Regressão Múltipla

Parte I: Modelo Geral e Estimação

Regressão linear múltipla

Exemplos:

• Num estudo sobre a produtividade de trabalhadores (em aeronave, navios) o pesquisador deseja controlar o número desses trabalhadores e o bônus pago (remuneração).

• Num estudo sobre a resposta à uma droga, o pesquisador deseja controlar as doses da droga e o método de aplicação.

▪ Num estudo sobre o tempo de CPU, para avaliar a demanda por recursos, o pesquisador decidiu verificar o efeito de X_1=*disk I/O* e X_2=*memory size.*

▪ Em todos os exemplos foram necessárias várias variáveis preditoras no modelo para um bom ajuste do mesmo.

▪ Um modelo contendo várias variáveis preditoras resulta numa estimação mais acurada.

Modelo de regressão com duas variáveis preditoras

O modelo de regressão linear é dado por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad (1)$$

onde Y_i é a resposta da i-ésima observação, X_{i1} e X_{i2} são os valores das duas variáveis preditoras da i-ésima observação. Os parâmetros do modelo são β_0, β_1, β_2 e o termo do erro é ε_i.

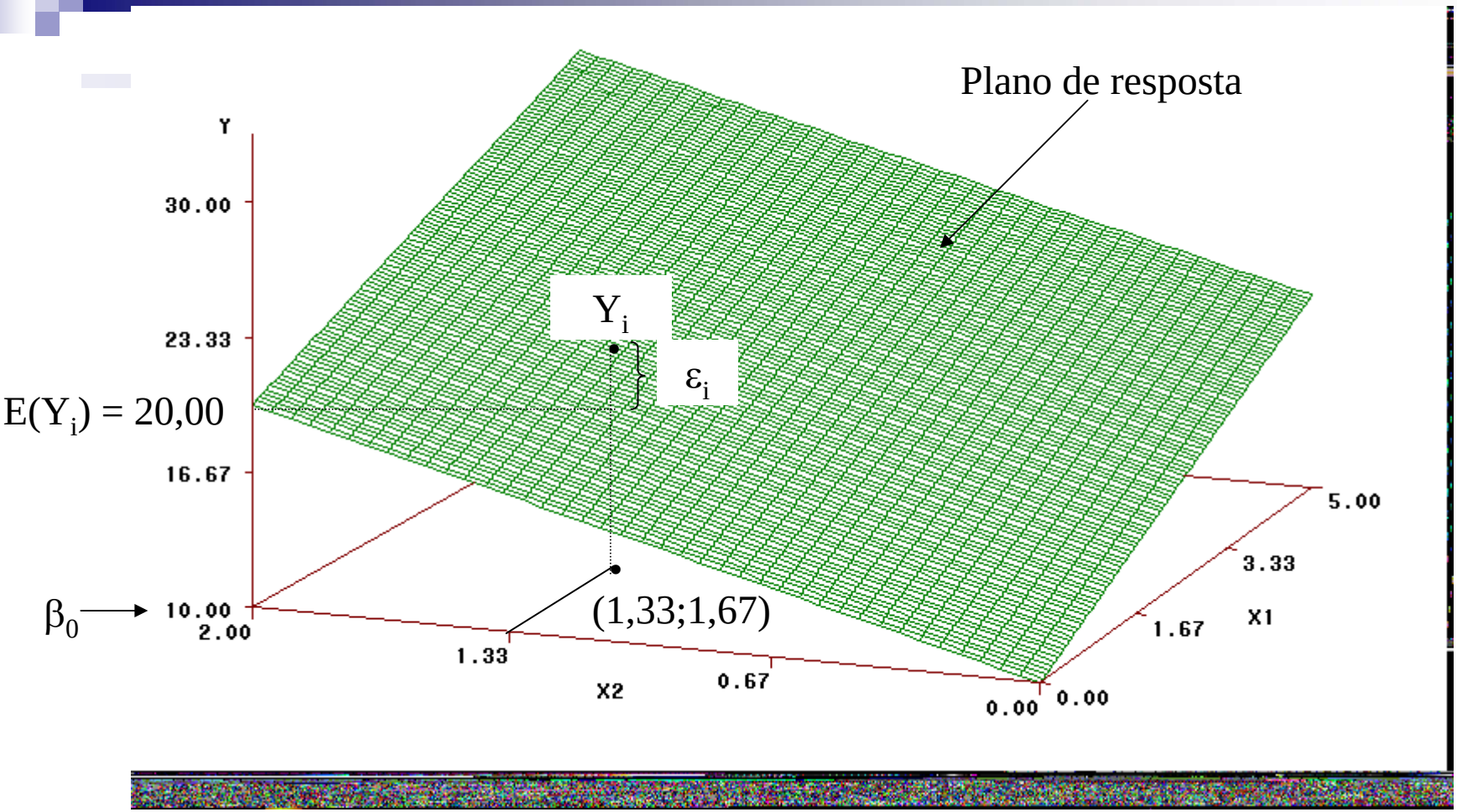
Vamos assumir que $E(\varepsilon_i)=0$, portanto, a função de regressão do modelo de primeira ordem é:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \qquad (2)$$

