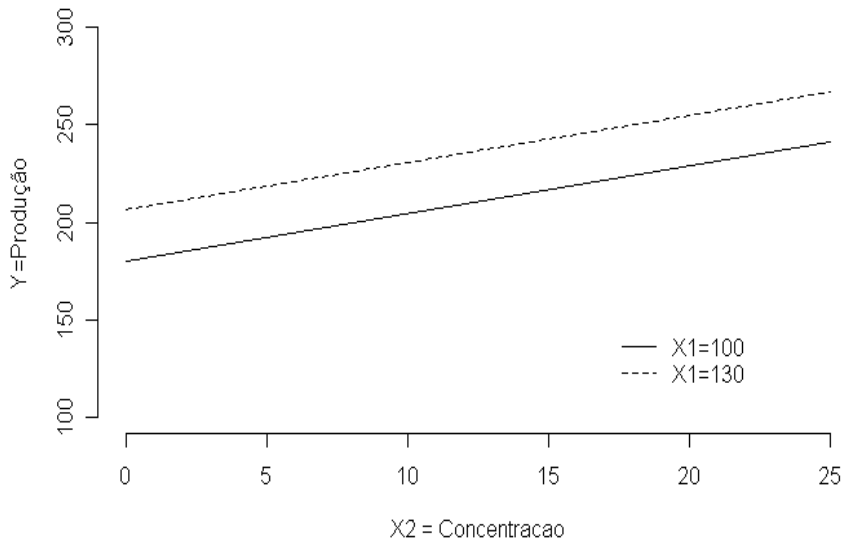
$$E(Y \mid x_1 = 100, x_2) = 93{,}18 + 0{,}87*100 + 2{,}424x_2 = 180{,}18 + 2{,}424x_2$$

ii) Fixe $x_1 = 130$ e obtenha $E(Y \mid x_1 = 130, x_2)$.

$$E(Y \mid x_1 = 100, x_2) = 93{,}18 + 0{,}87*130 + 2{,}424x_2 = 206{,}28 + 2{,}424x_2$$

iii) Represente graficamente os resultados obtidos em i) e ii).

```
x2<-c(0,25)
y<-c(180.18, 240.78)
plot(x2,y,type="l",lty=1,ylim=range(c(100,300)),ylab="Y=Produção",
     xlab = "X2 = Concentracao", bty="n")
x2<-c(0,25)
y<-c(206.28, 266.88)
lines(x2,y, lty=2)
legend(17,150,lty=c(1,2),c("X1=100","X1=130"), lwd=1, bty="n")
```



(e) Interprete os parâmetros no contexto dos dados.

i) β_o é o intercepto do plano com o eixo Y. Não apresenta interpretação prática nesse estudo.
ii) $\beta_1 = 0,87$ indica um aumento esperado na produção do processo químico de 0,87 unidades a cada acréscimo de uma unidade na temperatura, mantida a concentração fixa em 10, 25 ou outro valor possível.
iii) $\beta_2 = 2,424$ indica um aumento esperado na produção do processo químico de 2,424 unidades a cada acréscimo de uma unidade na concentração, mantida a temperatura fixa em 100, 130 ou outro valor possível.

(f) Qual a estimativa obtida para σ^2?

$\hat{\sigma}^2 = QMres = 36,5$.

(g) Obtenha, usando o modelo indicado, os valores preditos da produção Y.

```
obspred<-cbind(Y,mod3$fitted.values)
obspred
```

```
i  Obs Pred
1  189 187.02
2  203 204.42
3  222 221.82
4  234 239.22
5  261 256.62
6  204 199.14
7  212 216.54
8  223 233.94
9  246 251.34
10 273 268.74
11 220 211.26
12 228 228.66
13 252 246.06
14 263 263.46
15 291 280.86
16 226 223.38
17 232 240.78
18 259 258.18
19 268 275.58
20 294 292.98
```

(h) Teste as hipóteses: H_o: $\beta_1 = \beta_2 = 0$
H_a: pelo menos um β_j (j = 1, 2) difere de zero.

```
summary(mod3)
```

```
F-statistic: 215.9 on 2 and 17 DF, p-value: 8.234e-13
```

