Modelo de Regressão Linear Simples

O modelo de regressão linear simples considera apenas uma variável explicativa e a função de regressão é linear. O modelo é definido por:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \tag{1}$$

em que:

- Y_i é o valor da variável resposta para a i-ésima observação,
- β_0 é o intercepto e β_1 é o coeficiente angular, ambos são parâmetros desconhecidos,
- X_i é uma constante conhecida, o valor da variável explicativa para a i-ésima observação. X_i é uma variável fixa (ou sem erro ou determinística),
- ε_i são independentes e $N(0, \sigma^2)$.

Alguns Exemplos

A) Um estudo foi realizado para verificar a influência das variáveis: X_1 capital investido e X_2 gasto em publicidade no lucro anual, Y, de empresas.

B) Um experimento foi realizado para estudar a influência das variáveis: X_1 quantidade de adubo, X_2 quantidade de chuva e X_3 tipo de solo, na produção de cana de açúcar, Y.

C) Com o objetivo de verificar o perfil dos vestibulando, uma Universidade estudou quais fatores, como, X_1 nota na redação, X_2 notas médias nas provas de Matemática, Química e Física, X_3 curso pretendido, X_4 sexo e X_5 relação candidato/vaga influênciam na nota, Y, obtida na prova do Vestibular.

Modelo de regressão linear com 2 variáveis explicativas

Quando há duas variáveis explicativas, X_1 e X_2, o modelo de regressão é definido por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i,$$

é conhecido como modelo de primeira ordem, pois é linear nos parâmetros e linear nas variáveis explicativas.

- Os parâmetros desconhecidos do modelo são β_0, β_1 e β_2.
- ε_i é o erro do modelo com $N(0, \sigma^2)$.
- $\mathrm{E}(Y) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$ é conhecida como superfície de regressão ou superfície de resposta.

Modelo de regressão linear com 2 variáveis explicativas

- Quando o escopo do modelo inclui $X_1 = 0$ e $X_2 = 0$, então β_0 representa a média da variável resposta, E(Y). Caso contrário, β_0 não tem interpretação no modelo.
- O parâmetro β_1 indica a mudança no valor esperado de Y, E(Y), com o aumento em uma unidade em X_1, mantendo X_2 constante.
- O parâmetro β_2 indica a mudança no valor esperado de Y, E(Y), com o aumento em uma unidade em X_2, mantendo X_1 constante.
- Quando o efeito de X_1 na resposta média não depende de X_2 e quando o efeito de X_2 na resposta média não depende de X_1, as variáveis explicativas são consideradas ter efeito aditivo.
- Dessa forma, os coeficientes β_1 e β_2 são também conhecidos como coeficientes de regressão parcial.

Modelo de regressão linear geral

Em geral, a variável resposta Y pode estar relacionada com p variáveis explicativas. O modelo de regressão é definido por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i, \qquad (2)$$

é conhecido como modelo de primeira ordem, pois é linear nos parâmetros e linear nas variáveis explicativas. Esse modelo é conhecido como um modelo de regressão linear múltipla ou também como um modelo de regressão linear geral.

- Y_i é o valor da variável resposta para a i-ésima observação,
- β_0, $\beta_1, \ldots, \beta_p$ são parâmetros desconhecidos,
- $X_{i1}, X_{i2}, \ldots, X_{ip}$ são constantes conhecidas,
- ε_i são independentes e $N(0, \sigma^2)$, $i = 1, \ldots, n$.

Modelo de regressão linear geral

Assumindo que $E(\varepsilon_i) = 0$, a função da resposta para o modelo (2), é:

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}. \quad (3)$$

- Essa função resposta é um hiperplano, que é um plano em mais do que duas dimensões.
- O parâmetro β_k indica a mudança no valor esperado de Y, E(Y), com o aumento em uma unidade em X_k, mantendo todas as outras variáveis constantes.
- O efeito das variáveis explicativas no modelo são considerados aditivos.
- O modelo (2) com erro normal implica que as observações Y_i são variáveis normais independentes com $E(Y_i)$ dado em (3) e variância constante, $\text{Var}(Y_i) = \sigma^2$.

