Índice

2. REGRESSÃO LINEAR MÚLTIPLA

2.1 Introdução

Na regressão linear simples exploram-se os conceitos e técnicas para se analisar e utilizar a relação linear entre duas variáveis. Esta análise conduz a uma equação que pode ser utilizada para se "predizerem" valores de uma variável dependente (a variável resposta) dados valores de uma variável independente associada (o regressor).

A intuição deixa adivinhar que, geralmente, se pode melhorar esta "predição" se incluirmos novas variáveis independentes ao modelo (à equação de regressão). Deve, contudo, ter-se em conta o principio da parcimónia, ou seja, deve haver "equilíbrio" entre o número de parâmetros do modelo. Num modelo de regressão múltiplo, enquanto um número excessivo de parâmetros pode levar a um sobreajustamento dos dados, um número reduzido de parâmetros pode levar a um sobajustamento.

Os conceitos e técnicas para se analisarem as relações lineares entre uma variável dependente e várias variáveis independentes são uma extensão natural do que foi apresentado no capítulo da regressão linear simples. Contudo, como é de esperar, os cálculos tornam-se mais complexos. É vulgar encontrar investigadores que trabalham com inúmeras variáveis, o que hoje é bastante facilitado com a evolução dos meios informáticos.

2.2 O modelo de regressão linear múltipla

Na regressão linear múltipla assume-se que existe uma relação linear entre uma variável Y (a variável dependente) e k variáveis independentes, x_j ($j = 1,...,k$). As variáveis independentes são também chamadas variáveis explicatórias ou regressores, uma vez que são utilizadas para explicarem a variação de Y. Muitas vezes são também chamadas variáveis de predição, devido à sua utilização para se predizer Y .

As condições subjacentes à regressão linear múltipla são análogas à da regressão linear simples, resumidamente:

1. As variáveis independentes x_j são não aleatórias (fixas);
2. Para cada conjunto de valores de x_j há uma subpopulação de valores de Y. Para a construção dos intervalos de confiança e dos testes de hipóteses deve poder-se assumir que estas subpopulações seguem a distribuição normal;
3. As variâncias das subpopulações de Y são iguais;

4. Os valores de *Y* são estatisticamente independentes. Por outras palavras, quando se extrai a amostra, assume-se que os valores de *Y* obtidos para um determinado conjunto de valores de x_j são independentes dos valores de *Y* obtidos para outro qualquer conjunto de valores de x_j.

Os dados podem ser organizados numa tabela do tipo da que se segue:

Y	x_1	x_2	$\cdots$	x_k
y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

Tabela2.1 – Dados utilizados para a regressão linear múltipla

Na tabela2.1 estão representados os valores de *k* variáveis independentes (não aleatórias) e os valores da variável resposta (aleatória) depois de efectuada uma determinada experiência para uma amostra de tamanho *n*.

Um modelo de regressão linear múltiplo descreve uma relação entre as *k* variáveis independentes, x_j, e a variável dependente, *Y*, da seguinte maneira

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon, \qquad \textbf{(2.1)}$$

designado por modelo de regressão múltipla (convencional) com *k* regressores. Os, $p = k+1$, parâmetros β_j, $j = 0,1,\ldots,k$, são os coeficientes de regressão (parciais) e ε é o erro aleatório. Este modelo descreve um hiperplano no espaço *k*-dimensional dos regressores $\{x_j\}$. Em tudo o que se segue iremos supor a presença de β_0 no modelo. Os parâmetros β_j, $j = 1,\ldots,k$ representam a variação esperada na resposta *Y* para cada unidade de variação em x_j quando todos os restantes regressores x_i $(i \neq j)$ são considerados constantes em termos experimentais. Assume-se, assim, que o modelo que nos permite descrever a i-ésima resposta y_i é

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i,\ i = 1,\ldots,n \qquad \textbf{(2.2)}$$

isto é, cada observação $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$, $i = 1,\ldots,n$ $(n > k)$ satisfaz esta equação. Supõe-se, portanto, que a variável *Y* é função linear de *k* regressores, do termo independente ($p = k+1$ parâmetros) e do erro aleatório.

Para se proceder ao ajustamento deste modelo de regressão, devido às dificuldades de cálculo no manuseamento do elevado números de parâmetros, é conveniente expressar as operações matemáticas utilizando notação matricial.

O modelo apresentado na equação (2.2) é um sistema de n equações que pode ser representado matricialmente por

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad \textbf{(2.3)}$$

onde

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},\ \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix},\ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \text{ e } \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Sobre a constituição das diferentes matrizes pode referir-se o seguinte:

1) Matriz $\boldsymbol{Y}$, é o vector coluna ($n\times 1$) constituído pelas observações da variável resposta.
2) Matriz $\boldsymbol{X}$ ($n\times p$), as linhas são constituídas pelos valores das variáveis independentes, isto é, na linha i ($i=1,\ldots,n$) aparecem sucessivamente os coeficientes de β_j ($j=0,1,\ldots,k$) da j-ésima equação do sistema. Alternativamente, pensando em termos de colunas ter-se-á:
   1ª coluna – todos os valores iguais a 1, os coeficientes de β_0 em cada equação $i=1,\ldots,n$.
   2ª coluna – surgem directamente as observações da variável x_1 $(x_{11},x_{12},\ldots,x_{1n})$, são os coeficientes de β_1 em cada equação $i=1,\ldots,n$.
   Colunas seguintes – Aparecem as observações das variáveis $x_2,x_3,\ldots,x_k$ pelas mesmas razões.
3) Matriz $\boldsymbol{\beta}$, é o vector coluna ($p\times 1$) dos coeficientes de regressão.
4) Matriz $\boldsymbol{\varepsilon}$, é o vector coluna ($n\times 1$) dos erros aleatórios.

Pretende-se, agora, encontrar o vector de estimadores dos mínimos quadrados $\hat{\boldsymbol{\beta}} = \boldsymbol{B}$ que minimize a soma de quadrados do erro. Da equação(2.3) tem-se $\boldsymbol{\varepsilon} = \boldsymbol{Y} - \boldsymbol{X\beta}$ e, consequentemente

$$SQ_E = L = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = \left(\boldsymbol{Y} - \boldsymbol{X\beta}\right)^T\left(\boldsymbol{Y} - \boldsymbol{X\beta}\right) = \boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X\beta}$$

pois sendo $\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y}$ do tipo (1×1) ou escalar, a sua transposta, $\boldsymbol{Y}^T\boldsymbol{X\beta}$ tem o mesmo valor. O estimador dos mínimos quadrados $\hat{\boldsymbol{\beta}}$ será a solução (em ordem a $\hat{\boldsymbol{\beta}}$) das seguintes equações

$$\frac{\partial L}{\partial \hat{\boldsymbol{\beta}}} = 0 \Leftrightarrow -2\boldsymbol{X}^T\boldsymbol{Y} + 2\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = 0 \Leftrightarrow \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y},$$

as $p = k+1$ equações normais na forma matricial. Para resolver estas equações (em ordem a $\hat{\boldsymbol{\beta}}$) multiplicam-se ambos os membros, à esquerda, por $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$ (supondo que esta matriz é regular) obtendo-se o estimador

$$\hat{\boldsymbol{\beta}} = \boldsymbol{B} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}\,. \qquad \textbf{(2.4)}$$

As matriz $\boldsymbol{X}^T\boldsymbol{X}$ e $\boldsymbol{X}^T\boldsymbol{Y}$ são

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i2}x_{i1} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{ik}x_{i1} & \sum_{i=1}^n x_{ik}x_{i2} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{bmatrix}$$

e

$$\boldsymbol{X}^T\boldsymbol{Y} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & \cdots & x_{n1} \\ x_{12} & x_{22} & x_{32} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{1k} & x_{2k} & x_{3k} & \cdots & x_{nk} \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \sum_{i=1}^n x_{i2}y_i \\ \vdots \\ \sum_{i=1}^n x_{ik}y_i \end{bmatrix}$$

Então

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i2}x_{i1} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{ik}x_{i1} & \sum_{i=1}^n x_{ik}x_{i2} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{bmatrix}^{-1}\begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \sum_{i=1}^n x_{i2}y_i \\ \vdots \\ \sum_{i=1}^n x_{ik}y_i \end{bmatrix}.$$

A matriz $\boldsymbol{X}^T\boldsymbol{X}$ é uma matriz simétrica ($p \times p$) e $\boldsymbol{X}^T\boldsymbol{Y}$ é um vector coluna ($p \times 1$), ou seja, como seria de esperar, a matriz $\hat{\boldsymbol{\beta}}$ é um vector coluna ($p \times 1$).