<u>Conclusão</u>*: Há evidências para a rejeição de H_o em favor de H_a*

(i) Teste as hipóteses: H_0: $\beta_j = 0$ *versus* H_a: $\beta_j \neq 0$ (j = 0, 1, 2).

```
summary(mod3)
             Estimate   Std.Error t value   Pr(>|t|)
(Intercept)  93.1800    7.25415   12.85     3.53e-10 ***
X1           0.87000    0.04779   18.20     1.38e-12 ***
X2           2.42400    0.24180   10.03     1.49e-08 ***
```

- fazendo uso dos intervalos de confiança (1- α = 0,95): $\hat{\beta}_j \pm t_{\alpha/2,\, n-p}\sqrt{\hat{\sigma}^2 C_{ii}}$

  $IC(\beta_0)_{95\%} = (93,18 - 2,11(7,25); 93,18 + 2,11(7,25)) = (77,88; 108,47)$
  $IC(\beta_1)_{95\%} = (0,87 - 2,11(0,047); 0,87 - 2,11(0,047)) = (0,77; 0,97)$
  $IC(\beta_2)_{95\%} = (2,424 - 2,11(0,2418); 2,424 + 2,11(0,2418)) = (1,91; 2,93)$

Conclusão: Como o valor "zero" não pertence a nenhum dos intervalos, há evidências para rejeição de H_0: $\beta_j = 0$ para j = 0, 1 e 2.

- fazendo uso do teste $t^* = \dfrac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{ii}}} = \dfrac{\hat{\beta}_j}{\text{d.p.}(\hat{\beta}_j)} \sim t_{n-p}$

  j = 0 → t^* = 93,18 / 7,25 = 12,85 p-valor = 3,53e-10 (bilateral)
  j = 1 → t^* = 0,87 / 0,04779 = 18,20 p-valor = 1,38e-12 (bilateral)
  j = 2 → t^* = 2,424 / 0,2418 = 10,03 p-valor = 1,49e-08 (bilateral)

Conclusão: os p-valores obtidos são pequenos o suficiente para que se conclua pela rejeição das hipóteses nulas H_0: $\beta_j = 0$ para j = 0, 1 e 2.

- fazendo uso do teste F* parcial

```
anova(mod3)
Analysis of Variance Table
Response: Y
             Df     Sum Sq       Mean Sq      F value       Pr(>F)
X1           1      12110.4      12110.4      331.40        1.384e-12 ***
X2           1       3672.4       3672.4      100.49        1.493e-08 ***
Residuals    17        621.2         36.5
```

$$\begin{cases} j = 1 \rightarrow F^* = \dfrac{SQ_E(X_1 \mid nenhuma)/1}{QMres(X_1, X_2)} \sim F_{1;\,(n-3)} = F_{1;\,17} \\ \qquad F^* = 12110{,}4 / 36{,}5 = 331{,}40 \quad (\text{p-valor} = 1{,}384\text{e-}12) \end{cases}$$

$$\begin{cases} j = 2 \rightarrow F^* = \dfrac{SQ_E(X_2 \mid X_1)/1}{QMres(X_1, X_2)} \sim F_{1;\,(n-3)} = F_{1;\,17} \\ \qquad F^* = 3672{,}4 / 36{,}5 = 100{,}49 \quad (\text{p-valor} = 1{,}49\text{e-}08) \end{cases}$$

Conclusão: os p-valores obtidos são pequenos o suficiente para que se conclua pela rejeição das hipóteses nulas H_0: $\beta_j = 0$ para j = 1, 2.

(j) Obtenha um IC de 95% para a esperança de Y em $x_1 = 80$ e $x_2 = 10$.

Para $x_1 = 80$ e $x_2 = 10$ tem-se: $\hat{Y} = 187{,}02$ e $V(\hat{Y}) = \hat{\sigma}^2 \mathbf{x_0}'(X'X)^{-1}\mathbf{x_0} = 8{,}76$. Assim, $I.C._{95\%} = (187{,}02 - 2{,}11(2{,}96);\ 187{,}02 + 2{,}11(2{,}96)) = (180{,}77;\ 193{,}26)$

```
new<-data.frame(cbind(80,10))
predict(lm(Y~X1+X2), new,interval="confidence")
predict(lm(Y~X1+X2), interval="confidence")
```

(k) Obtenha os coeficientes de determinação simples e parcial.