Modelo de regressão linear geral

O modelo de regressão linear geral engloba uma variedade de situações. Algumas serão consideradas a seguir:

Variável explicativa Qualitativa

- O modelo (2) considera não apenas variáveis explicativas quantitativas, mas também variáveis qualitativas como, por exemplo: sexo, classe social, entre outras.
- Para a inclusão das variáveis explicativas qualitativas no modelo é necessário usar variável indicadora para identificar a classe/categoria da variável qualitativa.
- Considere uma análise de regressão para prever a duração da internação em um hospital (Y) baseada na idade (X_1) e sexo (X_2) do paciente.
- A variável sexo é definida por:

$$X_2 = \begin{cases} 1 & \text{se sexo feminino} \\ 0 & \text{se sexo masculino} \end{cases}$$

Variável explicativa Qualitativa

- O modelo de regressão linear é definido por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i,$$

em que:

- Y_i é a duração da internação em um hospital,
- X_{i1} é a idade e X_{i2} é o sexo do i-ésimo paciente.
- A função de regressão é: $E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$.

Variável explicativa Qualitativa

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

- Para pacientes do sexo masculino, $X_2 = 0$, a função de regressão é: $E(Y_i) = \beta_0 + \beta_1 X_{i1}$
- Para pacientes do sexo feminino, $X_2 = 1$, a função de regressão é: $E(Y_i) = (\beta_0 + \beta_2) + \beta_1 X_{i1}$
- Essas duas funções de regressão representam linhas retas paralelas com interceptos diferentes.

Variável explicativa Qualitativa

Um modelo de regressão linear para a variável resposta Y com idade, sexo e estado do paciente como variáveis explicativas é definido por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i,$$

em que:

- Y_i é a duração da internação em um hospital,
- X_{i1} é a idade
- $X_{i2} = \begin{cases} 1 & \text{se sexo feminino} \\ 0 & \text{se sexo masculino} \end{cases}$
- a variável estado do paciente: bom, estável ou grave, é representada pelas variáveis indicadoras:
- $X_{i3} = \begin{cases} 1 & \text{se grave} \\ 0 & \text{caso contrário} \end{cases}$ $\quad X_{i4} = \begin{cases} 1 & \text{se estável} \\ 0 & \text{caso contrário} \end{cases}$

Variável explicativa Qualitativa

- a variável estado do paciente: bom, estável ou grave, é representada pelas variáveis indicadoras:

Estado do paciente	X_{i3}	X_{i4}
grave	1	0
estável	0	1
bom	0	0

- Em geral, variáveis qualitativas com c categorias/classes são representadas por $c - 1$ variáveis indicadoras.

Modelo de Regressão Polinomial

Modelos de regressão polinomial são um caso particular do modelo de regressão linear geral. Eles consideram termo quadrático ou de maior ordem da variável explicativa. Um modelo de regressão polinomial com uma variável explicativa é definido por:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i,$$

se definirmos:

- $X_{i1} = X_i$ e $X_{i2} = X_i^2$
- O modelo pode ser escrito por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i.$$

Modelo de Regressão com variáveis transformadas

Considere o modelo com a seguinte variável transformada:

$$\log Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i.$$

- Se considerarmos $Y_i' = \log Y_i$, o modelo pode ser escrito por:

$$Y_i' = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i,$$

que está na forma do modelo de regressão linear geral (2).

Modelo de Regressão com variáveis transformadas

Agora, considere o modelo com a seguinte variável transformada:

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

- Se considerarmos $Y_i' = 1/Y_i$, o modelo pode ser escrito por:

$$Y_i' = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i,$$

que está na forma do modelo de regressão linear geral (2).