A representação gráfica desta função é um plano no espaço. A figura seguinte mostra este plano para a função:

$$E(Y) = 10 + 2X_1 + 5X_2 \qquad (3)$$

A função de regressão na regressão múltipla é chamada de ***superfície de resposta.***


Plano de resposta
Y
30.00
23.33
16.67
10.00
Y_i
ε_i
$E(Y_i) = 20,00$
β_0
(1,33;1,67)
2.00
1.33
0.67
0.00
X2
0.00
1.67
3.33
5.00
X1

Significado dos coeficientes de regressão:

• O parâmetro β_0 é o intercepto do plano de regressão. Se a abrangência do modelo inclui $X_1=0$ e $X_2=0$ então $\beta_0=10$ representa a resposta média $E(Y)$ neste ponto.

• O parâmetro β_1 indica a mudança na resposta média $E(Y)$ por unidade de acréscimo em X_1 quando X_2 é mantido constante. Da mesma forma β_2 indica a mudança na resposta média por unidade de aumento em X_2 quando X_1 é mantido constante.

• Neste modelo, o efeito de X_1 sobre a resposta média não depende de X_2 e vice-versa, assim, dissemos que as variáveis preditoras tem efeito aditivo ou não interagem. Temos um modelo sem interação.

Modelo linear geral de regressão

Vamos supor que temos $X_1, X_2,..., X_{p-1}$ variáveis preditoras. Vamos definir o modelo de regressão, com erros normais, em termos das variáveis preditoras:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1} + \varepsilon_i \qquad (4)$$

onde: $\beta_0, \beta_1,..., \beta_{p-1}$, são os parâmetros;

$X_{i1},..., X_{i,p-1}$ são constantes conhecidas;

ε_i são independentes com distribuição $N(0, \sigma^2)$

$i=1,2,...,n.$

A função resposta para o modelo, como $E(\varepsilon_i)=0$, é dada por:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{p-1} X_{p-1} \qquad (5)$$

Algumas situações em que podemos usar o modelo em consideração.

Modelo linear geral de regressão

1) Temos p-1 variáveis preditoras: todas as variáveis preditoras apresentam efeito aditivo, ou seja, não apresentam um efeito de interação entre elas (o efeito de uma variável preditora não depende dos níveis da outra variável preditora).

2) As variáveis preditoras são qualitativas: neste caso temos variáveis como: sexo, invalidez (normal, parcialmente inválido, inválido). Usamos variáveis indicadoras, que recebem valores 0 e 1 para identificar as categorias de uma variável qualitativa.

Exemplo: desejamos fazer uma análise de regressão para estimar a distância de um hospital (*Y*), baseado na cidade dos pacientes (X_1) e sexo (X_2). O modelo de regressão é:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad (6)$$ onde

$$X_{i1} = idade \quad dos \quad pacientes;$$

$$X_{i2} = \begin{cases} 1 \; se \; o \; paciente \quad \acute{e} \; do \; sexo \quad feminino \\ 0 \; se \; o \; paciente \quad \acute{e} \; do \; sexo \quad masculino \end{cases}$$

Modelo linear geral de regressão

A resposta média do modelo (6) é:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad (7)$$

Para pacientes do sexo masculino, $X_2=0$, temos:

$$E(Y) = \beta_0 + \beta_1 X_1 \quad (8)$$

Para pacientes do sexo feminino, $X_2=1$, temos:

$$E(Y) = (\beta_0 + \beta_2) + \beta_1 X_1 \quad (9)$$

As duas funções respostas representam duas retas paralelas com diferentes interceptos.