$r^2_{Y1} = 0{,}7382$
$r^2_{Y2} = 0{,}2238$
$r^2_{Y2 \cdot 1} = SQ_E(X_2 \mid X_1)/SQres(X_1) = 3672{,}4 / 4293{,}6 = 0{,}855$
$r^2_{Y1 \cdot 2} = SQ_E(X_1 \mid X_2)/SQres(X_2) = 12110{,}4 / 12731{,}6 = 0{,}951$

Assim, ao ser adicionada a regressora X_1 ao modelo que não contém nenhuma regressora, a SQres (que, nesse caso, é a SQtotal) é reduzida em 73,82% e, quando X_2 é adicionada ao modelo que contém X_1, a SQres(X_1) é reduzida em 85,5%. Analogamente, adicionar X_1 ao modelo que contém X_2, faz com que a SQres(X_2) seja reduzida em 95,1%.

(l) Obtenha os coeficientes de correlação simples e parcial.

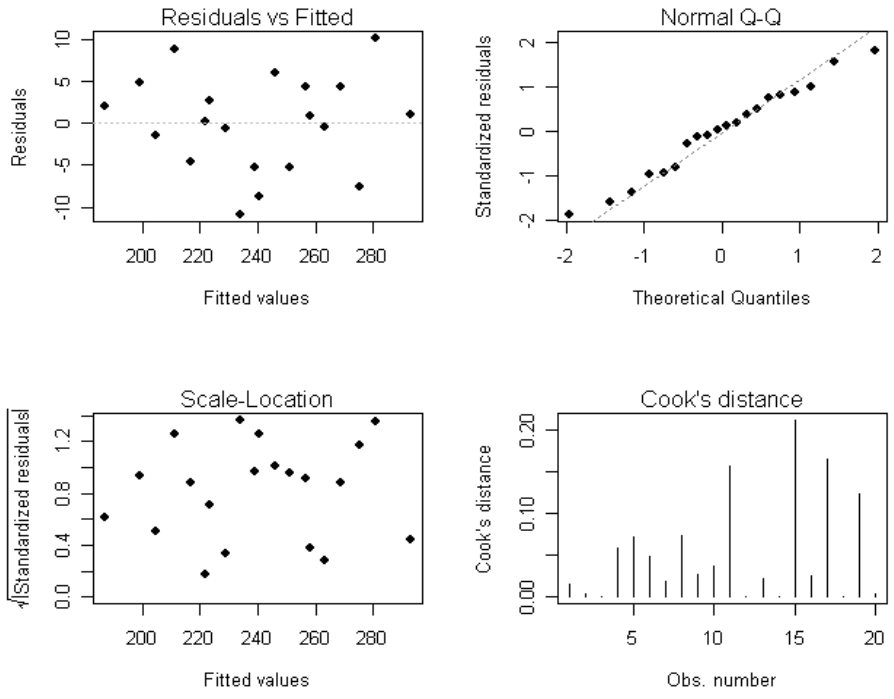
$r_{Y1} = 0{,}859$
$r_{Y2} = 0{,}473$
$r_{Y2 \cdot 1} = 0{,}9246$
$r_{Y1 \cdot 2} = 0{,}9752$

(m) Obtenha os resíduos e faça uma análise gráfica dos mesmos.

```
resid<- mod3$residuals
resid
```

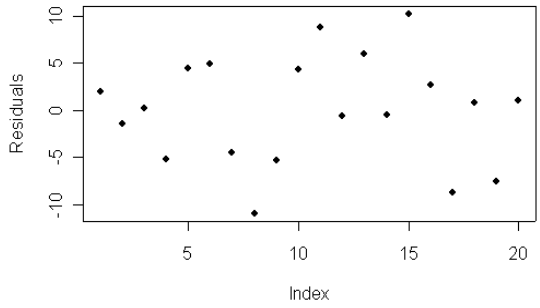
- **Resíduos *versus* preditos, sqrt{|resíduos|} *versus* preditos, Normal Q-Qplot e distância de Cook**

```
par(mfrow=c(2,2))
plot(mod3, which=c(1:4),pch=16, add.smooth=FALSE, id.n = 0)
# help(plot.lm)
```