Modelo de Regressão com Interação

Quando o efeito da variável explicativa na variável resposta não é aditivo e o efeito de uma variável explicativa depende dos níveis de outras variáveis explicativas, tem-se modelos de regressão não aditivos:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i.$$

- Se considerarmos $X_{i3} = X_{i1} X_{i2}$, o modelo pode ser escrito por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i.$$

que está na forma do modelo de regressão linear geral (2).

Combinação de casos

- Um modelo de regressão pode combinar alguns elementos já citados anteriormente e ainda pode ser tratado como um modelo de regressão linear geral.
- Seja um modelo de segunda ordem e com interação:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2} + \beta_4 X_{i2}^2 + \beta_5 X_{i1} X_{i2} + \varepsilon_i. \quad (4)$$

- Seja: $Z_{i1} = X_{i1}$, $Z_{i2} = X_{i1}^2$, $Z_{i3} = X_{i2}$, $Z_{i4} = X_{i2}^2$ e $Z_{i5} = X_{i1} X_{i2}$.
- O modelo de regressão (4) pode ser expresso por:

$$Y_i = \beta_0 + \beta_1 Z_{i1} + \beta_2 Z_{i2} + \beta_3 Z_{i3} + \beta_4 Z_{i4} + \beta_5 Z_{i5} + \varepsilon_i,$$

que está na forma do modelo de regressão linear geral (2).

Modelo de Regressão Linear Geral

Observação: Por meio dos modelos apresentados, verifica-se que o *Modelo de Regressão Linear Geral* não é restrito à superfície de resposta linear. O termo *modelo linear* se refere ao fato que o modelo (2) é linear nos parâmetros.

Modelo de Regressão Linear Geral em termos matriciais

- O modelo de regressão linear geral:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i \qquad i = 1, \ldots, n$$

pode ser escrito por:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \ldots + \beta_p X_{1p} + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \ldots + \beta_p X_{2p} + \varepsilon_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \ldots + \beta_p X_{np} + \varepsilon_n \end{aligned} \qquad (5)$$

Modelo de Regressão Linear Geral em termos matriciais

- E por ser definido pelos vetores das observações, matriz de variáveis explicativas, vetor dos parâmetros e vetor de erros:

$$\mathbf{Y}_{nx1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X}_{nx(p+1)} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix}$$

$$\boldsymbol{\beta}_{(p+1)x1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \boldsymbol{\varepsilon}_{nx1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- Dessa forma, o modelo (2) pode ser escrito em termos matriciais:

$$\mathbf{Y}_{nx1} = \mathbf{X}_{nx(p+1)}\boldsymbol{\beta}_{(p+1)x1} + \boldsymbol{\varepsilon}_{nx1}$$

Modelo de Regressão Linear Geral em termos matriciais

- As condições do erro do modelo (2), $E(\varepsilon_i) = 0$ e $Var(\varepsilon_i) = \sigma^2$, podem ser escritas em termos matriciais por:

$$E(\boldsymbol{\varepsilon})_{nx1} = \mathbf{0}_{nx1} \tag{6}$$

que implica em:

$$\begin{bmatrix} E(\varepsilon_1) \\ E(\varepsilon_2) \\ \vdots \\ E(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Modelo de Regressão Linear Geral em termos matriciais

- A condição que os erros têm variância constante σ^2 e que todas as covariâncias, $Cov(\varepsilon_i, \varepsilon_j) = 0$ para todo $i \neq j$ (e que os erros são normais e independentes) é expressa em termos matriciais através da matriz de covariâncias:

$$Var(\varepsilon)_{nxn} = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

$$Var(\varepsilon)_{nxn} = \sigma^2 \mathbf{I}_{nxn} \tag{7}$$

Modelo de Regressão Linear Geral em termos matriciais

- Assim, o modelo de regressão (2), em termos matriciais é definido por:

$$\mathbf{Y}_{nx1} = \mathbf{X}_{nx(p+1)}\beta_{(p+1)x1} + \varepsilon_{nx1} \tag{8}$$

em que,

- ε é um vetor de variáveis aleatórias independentes e identicamente distribuídas com distribuição Normal e $E(\varepsilon) = \mathbf{0}$ e $Var(\varepsilon) = \sigma^2\mathbf{I}$.
- Consequentemente,

$$E(\mathbf{Y}) = E(\mathbf{X}\beta + \varepsilon) = \mathbf{X}\beta + E(\varepsilon) = \mathbf{X}\beta$$
$$Var(\mathbf{Y}) = Var(\mathbf{X}\beta + \varepsilon) = \mathbf{0} + Var(\varepsilon) = \sigma^2\mathbf{I}$$