O modelo de regressão ajustado, correspondente a (2.2) é

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \ , \ i = 1, \ldots, n \qquad \textbf{(2.5)}$$

ou, em notação matricial,

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \, . \qquad \textbf{(2.6)}$$

A diferença entre a observação y_i e o correspondente valor ajustado (a estimativa) $\hat{y}_i$ é o resíduo (erro), $e_i = y_i - \hat{y}_i$. O vector $(n \times 1)$ dos resíduos é $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$.

Deve ter-se em atenção que, as unidades das variáveis independentes (regressores) são, em regra, diferentes, portanto, não se pode interpretar os valores dos seus parâmetros associados como uma medida de contribuição de cada regressor para a explicação da variação da variável resposta. Pode, contudo, estandardizar a equação de regressão convencional fazendo a seguinte transformação

$$y_i' = \frac{y_i - \overline{y}}{s_y} \qquad \textbf{(2.7)}$$

e

$$x_{ij}' = \frac{x_{ij} - \overline{x}_j}{s_{x_j}} \, , \qquad \textbf{(2.8)}$$

isto é, subtraindo a cada valor observado a média das observações e dividindo esta quantidade pelo desvio padrão respectivo. Obtém-se, então, a equação da regressão linear múltipla estandardizada

$$y_i' = \beta_1' x_{i1}' + \beta_2' x_{i2}' + \ldots + \beta_k' x_{ik}' + \varepsilon_i \ , \ i = 1, \ldots, n \, , \qquad \textbf{(2.9)}$$

repare-se que $\beta_0' = 0$. Os coeficientes β_j', $(j = 1, \ldots, k)$, são os coeficientes de regressão parciais estandardizados e estão relacionados com os coeficientes de regressão convencionais, os β_j's, da seguinte maneira

$$\beta_j' = \beta_j \frac{s_{x_j}}{s_y} \, . \qquad \textbf{(2.10)}$$

As quantidades $\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $s_{x_j} = \frac{1}{n-1}\sum_{i=1}^{n}\sqrt{(x_{ij} - \overline{x}_j)^2}$ e $s_y = \frac{1}{n-1}\sum_{i=1}^{n}\sqrt{(y_i - \overline{y})^2}$ são, respectivamente, as médias amostrais e os desvios padrão amostrais das variáveis x_j e y.

Estes coeficientes, em vez de expressarem a taxa de variação nas medidas originais, padronizam as unidades das diferentes variáveis independentes para unidades de desvio padrão, ou seja, o coeficiente de regressão estandardizado β_j' dá a taxa de variação em unidades de desvio padrão

para y por cada variação de uma unidade de desvio padrão para x_j (mantendo constantes todos as outras variáveis). Uma vantagem destes coeficientes estandardizados é o facto dos seus valores poderem ser comparados directamente (uma vez que as variáveis independentes passam a ter a mesma unidade de medida), dando, assim, uma antevisão das variáveis independentes que mais contribuem para a explicação da variação da variável dependente.

Exemplo2.1: Pretende-se investigar a utilização de um modelo de regressão linear múltiplo para se tentar explicar a variação da viscosidade de um polímero (*Y*) em função da temperatura de reacção, x_1, e da taxa de alimentação do catalisador, x_2. Realizando-se uma experiência, para os diferentes valores de x_1 e x_2, obtiveram-se os valores de *Y*, os y_i´s, que se apresentam na tabela2.2.

N.º da observação	Viscosidade (y)	Temperatura (x_1, ºC)	Catalisador (x_2, lb/h)
1	2256	80	8
2	2340	93	9
3	2426	100	10
4	2293	82	12
5	2330	90	11
6	2368	99	8
7	2250	81	8
8	2409	96	10
9	2364	94	12
10	2379	93	11
11	2440	97	13
12	2364	95	11
13	2404	100	8
14	2317	85	12
15	2309	86	9
16	2328	87	12

Tabela2.2 – Dados referentes à experiência com a viscosidade de um polímero

O modelo a ser ajustado é do tipo $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, onde se deve estimar os coeficientes de regressão. Em notação matricial, $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{Y}$, considerando a amostra obtém-se

$$\boldsymbol{X}^T \boldsymbol{X} = \begin{bmatrix} 16 & 1458 & 164 \\ 1458 & 133560 & 14946 \\ 164 & 14946 & 1726 \end{bmatrix} \text{ (matriz é simétrica),}$$

$$\left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} = \begin{bmatrix} 14{,}176004 & -0{,}129746 & -0{,}223453 \\ -0{,}129746 & 1{,}429184\times10^{-3} & -4{,}763947\times10^{-5} \\ -0{,}223453 & -4{,}763947\times10^{-5} & 2{,}222381\times10^{-2} \end{bmatrix} \text{ e } \boldsymbol{X}^T \boldsymbol{Y} = \begin{bmatrix} 37577 \\ 3429550 \\ 385562 \end{bmatrix}, \text{ donde}$$

$$\hat{\boldsymbol{\beta}}_0 = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 1566{,}07777 \\ 7{,}62129 \\ 8{,}58485 \end{bmatrix}.$$

Assim, o modelo de regressão ajustado aos dados é, com quatro casas decimais,

$$y = 1566,0777 + 7,6213x_1 + 8,5848x_2 .$$

A partir desta equação é possível obter os valores estimados (esperados através do modelo) de *Y* e predizer observações futuras para a mesma variável. Por exemplo, para a primeira observação $x_{11} = 80$ e $x_{12} = 8$, o valor ajustado será $\hat{y}_1 = 1566,00777 + 7,6213x_{11} + 8,5848x_{12} = 2244,46$, o valor observado correspondente é $y_1 = 2256$, o resíduo para esta observação é $e_1 = y_1 - \hat{y}_1 = 11,54$.

Apresentam-se na tabela seguinte os valores ajustados (estimativas) da variável resposta a partir deste modelo de regressão e os respectivos erros de ajustamento para cada observação.

N.º da observação	y_i	$\hat{y}_i$	e_i
1	2256	2244,46	11,54
2	2340	2352,12	-12,12
3	2426	2414,06	11,94
4	2293	2294,04	-1,04
5	2330	2346,43	-16,43
6	2368	2389,26	-21,26
7	2250	2252,08	-2,08
8	2409	2383,57	25,43
9	2364	2385,50	-21,50
10	2379	2369,29	9,71
11	2440	2416,95	23,05
12	2364	2384,53	-20,53
13	2404	2396,89	7,11
14	2317	2316,91	0,09
15	2309	2298,77	10,23
16	2328	2332,15	-4,15

Tabela2.3 – Observações e estimativas da variável resposta e respectivos resíduos

Obs.2.1: A título de exemplo, obtivemos as estimativas dos valores esperados sem ter em conta se o modelo é adequado.