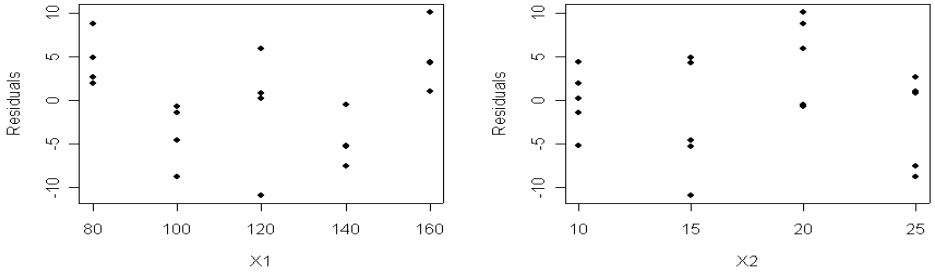
- **Resíduos *versus* ordem de coleta**

```
plot(mod3$residuals,pch=16,ylab="Residuals")
```



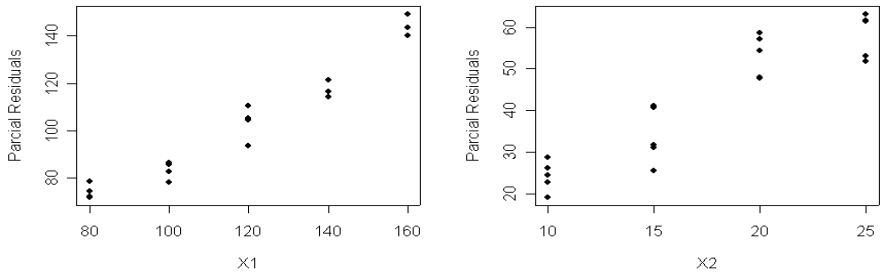
- **Resíduos *versus* cada regressora incluída no modelo**

```
par(mfrow=c(1,2))
plot(X1,mod3$residuals,pch=16,ylab="Residuals")
plot(X2,mod3$residuals,pch=16,ylab="Residuals")
```



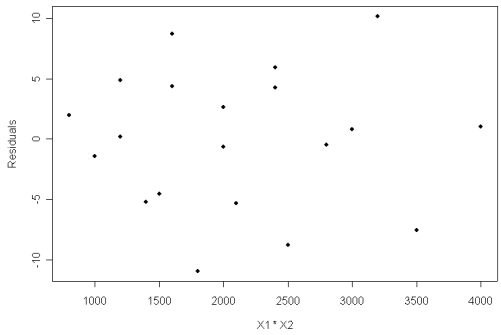
- **Resíduos parciais *versus* X_{ij} para cada X_j no modelo**

```
par(mfrow=c(1,2))
resparc1<-mod3$residuals + 0.87*X1
resparc2<-mod3$residuals + 2.424*X2
plot(X1,resparc1,pch=16,ylab="Parcial Residuals")
plot(X2,resparc2,pch=16,ylab="Parcial Residuals")
```



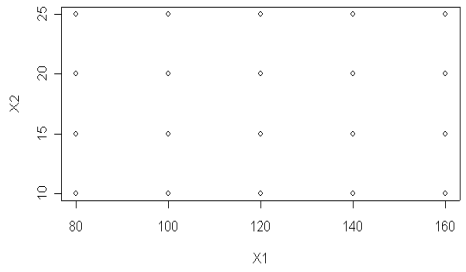
- **Resíduos *versus* interações não incluídas no modelo**

```
plot(X1*X2,mod3$residuals,pch=16,ylab="Residuals")
```



- **Gráfico de X_i versus X_j**

```
plot(X1,X2)
cor(X1,X2)
```



(n) Medidas para detecção de multicolinearidade

```
X<-as.matrix(cbind(X1,X2))
rxx<- cor(X)
det(rxx)
eigen(rxx)[1]
```

$VIF_1 = 1/(1-0) = 1$ e $VIF_2 = 1/(1-0) = 1$

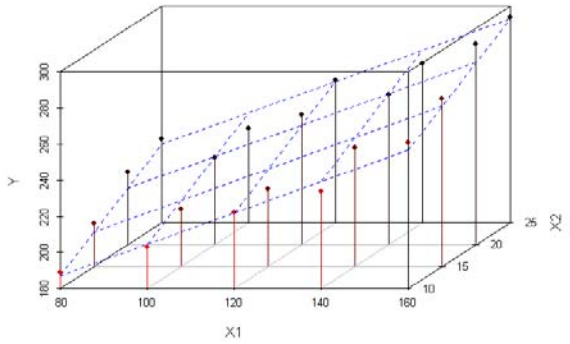
$\det(\mathbf{r}_{xx}) = 1$ e autovalores: $\lambda_1 = \lambda_2 = 1$.