Estimação dos parâmtros do modelo em forma matricial

- O método de mínimos quadrados para estimar os parâmetros do modelo de regressão linear geral, cujo objetivo é minimizar a soma do quadrado dos erros:

$$\begin{aligned} SQ(\beta_0, \beta_1, \ldots, \beta_p) &= \sum_{i=1}^{n} \varepsilon_i^2 \\ &= \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}))^2. \end{aligned}$$

- Em notação matricial, tem-se que:

$$SQ(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}' \boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \qquad (9)$$

Estimação dos parâmtros do modelo em forma matricial

- Ao expandir (9), tem-se que:

$$SQ(\beta) = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\beta - \beta'\mathbf{X}'\mathbf{Y} + \beta'\mathbf{X}'\mathbf{X}\beta$$

- $\mathbf{Y}'\mathbf{X}\beta$ é um escalar $(1x1)$, que é igual a sua transposta: $(\mathbf{Y}'\mathbf{X}\beta)' = \beta'\mathbf{X}'\mathbf{Y}$. Dessa forma, escrevemos:

$$SQ(\beta) = \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta \qquad (10)$$

- Para encontrar o vetor β que minimiza (10), faz-se: $\frac{\partial SQ(\beta)}{\partial \beta}$

Estimação dos parâmtros do modelo em forma matricial

$$\frac{\partial SQ(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \tag{11}$$

- Igualando (11) a zero e substituindo $\boldsymbol{\beta}$ por $\hat{\boldsymbol{\beta}}$, encontra-se as equações normais em forma matricial:

$$\begin{aligned} -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= 0 \\ -\mathbf{X}'\mathbf{Y} + \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= 0 \\ \text{equações normais} \Rightarrow \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{X}'\mathbf{Y} \end{aligned}$$

Estimação dos parâmtros do modelo em forma matricial

- Para obter os coeficientes estimados através das equações normais por métodos matriciais, multiplica-se ambos os lados pela inversa de $\mathbf{X}'\mathbf{X}$:

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

- Sabendo que $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$, matriz identidade, tem-se que:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{12}$$

Estimação dos parâmtros do modelo em forma matricial

- A matriz $(\mathbf{X}'\mathbf{X})^{-1}$ sempre existirá se as variáveis regressoras forem **linearmente independentes**.

- As variáveis regressoras são linearmente independente, se nenhuma coluna da matriz $\mathbf{X}$ for uma combinação linear das outras colunas.

Estimação dos parâmtros do modelo em forma matricial

Observação:

- Assume-se que as variáveis explicativas são fixas (não aleatórias), medidas sem erro.
- Entretanto, quando os dados são de estudos observacionais, algumas variáveis podem ser aleatórias. Porém, todos os resultados encontrados quando assume-se que as variáveis explicativas são fixas, são válidos para o caso em que essas variáveis são aleatórias.
- Quando as variáveis $X_1, X_2, \ldots, X_p$ são aleatórias é necessário que as observações de cada variável sejam independentes e que a distribuição não dependa dos coeficientes β ou de σ^2.
- Para construir IC e testes de hipóteses, tem que assumir que a distribuição condicional de Y dado $X_1, X_2, \ldots, X_p$ é normal com média $\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ e variância σ^2.