Outro exemplo: vamos considerar uma terceira variável no modelo, o status sobre a invalidez dos pacientes, a qual apresenta três categorias. Em geral, representamos uma variável qualitativa com *c* categorias, por meio de *c-1* variáveis indicadoras. Portanto, no exemplo, vamos definir as variáveis X_3 e X_4 como:

Modelo linear geral de regressão

$$X_3 = \begin{cases} 1 & \text{Se o paciente é normal} \\ 0 & \text{Se o paciente está em outra categoria} \end{cases}$$

$$X_4 = \begin{cases} 1 & \text{Se o paciente é parcialmente inválido} \\ 0 & \text{Se o paciente está em outra categoria} \end{cases}$$

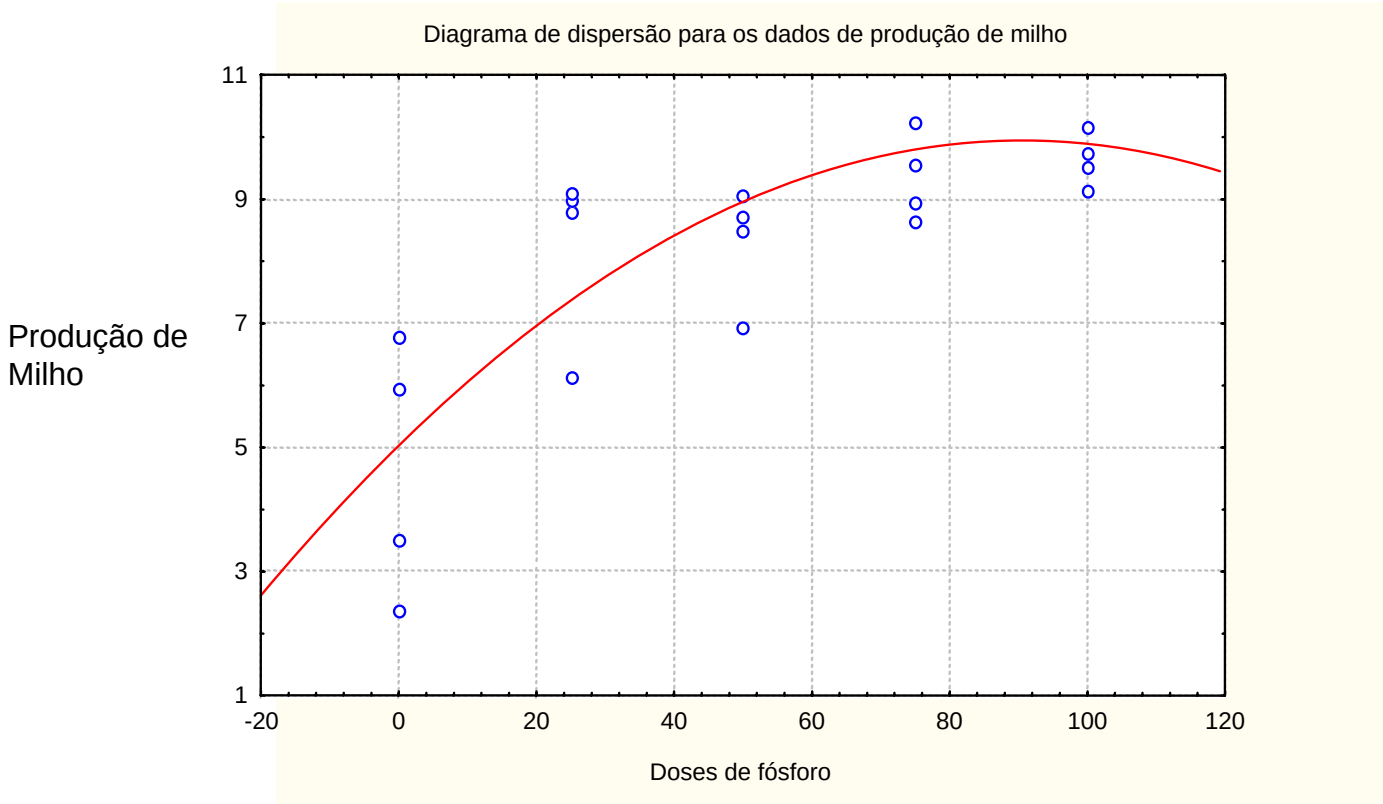
O modelo com idade, sexo e status da invalidez fica:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i \qquad (10)$$

3) Regressão polinomial: contém termos quadráticos e de maior ordem nas variáveis preditoras. Exemplo:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i \qquad (11)$$