Para se ver qual o regressor que mais contribui para a explicação da variação da variável resposta utiliza-se a equação de regressão estandardizada $y' = \beta_1' x_1' + \beta_2' x_2'$, onde os coeficientes β_j' se obtêm a partir da igualdade $\beta_j' = \beta_j \dfrac{s_{x_j}}{s_y}$, ou estandardizar os valores das variáveis através de (2.7) e (2.8).

Da tabela2.2, vem $s_y = 56,3536$, $s_{x_1} = 6,8301$ e $s_{x_2} = 1,7321$, donde, $\beta_1' = 0,9237$ (quando a variação de x_1 for de um desvio padrão, a variação de *Y* será de 0,9237 unidades de desvio padrão) e $\beta_2' = 0,2639$ (quando a variação de x_2 for de um desvio padrão, a variação de *Y* será de 0,2639 unidades de desvio padrão), como $|\hat{\beta}_1'| > |\hat{\beta}_2'|$ portanto, a variável x_1 contribui mais na explicação da variação de *Y* do que x_2. A equação padrão será $y' = 0,9237x_1' + 0,26392x_2'$.

2.3 Validação do modelo de regressão múltipla

Antes se utilizar um modelo de regressão múltipla para a predição e estimação, é aconselhável, saber se vale a pena aplicar tal modelo (se o modelo é adequado), ou seja, se através do modelo os regressores (ou pelo menos algum) contribuem para explicar (linearmente) a variação da variável resposta. Para isso, vamos utilizar testes de hipóteses e o coeficiente de determinação, o raciocínio em tudo o que se segue é análogo ao utilizado para o modelo de regressão simples.

2.3.1 Estimadores de σ^2 e dos erros padrão para a regressão linear múltipla

Antes de se passar à validação, propriamente dita, do modelo e à construção dos intervalos de confiança e de predição, faz-se uma breve referencia aos estimadores de σ^2 e dos erros padrão para a regressão linear múltipla.

Na regressão linear múltipla as condições impostas aos erros ε_i, $i = 1,\ldots,n$ são:

i) $E[\varepsilon_i] = 0$ donde $E[\boldsymbol{\varepsilon}] = 0$;

ii) $V[\varepsilon_i] = \sigma^2, i = 1,\ldots,n$, ε_i e ε_j não correlacionados (independentes), $\text{cov}[\varepsilon_i, \varepsilon_j] = 0$, $(i \neq j)$, donde a matriz de variâncias-covariâncias do erro é $\Sigma[\boldsymbol{\varepsilon}] = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \sigma^2 I$, I é a matriz identidade;

iii) Como, $\varepsilon_i \sim N(0,\sigma^2)$, $i = 1,\ldots,n$, então, $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0},\sigma^2 I)$, ou seja, o vector dos erros tem distribuição normal multivariada com vector médio $\boldsymbol{0}$ (a matriz nula) e matriz de variâncias-covariâncias $\sigma^2 I$.

Nestas condições e atendendo a que $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, tem-se para $\boldsymbol{Y}$,

$$E[\boldsymbol{Y}] = E[\boldsymbol{X\beta} + \boldsymbol{\varepsilon}] = E[\boldsymbol{X\beta}] + E[\boldsymbol{\varepsilon}] = \boldsymbol{X\beta} \qquad \textbf{(2.11)}$$

e

$$\Sigma[\boldsymbol{Y}] = \Sigma[\boldsymbol{X\beta} + \boldsymbol{\varepsilon}] = \Sigma[\boldsymbol{X\beta}] + \Sigma[\boldsymbol{\varepsilon}] = \sigma^2 \boldsymbol{I}, \qquad \textbf{(2.12)}$$

uma vez que $\Sigma[\boldsymbol{X\beta}] = \boldsymbol{0}$. Simbolicamente, $\boldsymbol{Y} \sim N\left(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}\right)$.

Passemos às propriedades dos estimadores $\hat{\boldsymbol{\beta}}$, como o estimador dos mínimos quadrados $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ é uma combinação linear das observações (variáveis normais independentes), tem distribuição normal multivariada. Atendendo a que $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$ é uma matriz constante e que $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$, vem para vector média

$$E\left[\hat{\boldsymbol{\beta}}\right]=E\left[\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}\right]=\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T E[\boldsymbol{Y}]=\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}=\boldsymbol{\beta}$$

donde se conclui que $\hat{\boldsymbol{\beta}}$ é um estimador não enviesado ou centrado de $\boldsymbol{\beta}$. A matriz simétrica ($p\times p$) de variâncias-covariâncias de $\hat{\boldsymbol{\beta}}$ é

$$\Sigma\left[\hat{\boldsymbol{\beta}}\right]=\Sigma\left[\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}\right]=\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{I}\sigma^2=\sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}=\sigma^2 C$$

onde os elementos da diagonal principal são as variâncias dos estimadores e os restantes elementos as covariâncias entre estimadores. Simbolicamente, $\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta},\sigma^2\boldsymbol{C}\right)$, com $\boldsymbol{C}=\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$.

Por exemplo, considerando $k=2$ (2 regressores e portanto 3 parâmetros),

$$\Sigma\left[\hat{\boldsymbol{\beta}}\right]=\sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}=\sigma^2\boldsymbol{C} \qquad \textbf{(2.13)}$$

com

$$\boldsymbol{C}=\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}=\begin{bmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{bmatrix} \qquad \textbf{(2.14)}$$

que é uma matriz simétrica, então

$$V\left[\hat{\beta}_j\right]=\sigma^2 C_{jj},\ j=0,1,2 \text{ e } \operatorname{cov}\left[\hat{\beta}_i,\hat{\beta}_j\right]=\sigma^2 C_{ij},\ i\neq j$$

donde $\hat{\beta}_j \sim N\left(\beta_j,\sigma^2 C_{jj}\right)$.

As estimativas das variâncias destes coeficientes de regressão são obtidas substituindo σ^2 por um seu estimador apropriado. Quando σ^2 é substituído por $\hat{\sigma}^2=S^2$, a raiz quadrada da variância estimada para o *j*-ésimo coeficiente de regressão é chamado o erro padrão estimado de $\hat{\beta}_j$, ou seja

$$se\left(\hat{\beta}_j\right)=\sqrt{\hat{\sigma}^2 C_{jj}} \qquad \textbf{(2.15)}$$

sendo C_{jj} o *j*-ésimo elemento da diagonal principal da matriz $\boldsymbol{C}$ (2.14), correspondente a $\hat{\beta}_j$.

Uma estimativa de σ^2 é obtida a partir dos resíduos. Como $SQ_E=\boldsymbol{Y}^T\boldsymbol{Y}-\hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y}$ prova-se que $E\left[SQ_E\right]=\sigma^2(n-p)$ e assim um estimador não enviesado de σ^2 é

$$\hat{\sigma}^2=S^2=\frac{SQ_E}{n-p}=MQ_E. \qquad \textbf{(2.16)}$$

2.3.2 Significância do modelo de regressão múltipla

Até agora assumiu-se um modelo linear da forma $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$ e obtivemos estimadores para os coeficientes de regressão. Queremos verificar se as variáveis independentes, $x_1, x_2, \ldots, x_k$, contribuem significativamente com informação para explicar linearmente a variação da variável resposta (variável dependente) *Y*. Pois quanto maior for esta contribuição melhores serão os resultados da estimação e da predição.