Não há, portanto, problemas de multicolinearidade nesse estudo.

(o) Obtenha o gráfico dos dados observados e do plano ajustado.

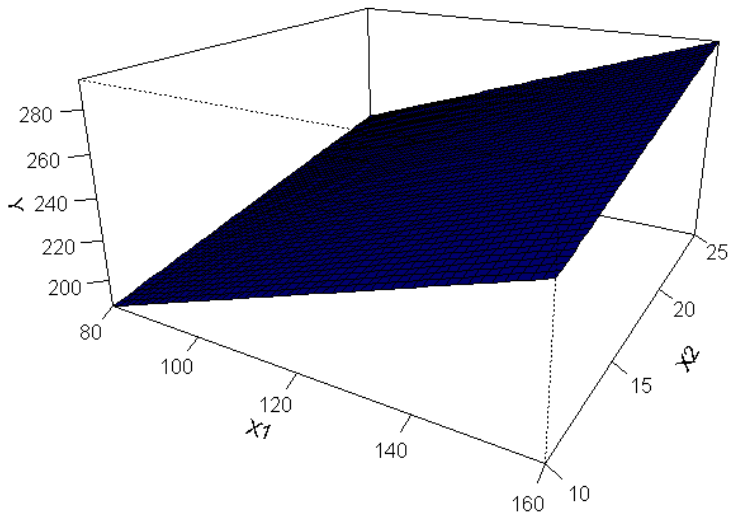
***Obs**: necessário baixar e instalar o pacote scatterplot3d*

```
require(scatterplot3d)
s3d <- scatterplot3d(X1,X2,Y,type="h",highlight.3d=TRUE,
       angle=55, scale.y =0.7, pch=16)
my.lm <- lm(Y~X1+X2)
s3d$plane3d(my.lm, col=4)
```



(j) Obtenha o gráfico do plano ajustado.

```
x1 <- seq(80, 160, length=50)
x2 <- seq(10,25,length=50)
 f <- function(x1,x2){r <- 93.18+0.87*x1+2.424*x2}
 y <- outer(x1,x2,f)
y[is.na(y)] <- 1
par(bg = "white")
persp(x1,x2,y,theta = 30, phi = 20,expand = 0.5,col = "blue",
                ltheta = 120, shade = 0.75, ticktype = "detailed",
                xlab = "X1", ylab = "X2", zlab = "Y")
```



EXEMPLO 2: Um engarrafador de bebidas está analisando os serviços de rotina realizados no sistema de distribuição de máquinas acionadas por moedas. Ele está interessado em predizer o tempo requerido para esses serviços de rotina que incluem: estocagem da máquina com bebidas e manutenções pequenas. O engenheiro industrial responsável sugeriu duas variáveis como as que mais afetam o tempo requerido por estes serviços:quantidade de bebida estocada e a distância percorrida pelo profissional responsável pelos serviços.

Tabela 1: Dados observados

Y = Tempo requerido (em minutos)	X_1 = Quantidade estocada (em unidades)	X_2 = Distância percorrida (em pés)
16.68	7	560
11.50	3	220
12.03	3	340
14.88	4	80
13.75	6	150
18.11	7	330
8.00	2	110
17.83	7	210
79.24	30	1460
21.50	5	605
40.33	16	688
21.00	10	215
13.50	4	255
19.75	6	462
24.00	9	448
29.00	10	776
15.35	6	200
19.00	7	132
9.50	3	36
35.10	17	770
17.90	10	140
52.32	26	810
18.75	9	450
19.83	8	635
10.75	4	150

Fonte: Montgomery and Peck (1992)

Comandos em R para obtenção dos resultados apresentados no texto [p. 23-27]