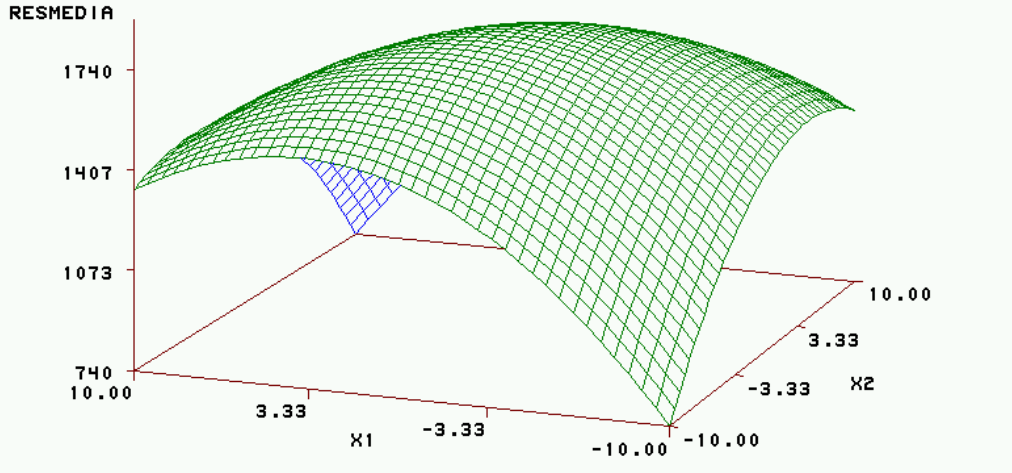
Gráfico: Uma parábola.



Apesar da natureza curvilínea da função resposta do modelo (11) ele é um caso especial do modelo (4). Fazendo-se $X_{i1}=X_i$ e $X_{i2}=X_i^2$, temos o modelo (1).

Modelo usado de segunda ordem com interação:

$$E(Y) = 1740 - 4x_1^2 - 3x_2^2 - 3x_1x_2$$



Observe que o modelo apresenta ponto de máximo em $x_1=0$ $x_2=0$.

Modelo linear geral de regressão

4)Variáveis transformadas: uma transformação bastante utilizada é a logarítmica:

$$Y_i^{'} = \log Y_i$$

O modelo fica:

$$Y_i^{'} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \qquad (12)$$

A função resposta é complexa. Porém, o modelo (12) é da forma do modelo linear geral de regressão.

Exercício: coloque o modelo (13) na forma do modelo de regressão linear geral (4).

$$Y_i = \frac{1}{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}} + \varepsilon_i \qquad (13)$$

Basta fazer:

$$Y_i^{'} = \frac{1}{Y_i} \Rightarrow Y_i^{'} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

Modelo linear geral de regressão

5) Modelos com efeito da interação entre variáveis preditoras. O efeito de uma variável preditora depende dos níveis das outras variáveis preditoras. Exemplo:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i \qquad (14)$$

Observe que fazendo-se $X_{i3}=X_{i1}X_{i2}$ obtemos um caso do modelo linear geral de regressão (4).

6) Combinando modelos:

Exemplo:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2} + \beta_4 X_{i2}^2 + \beta_5 X_{i1} X_{i2} + \varepsilon_i \qquad (15)$$

Fazendo-se:

$$Z_{i1} = X_{i1} \quad Z_{i2} = X_{i1}^2 \quad Z_{i3} = X_{i2} \quad Z_{i4} = X_{i2}^2 \quad Z_{i5} = X_{i1} X_{i2}$$

temos o modelo linear geral de regressão (4).

Modelo de regressão linear múltipla em termos matriciais

A expressão do modelo linear geral de regressão é dada por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \qquad (16)$$

Em termos matriciais, precisamos definir:

$$\underset{n \times 1}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{bmatrix} \qquad \underset{n \times p}{\mathbf{X}} = \begin{bmatrix} 1 & X_{11} & . & . & X_{1,p-1} \\ 1 & X_{21} & . & . & X_{2,p-1} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & X_{n1} & . & . & X_{n,p-1} \end{bmatrix}$$

$$\underset{p \times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_{p-1} \end{bmatrix} \qquad \underset{n \times 1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

Modelo de regressão linear múltipla em termos matriciais

Em termos matriciais, o modelo de regressão linear geral é dado por:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \textbf{(17)}$$