O teste de significância para a regressão é um teste para se determinar se há uma relação linear entre a função resposta *y* e os regressores *x*, para este efeito pode-se utilizar um teste de hipótese. Estes testes hipóteses acerca dos parâmetros do modelo de regressão, requerem que os termos do erro ε_i no modelo de regressão sejam normais e independentemente distribuídos com média zero e variância σ^2.

As hipóteses a testar são:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \text{ (hipótese nula)} \quad \textbf{(2.17)}$$

$$H_1 : \beta_j \neq 0 \text{ para algum } j,\ j = 1, \ldots, k \text{ ,(hipótese alternativa).} \quad \textbf{(2.18)}$$

Temos portanto, um teste bilateral face a *n* observações (a amostra). A hipótese nula que se pretende testar é a não existência de regressão. A rejeição de H_0 significa que pelo menos umas das variáveis independentes $x_1, x_2, \ldots, x_k$ (regressores) contribui significativamente para explicar a variação da variável dependente *Y*, e esta explicação pode ser representada por um modelo de regressão linear (o modelo diz-se significativo).

Caso não se rejeite H_0, ter-se-á o modelo $Y = \beta_0 + \varepsilon$, ou seja, $E[Y] = E[\beta_0 + \varepsilon] = \beta_0$ (constante), concluindo-se que os $x_1, x_2, \ldots, x_k$ não contribuem para explicar a variação de *Y*. O que leva à conclusão de que não há relação linear entre as variáveis (mau ajustamento do modelo linear em relação aos dados). O modelo diz-se não significativo e não deve ser utilizado.

Quando se pretende realizar um teste bilateral a análise de variância (Anova) pode ser utilizada para se analisar a significância do modelo de regressão. Para isso, utiliza-se a partição da soma de quadrados, a identidade da análise de variância, $SQ_T = SQ_R + SQ_E$. Em notação matricial,

- $SQ_T = \boldsymbol{Y}^T\boldsymbol{Y} - \dfrac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \boldsymbol{Y}^T\boldsymbol{Y} - n\bar{y}^2$, a soma de quadrados total, mede a variação total das observações em torno da sua média ;

- $SQ_R = \hat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{Y} - \dfrac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \hat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{Y} - n\bar{y}^2$, a soma de quadrados da regressão, mede a quantidade de variação da variável dependente explicada pela equação de regressão (o modelo);
- $SQ_E = \boldsymbol{Y}^T \boldsymbol{Y} - \hat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{Y}$, a soma de quadrados do erro (residual), é a variação devida ao erro, ou seja, mede a variação não explicada pela regressão (pelo modelo).

O procedimento da análise de variância para a regressão linear múltipla tem a seguinte estrutura:

Hipótese nula: $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$;

Hipóteses alternativa: $H_1 : \beta_j \neq 0$ para algum j, $j = 1, \ldots, k$;

Estatística de teste: $F_o = \dfrac{MQ_R}{MQ_E}$, onde $MQ_R = \dfrac{SQ_R}{gl_R}$ e $MQ_E = \dfrac{SQ_E}{gl_E}$;

Critério de rejeição: $f_o > f_t = f_\alpha[k, n-p]$.

Que pode ser sumariado na seguinte tabela Anova:

Fonte de variação (F.V)	Graus de liberdade (gl)	Soma de Quadrados	Média quadrática	F_0
Regressão (modelo)	k	SQ_R	MQ_R	$\dfrac{MQ_R}{MQ_E}$
Erro (residual)	$n-p$	SQ_E	MQ_E	
Total	$n-1$	SQ_T		

Tabela2.4 – Anova para a regressão linear múltipla

Obs.2.2: Atenção que $p = k+1$, $MQ_R = \dfrac{SQ_R}{k}$ e $MQ_E = \dfrac{SQ_E}{n-p}$.

Assim, relativamente à hipótese $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$, rejeita-se a hipótese nula, com um grau de significância α, se o valor da estatística de teste, F_o (valor de F observado a partir da regressão), for maior do que o valor de F_t (valor tabelado para a distribuição F) com $gl_R = k$ e $gl_E = n-k$ graus de liberdade. Resumindo:

i) Se $f_0 \leq f_\alpha[k, n-p] \Rightarrow$ não se rejeita H_0

ii) Se $f_0 > f_\alpha[k, n-p] \Rightarrow$ rejeição de H_0, ou seja, não rejeitar H_1

Caso se verifique ii) conclui-se, com $(1-\alpha)\times 100\%$ de confiança, que o modelo é significativo, isto é, que pelo menos um x_j, $j = 1, \ldots, k$, contribui significativamente para explicar a variação de Y.

Exemplo2.2: Considerando os dados da tabela2.2, testa-se a significância do modelo

$$\hat{y} = 1566{,}08 + 7{,}62x_1 + 8{,}58x_2 ,$$

ou seja, se, através da amostra apresentada na tabela2.2, há evidência de uma relação linear entre a viscosidade do polímero (*Y*), a temperatura de reacção (x_1) e a taxa de alimentação do catalisador (x_2) e se essa relação pode ser descrita por esta última equação.

As hipóteses a testar são

$$H_0 : \beta_1 = \beta_2 = 0 ;$$
$$H_1 : \beta_j \neq 0 \text{ para algum } j , j = 1,2 .$$

Obtém-se a seguinte tabela Anova:

F.V	gl	SQ	MQ	F_0
Regressão	2	44157,09	22078,54	82,50***
Erro	13	3478,85	267,60	
Total	15	47635,94		

Tabela2.5 – Anova para a equação referente aos dados ao exemplo2.1

Da tabela, $f_0 = 82{,}50$, como $f_0 > f_t = f_{0,05}[2,13] = 3{,}80$, rejeita-se a hipótese nula, H_0, para $\alpha = 0{,}05$. Conclui-se com 95% de confiança que a viscosidade do polímero está linearmente relacionada (pela equação de regressão linear) com a reacção da temperatura e com a taxa de alimentação do catalisador (a viscosidade do polímero é, de alguma maneira, explicada por este modelo pelo menos por uma desta duas variáveis independentes). Repare-se que $f_{0,01}[2,13] = 6{,}70$ e $f_{0,001}[2,13] = 12{,}31$ o que quer dizer que se rejeita H_0 com 99,9% de confiança, ou seja, a estatística de teste é altamente significativa $(f_0 = 82{,}50***)$ consequentemente o modelo de regressão é altamente significativo.

O facto do modelo de regressão ser altamente significativo não implica necessariamente que a relação encontrada seja o modelo mais adequado para estimar e/ou predizer a viscosidade do polímero (valores da variável resposta) em função da reacção da temperatura e da taxa de alimentação do catalisador. São necessários outros testes, para se concluir sobre a qualidade do ajustamento, antes de se utilizar este modelo numa situação prática (ou até mesmo, a comparação com outros modelos). Saliente-se, ainda, que a partir deste teste não é possível saber se se pode eliminar alguma das variáveis do modelo, tal facto poderá acontecer se a contribuição de um dos regressores na explicação da variável resposta seja estatisticamente não significativa.