Valores ajustados

- Seja o vetor de valores ajustados $\hat{Y}_i$ denotado por $\hat{\mathbf{Y}}$:

$$\hat{\mathbf{Y}}_{nx1} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}$$

- Em notação matricial tem-se que:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} \tag{13}$$

porque:

$$\mathbf{Y}_{nx1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

Valores ajustados

- A equação (13) pode ser escrita como:

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

ou, de forma equivalente:

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} \tag{14}$$

em que:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

é conhecida como a matriz chapéu. Essa matriz tem importante papel na análise de diagnóstico. A matriz $\mathbf{H}$ é simétrica e tem a propriedade de ser idempotência. Ou seja,

$$\mathbf{H}\mathbf{H} = \mathbf{H}$$

Resíduo

- Seja o vetor de resíduos $e_i = Y_i - \hat{Y}_i$ denotado por $\mathbf{e}$:

$$\mathbf{e}_{nx1} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

- Em notação matricial tem-se que:

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \tag{15}$$

ou,

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

Estimador de σ^2

- Como definido anteriormente, o estimador de σ^2 é dado por:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-(p+1)} = \frac{SQRes}{n-(p+1)} = MSRes$$

- Em termos matriciais $SQRes = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ pode ser definido por:

$$\begin{aligned} SQRes &= \mathbf{e}'\mathbf{e} \\ &= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= (\mathbf{Y}' - \hat{\boldsymbol{\beta}}'\mathbf{X}')(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{Y}'\mathbf{Y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \end{aligned}$$

Estimador de σ^2

- Desde que $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, tem-se que:

$$\begin{aligned} SQRes &= \mathbf{Y}'\mathbf{Y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{Y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} \end{aligned}$$

- Dessa forma, o estimador de σ^2 em termos matriciais é definido por:

$$\hat{\sigma}^2 = \frac{\mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}}{n-(p+1)} = MSRes$$

Modelo de regressão linear geral

Em geral, a variável resposta Y pode estar relacionada com p variáveis explicativas. O modelo de regressão é definido por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i, \qquad (2)$$

é conhecido como modelo de primeira ordem, pois é linear nos parâmetros e linear nas variáveis explicativas. Esse modelo é conhecido como um modelo de regressão linear múltipla ou também como um modelo de regressão linear geral.

- Y_i é o valor da variável resposta para a i-ésima observação,
- β_0, $\beta_1, \ldots, \beta_p$ são parâmetros desconhecidos,
- $X_{i1}, X_{i2}, \ldots, X_{ip}$ são constantes conhecidas,
- ε_i são independentes e $N(0, \sigma^2)$, $i = 1, \ldots, n$.

Modelo de Regressão Linear Geral em termos matriciais

- Dessa forma, o modelo (2) pode ser escrito em termos matriciais:

$$\mathbf{Y}_{nx1} = \mathbf{X}_{nx(p+1)}\boldsymbol{\beta}_{(p+1)x1} + \boldsymbol{\varepsilon}_{nx1}$$

- As condições do erro do modelo (2) são:

$$E(\boldsymbol{\varepsilon})_{nx1} = \mathbf{0}_{nx1} \quad \text{e} \quad Var(\boldsymbol{\varepsilon})_{nxn} = \sigma^2\mathbf{I}_{nxn}$$

- Consequentemente,

$$E(\mathbf{Y}) = E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta} + E(\boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta}$$
$$Var(\mathbf{Y}) = Var(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{0} + Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

- O estimador de mínimos quadrados de $\boldsymbol{\beta}$ é dada por:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad (3)$$

Estimador de σ^2

- Como definido anteriormente, o estimador de σ^2 é dado por:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-(p+1)} = \frac{SQRes}{n-(p+1)} = MSRes$$

- Em termos matriciais:

$$\hat{\sigma}^2 = \frac{\mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y}}{n-(p+1)} = \frac{SQRes}{n-(p+1)} = MSRes$$