==========

a) Leitura dos dados

```
ex2<- read.table("../~giolo/CE071/Exemplos/exemplo2.txt",h=T)
attach(ex2)
```

b) Obtenção da matriz r_{xx}

```
X<-as.matrix(cbind(X1,X2))
rxx<-cor(X)
rxx
```

c) Obtenção dos VIF's, det(r_{XX}) e autovalores de r_{XX}

```
vif1<- 1/(1-(rxx[1,2]^2))
vif1
vif2<- 1/(1-(rxx[2,1]^2))
vif2
det(rxx)
eigen(rxx)[11
```

d) Ajuste dos modelos de regressão: Y em X_1 e X_2; Y em X_1 e Y em X_2

```
mod1 <- lm(Y~X1+X2)
anova(mod1)
summary(mod1)

mod2 <- lm(Y~X1)
anova(mod2)
summary(mod2)

mod3 <- lm(Y~X2)
anova(mod3)
summary(mod3)
```

e) Gráficos dos resíduos

```
par(mfrow=c(2,2))
plot(mod1,which=c(1:4),pch=16,add.smooth=FALSE)

par(mfrow=c(1,1))
plot(mod1$residuals,pch=16,ylab="Resíduos")

par(mfrow=c(1,2))
plot(X1,mod1$residuals,pch=16,ylab="Resíduos")
plot(X2,mod1$residuals,pch=16,ylab="Resíduos")

resparc1<- mod1$residuals + 1.615*X1
resparc2<-mod1$residuals + 0.014*X2
plot(X1,resparc1,pch=16,ylab="Resíduos Parciais")
plot(X2,resparc2,pch=16,ylab="Resíduos Parciais")
```

f) Diagnóstico de influência

```
influ<-influence.measures(mod1)
influ
summary(influ)
```

g) Gráficos do diagnóstico de influência

```
par(mfrow=c(1,3))
plot(influ$infmat[,1], ylab="DFBeta(0)", pch=16)
plot(influ$infmat[,2], ylab="DFBeta(1)", pch=16)
plot(influ$infmat[,3], ylab="DFBeta(2)", pch=16)

par(mfrow=c(2,2))
plot(influ$infmat[,4], ylab="DFFits", pch=16)
plot(influ$infmat[,5], ylab="CovRatio", pch=16)
plot(influ$infmat[,6], ylab="D de Cook", pch=16)
plot(influ$infmat[,7], ylab="hii", pch=16)
```

h) Reajuste do modelo Y em X_1 e X_2

h_1) sem a observação 9

```
ex2s9<-ex2[-9,]
attach(ex2s9)
mod1s9<-lm(Y~X1+X2)
summary(mod1s9)
anova(mod1s9)
```

h_2) sem a observação 22

```
ex2s22<-ex2[-22,]
attach(ex2s22)
mod1s22<-lm(Y~X1+X2)
summary(mod1s22)
anova(mod1s22)
```

h_3) sem as observações 9 e 22

```
ex2s9s22<-ex2s9[-21,]
attach(ex2s9s22)
mod1s9s22<-lm(Y~X1+X2)
summary(mod1s9s22)
anova(mod1s9s22)
```

i) Gráficos dos resíduos do modelo Y em X_1 e X_2 sem a observação 9

```
par(mfrow=c(2,2))
plot(mod1s9,which=c(1:4),pch=16,add.smooth=FALSE)
plot(mod1s9,which=c(1:4),pch=16,add.smooth=FALSE,id.n=0)
```

j) Gráficos dos resíduos do modelo Y em X_1 sem a observação 9

```
ex2s9<-ex2[-9,]
attach(ex2s9)
mod2s9<-lm(Y~X1)
plot(mod2s9,which=c(1:4),pch=16,add.smooth=FALSE)
```

k) Gráficos: valores observados e modelo ajustado (Y em X_1 e X_2 sem obs 9)

obs: necessário instalar no diretório *library* do R a função scatterplot3d

```
attach(exe2)
require(scatterplot3d)
s3d <- scatterplot3d(X1,X2,Y, type="h", highlight.3d=TRUE,
                     angle=55, scale.y=0.7, pch=16)
my.lm <- lm(Y~X1+X2)
s3d$plane3d(my.lm,col=4)
```

k_1) gráfico somente do modelo ajustado

```
x1 <- seq(2,30,length=50)
x2 <- seq(30,1500,length=50)
      f <- function(x1,x2){
         r <- 4.447+1.498*x1+0.0103*x2
         }
y <- outer(x1,x2,f)
y[is.na(y)] <- 1
par(bg = "white")
persp(x1,x2,y, theta=30, phi=20, expand=0.5, col="blue",
      ltheta=120, shade=0.75, ticktype="detailed",
      xlab="X1", ylab="X2", zlab="Y")
```

l) valor predito em x_1=8 e x_2=275

```
new<-data.frame(cbind(8,275))
predict(mod1s9,new,interval="confidence")
predict(mod1s9,interval="confidence")
```