2.3.3 Coeficiente de determinação

Tal como no modelo de regressão simples o coeficiente de determinação é dado por

$$R^2 = \frac{SQ_R}{SQ_T} = 1 - \frac{SQ_E}{SQ_T}. \qquad \textbf{(2.19)}$$

Este coeficiente é uma medida da proporção da variação da variável resposta *Y* que é explicada pela equação de regressão quando estão envolvidas as variáveis independentes $x_1, x_2, \ldots, x_k$. Devido à identidade da análise de variância $0 \le R^2 \le 1$. Contudo, um grande valor de R^2 não implica necessariamente que o modelo de regressão seja um bom ajustamento, uma vez que a adição de uma variável aumenta sempre o valor deste coeficiente (a adição de uma variável ao modelo faz sempre com que a soma de quadrados da regressão aumente), sem ter em conta se a variável que se adiciona é ou não estatisticamente significativa. Assim, modelos com um elevado valor de R^2 podem produzir predições pouco fiáveis de novas observações ou estimativas pouco fiáveis do valor esperado de *Y*. Por este motivo R^2 não será um bom indicador do grau de ajustamento do modelo. Por este facto, alguns investigadores preferem utilizar o coeficiente de determinação ajustado

$$R^2_{\text{ajust.}} = 1 - \frac{\frac{SQ_E}{n-p}}{\frac{SQ_T}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right)\left(1 - R^2\right). \qquad \textbf{(2.20)}$$

Este coeficiente dá uma melhor ideia da proporção de variação de *Y* explicada pelo modelo de regressão uma vez que tem em conta o número de regressores. Ao contrário do que acontecia com o coeficiente de determinação múltiplo, $R^2_{\text{ajust.}}$ não aumenta sempre, quando uma nova variável é adicionada ao modelo. Este só aumenta se de alguma maneira houver vantagem na adição de uma nova variável. De facto, se forem adicionados termos desnecessários, o valor de $R^2_{\text{ajust.}}$, na maior parte dos casos decresce. Quando a diferença entre R^2 e $R^2_{\text{ajust.}}$ é acentuada, há uma boa hipótese de que tenham sido incluídos no modelo termos estatisticamente não significativos.

Exemplo2.3: No exemplo2.1 viu-se que a relação entre a viscosidade do polímero, a temperatura de reacção e a taxa de alimentação do catalisador pode ser representada pelo modelo $\hat{y} = 1566{,}08 + 7{,}62x_1 + 8{,}58x_2$. No exemplo2.2, viu-se que o modelo é altamente significativo. Queremos, agora, saber qual a percentagem de contribuição do modelo para a explicação da variação da viscosidade do polímero. O valor do coeficiente de determinação múltiplo é dado por

$$R^2 = \frac{SQ_R}{SQ_T} = \frac{44157{,}09}{47635{,}94} = 0{,}9270.$$

Apesar do valor de R^2 ser, de alguma forma, elevado este pode não ser um bom indicador do grau de ajustamento do modelo (como foi referido). Vamos calcular o coeficiente de determinação ajustado,

$$R^2_{\text{ajust.}} = 1 - \left(\frac{n-1}{n-p}\right)\left(1-R^2\right) = 1 - \frac{15}{13}\left(1-0{,}92697\right) = 0{,}9157\,,$$

que é um valor próximo de R^2. Podemos concluir que o modelo é um bom ajustamento aos dados. Cerca de 92% da variação da viscosidade do polímero é explicada pelo modelo (pela reacção da temperatura e pela taxa de alimentação do catalisador quando considerados em conjunto neste modelo de regressão).

Exercício2.1: Verifique se o modelo envolvendo as variáveis y e x_2 é significativo. Conclua relativamente à inclusão de x_1 no modelo.

No final da secção2.2 falou-se dos coeficientes de regressão padronizados a partir dos quais é possível comparar o peso da contribuição de cada variável independente na explicação da variação da variável dependente. Contudo, a partir da equação padronizada não é possível saber a percentagem dessa contribuição (sabemos qual o regressor que mais contribui mas não se sabe o valor dessa contribuição). Uma primeira abordagem para se ter uma ideia da percentagem de contribuição de cada regressor será recorrendo aos coeficientes de determinação parcial. Consideram-se regressões lineares simples entre a variável dependente e os diferentes regressores e depois calculam-se e comparam-se esses coeficientes de determinação. Ficaremos, assim, a saber a percentagem de contribuição de cada regressor, caso seja considerado separadamente dos restantes, para a explicação da variação de Y , e poderemos tirar ilações sobre a possível exclusão de determinado regressor da equação de regressão linear múltipla.

Exemplo2.4: No exemplo2.1, viu-se através da equação de regressão padronizada $y' = 0{,}9237x_1' + 0{,}26392x_2'$, que a contribuição da variável x_1 é maior na explicação da variação de Y do que a contribuição de x_2. A ordem de grandeza dos coeficientes de regressão estandardizados está de acordo com a ordem de grandeza dos coeficientes de determinação parcial, $r_1^2 = 0{,}8574$ e $r_2^2 \simeq 0{,}0738$.

Obs.2.3: A soma dos diferentes coeficientes de determinação parciais não é igual ao valor do coeficiente de determinação múltiplo.

Obs.2.4: Avaliando a soma de quadrados do erro é possível verificar qual a variável que mais contribui para a explicação da variação da variável resposta, isto acontece para a variável que tenha, como é lógico, a menor soma de quadrados do erro.

2.3.4 Testes de significância para os coeficientes de regressão

Na regressão múltipla, temos muitas vezes o interesse de testar hipóteses sobre os coeficientes de regressão. Estas constituem outra maneira de se determinar o potencial de cada regressor no modelo de regressão. Por exemplo, o modelo pode tornar-se mais eficaz com a inclusão de novas variáveis ou com a exclusão de uma ou mais variáveis existentes no modelo. A inclusão de uma variável ao modelo faz, sempre, aumentar a soma de quadrados da regressão e diminuir a soma de quadrados do erro. Deve-se decidir se o aumento na soma de quadrados do erro justifica a inclusão da nova variável. Para além disso, a inclusão de uma variável pouco importante ao modelo pode fazer com que a média quadrática do erro aumente, fazendo decrescer a utilidade do modelo (isto é, indicando que esta inclusão não faz sentido, uma vez que "empobrece" o grau de ajustamento do modelo aos dados).

Um teste de significância para os coeficientes de regressão, os β_j´s, é elaborado de modo seguinte

Hipótese nula $H_0 : \beta_j = 0$;

Hipótese alternativa: $H_1 : \beta_j \neq 0$;

Estatística de teste: $T_0 = \dfrac{\hat{\beta}_j}{se(\hat{\beta}_j)}$, onde $se\left(\hat{\beta}_j\right)$ é dado por (2.15);

Região de rejeição: $|t_0| > t_t = t_{\alpha/2}[n-p]$.

Se H_0 não for rejeitada, isto indica que o regressor x_j pode ser "eliminado" do modelo. Este teste é considerado parcial uma vez que o coeficiente de regressão $\hat{\beta}_j$ depende de todos os outros regressores x_i $(i \neq j)$ existentes no modelo.

Obs.2.5: Caso na equação de regressão linear múltipla um dos coeficientes de regressão tenha um valor próximo de zero não quer dizer que a variável correspondente possa ser eliminado do modelo. Devemos ter em conta que as variáveis independentes podem ter diferentes unidades de medida e portanto os respectivos coeficientes de regressão ordens de grandeza diferentes.

Exemplo2.5: Vimos nos exemplos2.1 e 2.4 que a variável x_1 contribui mais para a explicação da variação da variável resposta do que a variável x_2, vamos proceder a um teste de hipóteses (teste *t*) para confirmar ou não a utilização da variável x_2 no modelo de regressão linear.