Soma de Quadrados

- Vamos definir a soma de quadrados total em termos matriciais. Lembrando que:

$$SQT = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}$$

- O termo $\frac{(\sum_{i=1}^{n} Y_i)^2}{n}$ pode ser escrito como:

$$\frac{(\sum_{i=1}^{n} Y_i)^2}{n} = (\frac{1}{n})\mathbf{Y}'\mathbf{J}\mathbf{Y}$$

em que $\mathbf{J}$ é uma matriz de 1s. Isto é:

$$\mathbf{J} = \begin{bmatrix} 1 & \dots & 1 \\ \vdots & & \vdots \\ 1 & \dots & 1 \end{bmatrix}$$

Soma de Quadrados

- Dessa forma, a soma de quadrados total é definida por:

$$SQT = \mathbf{Y}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y}.$$

- O número de graus de liberdade associado a SQT é $n - 1$.

- Dado as informações das SQT e SQRes, a soma de quadrados da regressão é definida por:

$$SQReg = SQT - SQRes = \hat{\beta}'\mathbf{X}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y}.$$

- O número de graus de liberdade associado a SQReg é p.

Análise de Variância

- Hipóteses:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$
$$H_1 : \text{existe pelo menos um} \beta_k \neq 0$$

ou,

$$H_0 : \text{ausência de regressão}$$
$$H_1 : \text{existe regressão}$$

Análise de Variância

Tabela : Tabela de Análise de Variância

Fonte de Variação	SQ	g.l.	QM	F
Regressão	SQReg	p	MSReg	MSReg/MSRes
Resíduo	SQRes	n-(p+1)	MSRes	
Total	SQT	n-1		

Análise de Variância

- Estatística do teste

$$F = \frac{MSReg}{MSRes},$$

tem distribuição F com p graus de liberdade no numerador e (n-(p+1)) graus de liberdade no denominador.

- Se $F \leq F_{(1-\alpha,p,n-(p+1))} \Rightarrow$ Não há evidência para rejeitar H_0
- Se $F > F_{(1-\alpha,p,n-(p+1))} \Rightarrow$ Há evidência para rejeitar H_0

Coeficiente de Determinação

- A medida R^2, chamada de **Coeficiente de Determinação**, representa a proporção da variação total explicada pela relação **X** e **Y** (regressão).

$$R^2 = \frac{SQReg}{SQT} = 1 - \frac{SQRes}{SQT}$$

- Desde de que $0 \leq SQRes \leq SQT$, segue que

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

- Valores grandes de R^2 indicam que a variação total de **Y** é reduzida pela introdução das variáveis explicativas $X_1, X_2, \ldots, X_p$.

Coeficiente de Determinação Ajustado

- Ao adicionar mais variáveis explicativas no modelo, a medida R^2 sempre aumentará. Porque, SQReg aumenta ao adicionar mais variáveis explicativas e a SQT é sempre a mesma para o conjunto de variável resposta.

- O Coeficiente de Determinação Ajustado, R^2_{ajust} é definido por:

$$R^2_{\text{ajust}} = 1 - \frac{SQRes/(n-(p+1))}{SQT/(n-1)},$$

- R^2_{ajust} apenas aumentará quando for adicionado uma variável que reduzirá o MSRes.

Propriedades dos estimadores de Mínimos Quadrados

- O Teorema de Gauss Markov, afirma que os estimadores de mínimos quadrados de β_0, $\beta_1, \ldots, \beta_p$ são não viesados. Em termos matriciais o Teorema é definido por:
- Lembrando que $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{C}\mathbf{Y}$
- Sendo que $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ é uma matriz de constante.
- Dessa forma, tem-se que:

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

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

Propriedades dos estimadores de Mínimos Quadrados

- A matriz de covariâncias de $\hat{\beta}$ é definida por:

$$Var(\hat{\beta}) = Var(\mathbf{CY}) = \mathbf{C}Var(\mathbf{Y})\mathbf{C}'$$

como $Var(\mathbf{Y}) = \sigma^2\mathbf{I}$

$$\begin{aligned} Var(\hat{\beta}) &= \mathbf{C}\sigma^2\mathbf{I}\mathbf{C}' \\ &= \sigma^2\mathbf{C}\mathbf{C}' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