$\boldsymbol{\varepsilon}$ é um vetor de variáveis aleatórias independentes e normalmente distribuídas com esperança (média), $\mathbf{E}(\boldsymbol{\varepsilon})=\mathbf{0}$ e matriz de variância-covariância dada por:

$$\boldsymbol{\sigma}^2(\boldsymbol{\varepsilon}) = \begin{bmatrix} \sigma^2 & 0 & . & 0 \\ 0 & \sigma^2 & . & 0 \\ . & . & . & . \\ 0 & 0 & . & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}$$

Assim, o vetor das observações $\mathbf{Y}$ tem esperança e variância dadas por:

$$\underset{n\ x\ 1}{\mathbf{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}} \qquad \underset{n\ x\ n}{\boldsymbol{\sigma}^2(\mathbf{Y}) = \sigma^2\mathbf{I}} \qquad \textbf{(18)}$$

Estimação dos coeficientes de regressão

O sistema de equações normais para o modelo (17) é:

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y} \qquad (19)$$

E os estimadores de mínimos quadrados são dados por:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad (20)$$

Método de máxima verossimilhança

Vamos considerar o modelo com erros normais (17). A função de máxima verossimilhança é dada por:

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_{p-1} X_{i,p-1})^2 \right] \qquad (21)$$

Os estimadores de máxima verossimilhança são exatamente os mesmos obtidos com o método de mínimos quadrados.

Valores estimados e resíduos

Os valores estimados são obtidos por:

$$\underset{n\ x\ 1}{\hat{\mathbf{Y}}} = \mathbf{Xb} \qquad (22)$$

Os resíduos são obtidos através da expressão matricial:

$$\underset{n\mathbf{x1}}{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{Xb} \qquad (23)$$

Análise de variância

Soma de quadrados e quadrados médios

$$SS_T = \mathbf{Y}'[\mathbf{I} - (\tfrac{1}{n})\mathbf{J}]\mathbf{Y} \quad \textit{com n - 1 graus de liberdade}$$

$$SS_M = \mathbf{Y}'[\mathbf{H} - (\tfrac{1}{n})\mathbf{J}]\mathbf{Y} \quad \textit{com p - 1 graus de liberdade}$$

$$SS_R = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y} \quad \textit{com n - p graus de liberdade}$$

Onde **J** é uma matriz *n* x *n* de un's e **H=X(X'X)$^{-1}$X'** é a matriz de projeção.

$$\hat{y} = Hy$$

Os quadrados médios são dados por:

$$MS_M = \frac{SSM}{p-1}$$

$$MS_R = \frac{SSR}{n-p}$$

Teste F para regressão

Hipóteses em teste:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1} = 0$$

$$H_a : pelo \ \ menos \ \ um \ \ \beta_k \ é \ diferente \ \ \ de \ \ zero.$$

A estatística de teste é dada por:

$$F^* = \frac{MS_M}{MS_R} \qquad (24)$$

Se $F^* > F(\alpha; p-1, n-p)$, rejeitamos a hipótese nula, caso contrário, não rejeitamos a hipótese nula. Não devemos esquecer de usar o valor p.

Coeficiente de determinação (R^2)

Define-se R^2 por:

$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSR}{SST} \qquad (25)$$

Mede a redução da variabilidade total de *Y* associada com o uso do conjunto de variáveis $X_1,...,X_{p-1}$. Como na regressão linear simples, temos:

$$0 \leq R^2 \leq 1$$

Assim, $R^2=0$ se todas as estimativas $\beta_k=0$ $(k=1,...,p-1)$, e $R^2=1$ quando todas as observações *Y* caírem exatamente na superfície de regressão ajustada, isto é, quando:

$$Y_i = \hat{Y}_i \quad para \quad todo \quad i.$$

Como R^2 aumenta com a adição de variáveis explanatórias, sugere-se utilizar o coeficiente de determinação ajustado (corrigido) para os graus de liberdade. O coeficiente de determinação ajustado é dado por:

$$R_a^2 = 1 - \frac{\frac{SSR}{n-p}}{\frac{SST}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSR}{SST} \qquad (26)$$

Coeficiente de determinação (R^2)

Um alto valor de R^2 não necessariamente implica que o modelo ajustado se presta para se fazer inferências precisas, pois apesar de um valor alto de R^2, o MS_R ainda pode ser grande. O modelo pode não ser exatamente linear.

Coeficiente de correlação múltipla (R)

O coeficiente de correlação múltipla mede o relacionamento linear entre Y e $\hat{Y}$.

$$R = \sqrt{R^2} \qquad (27)$$