As hipóteses a testar são

$$H_0 : \beta_2 = 0 \text{ vs } H_1 : \beta_2 \neq 0$$

Para o cálculo da estatística de teste são necessárias as quantidades, $\hat{\sigma}^2 = MQ_E = \frac{SQ_E}{n-p} = 267{,}60$ (que pode ser obtida da tabela Anova) e $C_{22} = 0{,}02222381$ que é o elemento da diagonal principal de $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ que corresponde a $\hat{\beta}_2$ (apresentada no exemplo2.1) e $\hat{\beta}_2 = 8{,}584846$. Assim

$$se(\hat{\beta}_2) = \sqrt{\hat{\sigma}^2 C_{22}} = \sqrt{267{,}60 \times 0{,}02222381} = 2{,}438684$$

vindo

$$T_0 = \frac{\hat{\beta}_2}{se(\hat{\beta}_2)} \Leftrightarrow t_0 = \frac{8{,}584846}{2{,}438684} = 3{,}5203 **.$$

Para $\alpha = 0{,}05$, como $t_{0{,}025}[13] = 2{,}1604$, rejeita-se H_0 com 95% de confiança. Por outro lado, como $t_{0{,}005}[13] = 3{,}0123$ e $t_{0{,}0005}[13] = 4{,}2209$, temos $t_{0{,}005}[13] < t_0 < t_{0{,}0005}[13]$ e, portanto, rejeitamos H_0 com 99%, mas não com 99,9% de confiança. Conclui-se que a variável x_2 contribui de alguma forma para o modelo, cabendo, ao investigador decidir se compensa ou não a sua utilização no modelo.

É interessante verificar que, para a viscosidade do polímero (*y*) e temperatura de reacção (x_1), obtém-se $r^2_{ajust.} = 0{,}8472$ (exercício), adicionando a variável x_2 ao modelo este coeficiente aumenta de 0,8472 para $R^2_{ajust.} = 0{,}9157$ consequentemente, poderá ser aconselhável incluir esta variável no modelo. Por outro lado, considerando o modelo original envolvendo a viscosidade do polímero (*y*) e a taxa de alimentação do catalisador (x_2), obtém-se $r^2_{ajust.} = 0{,}0076$, ou seja, a introdução da variável x_1, no modelo, faz com $R^2_{\text{ajust.}}$ aumente muito mais, passa de 0,0076 para 0,9157 (será bastante aconselhável a inclusão desta variável no modelo ou a sua não exclusão).

Exercício2.2: Fazer o mesmo teste para a variável x_1 e tirar as respectivas conclusões.

2.3.5 Intervalos de confiança

2.3.5.1 Intervalos de confiança para os coeficientes de regressão

Nestes modelos é útil construir intervalos de confiança para as estimativas dos coeficientes de regressão. Vimos que $\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2 \boldsymbol{C}\right)$ então cada estatística

$$\frac{\hat{\beta}_j - \beta_j}{se\left(\hat{\beta}_j\right)},\ j = 0,1,\ldots,k\,, \tag{2.21}$$

tem uma distribuição t com $n-p$ graus de liberdade, onde $se\left(\hat{\beta}_j\right)$, $j = 0,1,\ldots,k$, é dado por (2.15).

Os extremos do intervalo de confiança a $100(1-\alpha)\%$ para os coeficientes de regressão, os β_j´s, com $j = 0,1,\ldots,k$, são

$$\hat{\beta}_j \pm t_{\alpha/2}[n-p]se\left(\hat{\beta}_j\right). \tag{2.22}$$

Exemplo2.6: Vamos construir um intervalo de confiança a 95% para β_1 referente ao nosso exemplo de regressão linear múltipla (dados da tabela2.2). Como se viu, $\hat{\beta}_1 = 7,6213$, o elemento da diagonal de $\boldsymbol{C} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$ correspondente a β_1 é $C_{11} = 1,429184 \times 10^{-3}$ (ver exemplo2.1), $\hat{\sigma}^2 = MQ_E = 267,6039$. Como $t_{0,025}[13] = 2,1604$, os extremos do intervalo de confiança a 95% para β_1 são dados por

$$7,6213 \pm 2,1604\sqrt{267,6039 \times 1,429184 \times 10^{-3}}$$

donde o intervalo de confiança é $[6,2852; 8,9573]$. Conclui-se com 95% de confiança que o valor de β_1 (da população) se encontra neste intervalo .

Convém salientar que existe uma relação fundamental entre os testes de hipóteses e os intervalos de confiança, essa pode ser enunciada nos termos seguintes: uma hipótese nula $H_0 : \beta_j = 0$ pode ser rejeitada a um nível de significância α se, e só se, o intervalo de confiança de β_j a $100(1-\alpha)\%$ não incluir o valor 0. Note-se que esta condição impõe que o intervalo de confiança seja compatível com a natureza de H_1, ou seja, para testes bilaterais se construam intervalos de confiança bilaterais e para testes unilaterais (num sentido) se construam intervalos de confiança unilaterais (no mesmo sentido). A implicação essencial desta relação é que se pode proceder ao teste de hipóteses recorrendo a intervalos de confiança.

2.3.5.2 Intervalos de confiança para os valores esperados de *Y*

No caso da regressão linear múltipla pode obter-se um intervalo de confiança para o valor esperado da resposta dado um determinado $\boldsymbol{x}_0$. Este último, não é mais do que uma linha da matriz $\boldsymbol{X}$, ou seja, o vector, $\boldsymbol{x}_0^T = [1 \quad x_{01} \quad x_{02} \quad \ldots \quad x_{0k}]$. O valor esperado para Y considerando $\boldsymbol{x}_0$, é $E[Y \mid \boldsymbol{x}_0] = \mu_{Y|\boldsymbol{x}_0} = \boldsymbol{x}_0^T \boldsymbol{\beta}$, que é estimado por $\hat{\mu}_{Y|\boldsymbol{x}_0} = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}$. Este estimador é não enviesado, uma vez que $E\left[\hat{\mu}_{Y|\boldsymbol{x}_0}\right] = \mu_{Y|\boldsymbol{x}_0}$ e tem variância $V\left[\hat{\mu}_{Y|\boldsymbol{x}_0}\right] = \hat{\sigma}^2 \boldsymbol{x}_0^T \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{x}_0 = \hat{\sigma}^2 \boldsymbol{x}_0^T \boldsymbol{C}\, \boldsymbol{x}_0$, podemos definir o seguinte erro padrão

$$se\left(\hat{\mu}_{Y|\boldsymbol{x}_0}\right) = \sqrt{\hat{\sigma}^2 \boldsymbol{x}_0^T \boldsymbol{C}\, \boldsymbol{x}_0} \quad . \tag{2.23}$$

O intervalo de confiança a $100(1-\alpha)\%$ para $\mu_{Y|\boldsymbol{x}_0}$ pode ser construído a partir da estatística

$$\frac{\hat{\mu}_{Y|\boldsymbol{x}_0} - \mu_{Y|\boldsymbol{x}_0}}{se\left(\hat{\mu}_{Y|\boldsymbol{x}_0}\right)} \tag{2.24}$$

que segue uma distribuição t com $n-p$ graus de liberdade. Os extremos do intervalo de confiança para o valor esperado da resposta para um determinado ponto $\boldsymbol{x}_0$, são dados por

$$\hat{\mu}_{Y|\boldsymbol{x}_0} \pm t_{\alpha/2}[n-p]\, se\left(\hat{\mu}_{Y|\boldsymbol{x}_0}\right). \tag{2.25}$$

Exemplo2.7: Vamos construir um intervalo de confiança a 95% para o valor esperado da viscosidade do polímero quando $x_1 = 80$ e $x_2 = 8$, ou seja, para $\boldsymbol{x}_0^T = [1 \quad 80 \quad 8]$. Uma estimativa do valor esperado para a variável resposta é

$$\hat{\mu}_{Y|\boldsymbol{x}_0} = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} = [1 \quad 80 \quad 8] \begin{bmatrix} 1566,07777 \\ 7,62129 \\ 8,58485 \end{bmatrix} = 2244,46$$