Propriedades dos estimadores de Mínimos Quadrados

- A matriz de covariâncias estimada é obtida ao substituir σ^2 por seu estimador, $\hat{\sigma}^2$. Ou seja,

$$\hat{Var}(\hat{\beta}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} = MSRes(\mathbf{X}'\mathbf{X})^{-1}$$

Distribuição amostral

- Ao considerar as hipóteses do modelo (2):
  - ε são independentes e identicamente distribuídos com distribuição Normal,
  - $E(\varepsilon) = \mathbf{0}$ e $Var(\varepsilon) = \sigma^2\mathbf{I}$,
  - Portanto, $\mathbf{Y}$ são independentes e identicamente distribuídos com distribuição Normal com média $E(\mathbf{Y}) = \mathbf{X}\beta$ e variância $Var(\mathbf{Y}) = \sigma^2\mathbf{I}$.
- Desde que, os estimadores $\hat{\beta}$ são uma combinação linear das observações, $\mathbf{Y}$, segue que:

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

Intervalo de Confiança para β_k

- Para o modelo de regressão linear (2), tem-se a estatística:

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}^2 C_{kk}}} \sim \text{t-Student}(n - (p+1)),$$

em que $k = 0, 1, \ldots, p$, $Var(\beta_k) = \sigma^2 C_{kk}$ e C_{kk} é o k-ésimo elemento da diagonal da matriz $(\mathbf{X}'\mathbf{X})^{-1}$.

Intervalo de Confiança para β_k

- Um Intervalo de Confiança $(1-\alpha)$ para o parâmetro β_k do modelo de regressão linear (2) é dado por:

$$IC(\beta_k, (1-\alpha)) = (\hat{\beta}_k - t_{(1-\alpha/2;n-(p+1))}\sqrt{\hat{\sigma}^2 C_{kk}}; \hat{\beta}_k + t_{(1-\alpha/2;n-(p+1))}\sqrt{\hat{\sigma}^2 C_{kk}})$$

Teste de hipótese para β_k

- Para testar a significância de algum coeficiente de regressão, β_k, as hipóteses são:

$$H_0 : \beta_k = 0$$
$$H_1 : \beta_k \neq 0$$

- Supondo H_0 verdadeira, a estatística do teste é dada por:

$$t = \frac{\hat{\beta}_k - 0}{\sqrt{\hat{\sigma}^2 C_{kk}}} \sim \text{t-Student}(n - (p + 1)).$$

- A hipótese nula será rejeitada se $|t| > \text{t-Student}(n - (p + 1))$.

Intervalo de Confiança para a resposta média

- Para alguma valor de $X_1, \ldots, X_p$ denotado por $X_{01}, \ldots, X_{0p}$, a resposta média é $E(Y_0)$. O vetor $\mathbf{X}_0$ é definido por:

$$\mathbf{X}_0 = \begin{bmatrix} 1 \\ X_{01} \\ \vdots \\ X_{0p} \end{bmatrix}$$

- O valor ajustado é definido por:

$$\hat{Y}_0 = \mathbf{X}_0' \hat{\boldsymbol{\beta}}$$

- O estimador $\hat{Y}_0$ é não viesado:

$$E(\hat{Y}_0) = \mathbf{X}_0' \boldsymbol{\beta} = E(Y_0)$$

Intervalo de Confiança para a resposta média

- A variância de $\hat{Y}_0$ é dada por:

$$Var(\hat{Y}_0) = \sigma^2 \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0$$

- Um intervalo de confiança $(1-\alpha)$ para $E(Y_0)$ é dado por:

$$(\hat{Y}_0 - t_{(1-\alpha/2;n-(p+1))}\sqrt{\hat{\sigma^2}\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0};\ \hat{Y}_0 + t_{(1-\alpha/2;n-(p+1))}\sqrt{\hat{\sigma^2}\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0})$$