(este valor poderia ter sido obtido através da equação de regressão)

$$V\left[\hat{\mu}_{Y|\boldsymbol{x}_0}\right] = \hat{\sigma}^2 \boldsymbol{x}_0^T \boldsymbol{C} \boldsymbol{x}_0 = 267,6039 [1 \quad 80 \quad 8] \times \begin{bmatrix} 14,176004 & -0,129746 & -0,223453 \\ -0,129746 & 1,429184\times10^{-3} & -4,763947\times10^{-5} \\ -0,223453 & -4,763947\times10^{-5} & 2,222381\times10^{-2} \end{bmatrix} \begin{bmatrix} 1 \\ 80 \\ 8 \end{bmatrix}$$

ou seja,

$$V\left[\hat{\mu}_{Y|\boldsymbol{x}_0}\right] = 267,6039 \times 0,349519 = 93,53263 .$$

Os extremos do intervalo de confiança a 95% para o valor esperado da variável resposta para o ponto $\boldsymbol{x_0}$, isto é, para $x_1 = 80$ e $x_2 = 8$, são $2244,46 \pm 2,160368\sqrt{93,53263}$ logo o intervalo é $[2223,57; 2265,35]$. Tem-se, pois, 95% de confiança que o verdadeiro valor de *Y*, para estes valores particulares das variáveis independentes, está neste intervalo.

2.3.6 Predição de novas observações

Tal como na regressão linear simples, o modelo de regressão múltiplo pode ser utilizado para se "predizer" uma determinada resposta de *Y* que será observada no futuro, com base em valores particulares das variáveis independentes, $x_{01}, x_{02}, \ldots, x_{0k}$. Se $\boldsymbol{x_0^T} = [1 \quad x_{01} \quad x_{02} \quad \ldots \quad x_{0k}]$, uma estimativa de uma observação futura Y_0 para $x_{01}, x_{02}, \ldots, x_{0k}$ é dada por $\hat{y}_0 = \boldsymbol{x_0^T \hat{\beta}}$. O erro de predição, ε_0, para a observação futura, é $\varepsilon_0 = Y_0 - \hat{Y}_0$, onde Y_0 e $\hat{Y}_0$ são variáveis aleatórias normalmente distribuídas, assim como o erro. Nestes termos, prova-se que o erro de predição de uma valor particular de *Y* é uma variável aleatória normal com média zero $E[\varepsilon_0] = 0$ e variância $V[\varepsilon_0] = \sigma^2\left[1 + \boldsymbol{x_0^T}\left(\boldsymbol{X^T X}\right)^{-1}\boldsymbol{x_0}\right] = V[\varepsilon_0] = \sigma^2\left[1 + \boldsymbol{x_0^T C x_0}\right]$, pode, então, definir-se o erro padrão

$$se(\varepsilon_0) = \sqrt{\sigma^2\left[1 + \boldsymbol{x_0^T C x_0}\right]}\ . \tag{2.26}$$

Consequentemente,

$$Z = \frac{Y_0 - \hat{Y}_0}{se(\varepsilon_0)}, \tag{2.27}$$

é uma variável aleatória com distribuição normal reduzida. Substituindo σ^2 por $\hat{\sigma}^2$ obtém-se

$$T = \frac{Y_0 - \hat{Y}_0}{\sqrt{\sigma^2\left[1 + \boldsymbol{x_0^T C x_0}\right]}}, \tag{2.28}$$

que é uma variável aleatória com distribuição *t* de Student com $(n - p)$ graus de liberdade. Este resultado pode ser utilizado para limitar o erro de predição e construir intervalos de predição para a variável aleatória *Y* segundo os métodos de construção de intervalos de confiança. Prova-se, com probabilidade $(1 - \alpha)$, que é de esperar que o erro de predição seja menor em valor absoluto do que

$$t_{\frac{\alpha}{2}}[n - p]\sqrt{\hat{\sigma}^2\left[1 + \boldsymbol{x_0^T C x_0}\right]}\ . \tag{2.29}$$

Os extremos do intervalo predição a $100(1-\alpha)\%$ para uma observação "futura" de Y correspondente a um determinado valor de $\boldsymbol{x}_0$ na regressão linear múltipla são

$$\hat{y}_0 \pm t_{\frac{\alpha}{2}}[n-p]\sqrt{\hat{\sigma}^2\left[1+\boldsymbol{x}_0^T\boldsymbol{C}\,\boldsymbol{x}_0\right]} \quad . \qquad \textbf{(2.30)}$$

Exemplo2.8 : Supondo que se pretende construir um intervalo de predição a 95% para a viscosidade do polímero quando $x_1 = 80$ e $x_2 = 8$ (supondo que a experiência ainda não foi efectuada para estes valores), ou seja, para $\boldsymbol{x}_0^T = [1 \quad 80 \quad 8]$. Uma estimativa da viscosidade do polímero para estes valores das variáveis independentes é dada por $\hat{y}_0 = \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} = 2244,64$ e $\hat{\sigma}^2 = 267,6039$ tem-se ainda que $\boldsymbol{x}_0^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0 = 0,349519$ (ver exemplo2.7). Assim, os extremos para o intervalo de predição a 95% para o valor "futuro" da variável resposta dado $\boldsymbol{x}_0$ são

$$2244,64 \pm 2,160368\sqrt{267,6039(1+0,349519)}$$

sendo o, respectivo, intervalo

$$[2203,41;\ 2285,52].$$

Temos, então, 95% de confiança que o verdadeiro valor desta observação "futura" se encontra neste intervalo.

Saliente-se que a amplitude do intervalo de predição é maior que a amplitude do intervalo de confiança para o valor esperado quando calculado para os mesmos valores das variáveis independentes.

Na predição de novas observações e na estimação de valores esperados para a variável resposta, Y, para um dado ponto $x_{01}, x_{02}, \ldots, x_{0k}$, deve-se ter cuidado quando este ponto não se encontra na região que contém as observações originais ou iniciais (extrapolar fora da região) . É muito possível que um modelo que se ajuste bem para os dados iniciais, não se ajuste bem a dados fora a desta região (a outras quaisquer observações).

2.3.7 Análise de resíduos

Para a construção dos modelos de regressão linear foram consideradas algumas hipóteses relativamente aos resíduos. Prioritariamente os resíduos foram considerados independentes, e $\varepsilon_i \sim N(0,\sigma^2)$, $i=1,\ldots,n$,. Esta última suposição é absolutamente exigida para a construção, por exemplo, dos testes *F*. Em consequência, se o modelo ajustado for o correcto os resíduos devem evidenciar tendências que confirmem, ou pelo menos não desmintam, as suposições feitas. Assim, ainda que, eventualmente, com base num qualquer teste não haja razão para duvidar de que o modelo seja adequado não se deve prescindir da análise dos resíduos.

Os resíduos de um modelo de regressão representam as diferenças entre aquilo que foi realmente observado e o que foi estimado através da equação de regressão, ou seja, a quantidade que a equação de regressão não foi capaz de explicar, i.e., $e_i = y_i - \hat{y}_i$, $i=1,\ldots,n$, onde y_i é uma determinada observação e $\hat{y}_i$ é o seu correspondente valor ajustado (estimado) através do modelo de regressão. Pode-se, assim, pensar nos resíduos como sendo os erros observados se o modelo é correcto. A análise dos resíduos é útil para se verificar se estes erros têm, aproximadamente, uma distribuição normal com média zero e variância constante, bem como para se determinar se a inclusão/exclusão de novos termos ao modelo se torna útil.

Na análise dos resíduos, quando se tenta saber se as suposições feitas são violadas, ou se conclui que elas parecem ser violadas ou se conclui que essas não parecem ser violadas. Esta última situação não significa que estejamos a concluir que as suposições são correctas mas tão só que, tomando por base os dados, não temos razões para afirmar que elas sejam incorrectas. Mesmo que o modelo seja significativo e correcto não significa que ele seja o modelo adequado, mas apenas um plausível que não foi declarado incorrecto através dos dados. Se foi declarado não ajustado, existência de falta de ajustamento, torna-se necessário um modelo diferente, eventualmente um modelo quadrático.

Vamos considerar duas maneiras de analisar os resíduos:

i) Analíticamente

Analiticamente, poderiam aplicar-se testes de qualidade de ajustamento de qui-quadrado (χ^2) ou de Kolmogorov-Smirnov (K-S), por exemplo, para inferir em relação à suposição da normalidade. Se o modelo for correcto os *n* resíduos devem assemelhar-se a *n* observações de uma distribuição normal com média zero e variância σ^2.

Outra maneira, será estandardizando os resíduos, fazendo

$$d_i = \frac{e_i}{\hat{\sigma}}, \quad i = 1, \ldots, n,$$

com $\hat{\sigma} = \sqrt{MQ_E}$. Estes resíduos estandardizados tem média zero e variância aproximadamente 1 (consequentemente, são úteis na identificação de *outliers*). Se os erros forem normalmente distribuídos, então aproximadamente 95% dos resíduos estandardizados deverão estar no intervalo $[-2, 2]$. Os resíduos que se encontrem muito afastados dos extremos deste intervalo, poderão indicar a presença de *outliers*; isto é, observações (correspondentes) que não segue os "padrões" das restantes. Os *outliers* deverão ser examinados com muito cuidado, uma vez que podem representar erros, tais como; erros de registo, erros da própria natureza dos dados, ou outros mais ou menos graves. Nestes termos, providenciam, muitas vezes, informação importante acerca de circunstâncias fora do usual, de interesse para os investigadores e, como tal, deverão ser tomados em conta.

ii) Graficamente

A maneira mais simples de analisar os resíduos consiste na sua representação gráfica podendo esta ser feita de diferentes modos, entre os quais:

- Construção de um histograma de frequências para os resíduos;
- Representação dos resíduos num gráfico cartesiano, ou "normal probability plot of residual";
- Representação gráfica dos resíduos *versus* valores ajustados, ou, *versus* a variável independente.

Quando na regressão se trabalha com amostras de tamanho reduzido, o que acontece muitas vezes, o que faz com que o histograma da distribuição não seja muito conclusivo quanto à normalidade dos resíduos, a representação destes últimos num gráfico cartesiano é uma boa alternativa (muitos programas estatísticos apresentam estes gráficos como "normal probability plot residual"). Sendo o modelo correcto, uma vez que, os n resíduos deverão representar n observações de uma variável aleatória $N(0, \sigma^2)$, se eles forem representados em papel de probabilidade normal deverá essa representação corresponder, aproximadamente, a uma linha recta.

Por exemplo, na representação gráfica dos resíduos *versus* os valores ajustados (estimados) representam-se os pares ordenados $(\hat{y}_i, e_i)$, $i = 1, \ldots, n$, a forma da "nuvem" de pontos representada indicar-nos-á, em certas situações empiricamente detectadas, alguma informação específica sobre o modelo ajustado. Vejamos duas dessas situações:

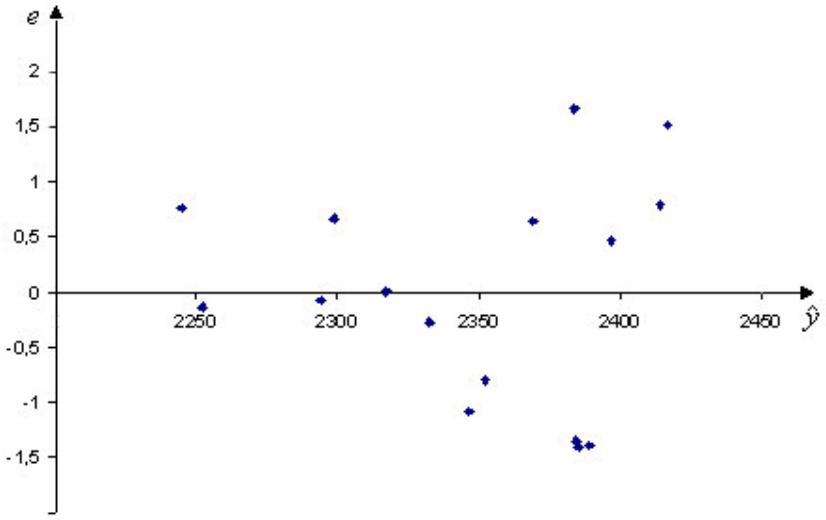

Figura2.1: Representação dos resíduos de um modelo adequado

A figura 2.1 ilustra que não há razão para duvidar do modelo proposto, nem da homocedasticidade (homogeneidade de variâncias), $V[e_i]=\sigma^2$, $i=1,\ldots,n$, (representa a situação ideal). A banda horizontal dos pontos do gráfico não revela falta de normalidade.

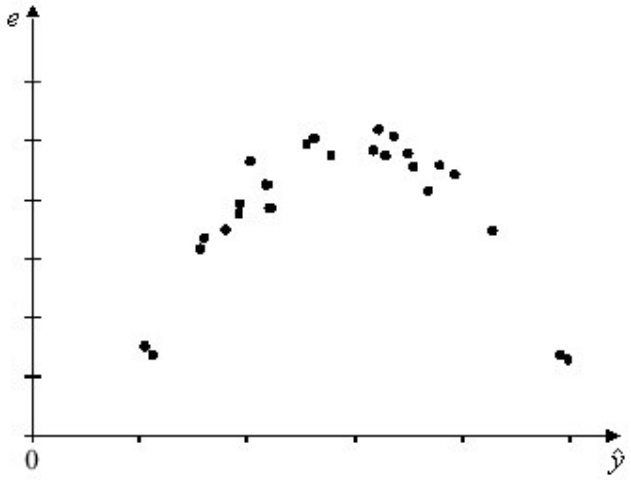

Figura2.2: Representação dos resíduos de um modelo não adequado

A figura2.2 ilustra um caso em que os resíduos não seguem uma distribuição normal. Aqui o modelo de regressão linear apresenta-se inadequado. Não deve ser utilizado, ou então, devem ser incluídos/excluídos, no modelo, novos regressores. Caso se proceda à inclusão de novos regressores esses poderão ser, por exemplo, termos em potências e/ou produtos das variáveis. Em conformidade, para um modelo da forma $Y=\beta_0+\beta_1x_1+\beta_2x_2+\varepsilon$ poderia tentar-se, por exemplo, um modelo da forma $Y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2+\beta_4x_2^3+\varepsilon$. Pode, ainda, proceder-se à transformação das observações y_i ou x_i, antes da análise.

Exercício2.3: Construa o gráfico de dispersão dos resíduos estandardizados obtidos para o modelo de regressão linear múltiplo associado aos dados da tabela2.2. Verifique se 95% dos destes resíduos se encontram no intervalo $[-2, 2]$.

Obs.2.6: Caso o gráfico de dispersão revele violação da hipótese de homocedastidade, (variância não constante), por exemplo, a variância das observações pode estar a aumentar com o tempo. Provável necessidade de transformação da variável resposta *y*, visando eliminar este "problema", antes de se realizar a análise de regressão. As transformações mais utilizadas são $\sqrt{y}$, $\ln y$ ou $\frac{1}{y}$, contudo existem métodos para se escolher uma transformação apropriada.