Teste de Hipótese em Regressão Normal Linear Múltipla

Análise do Teste $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}_p$

Curso de Inferência Estatística - PPGEST/UFPE

17 de novembro de 2025

1 Introdução e Especificação do Modelo

A regressão linear múltipla constitui uma das ferramentas fundamentais da análise estatística, permitindo modelar a relação entre uma variável resposta Y e múltiplas variáveis explicativas $X_1, X_2, \ldots, X_p$. Os testes de hipótese desempenham papel crucial nesta análise, permitindo avaliar a significância estatística dos parâmetros do modelo e, consequentemente, a relevância das variáveis explicativas na predição da variável resposta.

1.1 Especificação do Modelo

O modelo de regressão linear múltipla é especificado como:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{1}$$

onde:

- Y_i é a variável aleatória resposta para a observação i;
- $X_{1i}, X_{2i}, \ldots, X_{pi}$ são variáveis explicativas fixas (não aleatórias) para a observação i;
- β_0 é o intercepto (parâmetro desconhecido);
- $\beta_1, \beta_2, \ldots, \beta_p$ são os coeficientes de regressão (parâmetros desconhecidos);
- ε_i é o termo de erro aleatório para a observação i.

Em notação matricial, o modelo pode ser escrito de forma compacta como:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2}$$

onde:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{21} & \cdots & X_{p1} \\ 1 & X_{12} & X_{22} & \cdots & X_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{pn} \end{pmatrix}, \tag{3}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{4}$$

A matriz $\mathbf{X}$ é denominada matriz de planejamento (design matrix) e possui dimensão $n \times (p+1)$, onde a primeira coluna é composta por uns para incluir o intercepto.

1.2 Pressupostos do Modelo

Para que os resultados inferenciais sejam válidos, o modelo de regressão linear múltipla requer os seguintes pressupostos:

Definição 1.1 (Pressupostos Clássicos)**.**

1. **Normalidade**: Os erros ε_i são independentes e identicamente distribuídos como $\varepsilon_i \sim N(0, \sigma^2)$ para $i = 1, 2, \ldots, n$, onde $\sigma^2 > 0$ é desconhecido.

2. **Homocedasticidade**: A variância dos erros é constante, ou seja, $\text{Var}(\varepsilon_i) = \sigma^2$ para todo i.

3. **Independência**: Os erros são mutuamente independentes, isto é, $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ para $i \neq j$.

4. **Não-colinearidade**: A matriz $\mathbf{X}^T\mathbf{X}$ é não-singular (invertível), garantindo que não há dependência linear perfeita entre as variáveis explicativas.

5. **Linearidade**: A relação entre $E[Y_i]$ e as variáveis explicativas é linear nos parâmetros.

Sob esses pressupostos, temos que $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, onde $\mathbf{I}_n$ é a matriz identidade de ordem n.

2 Fundamentação Teórica

2.1 Estimadores de Mínimos Quadrados

O método de mínimos quadrados consiste em encontrar os valores de $\boldsymbol{\beta}$ que minimizam a soma dos quadrados dos resíduos:

$$S(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \tag{5}$$

Diferenciando $S(\boldsymbol{\beta})$ em relação a $\boldsymbol{\beta}$ e igualando a zero, obtemos as equações normais:

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{Y} \tag{6}$$

Assumindo que $\mathbf{X}^T\mathbf{X}$ é invertível (pressuposto de não-colinearidade), o estimador de mínimos quadrados é dado por:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{7}$$

Teorema 2.1 (Propriedades do Estimador de Mínimos Quadrados)**.** Sob os pressupostos clássicos do modelo de regressão linear múltipla:

1. **Não-viesado**: $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$.
2. **Eficiência (Teorema de Gauss-Markov)**: Entre todos os estimadores lineares não-viesados de $\boldsymbol{\beta}$, o estimador de mínimos quadrados possui variância mínima.
3. **Consistência**: $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$ quando $n \to \infty$.

2.2 Distribuição dos Estimadores

Como $\hat{\boldsymbol{\beta}}$ é uma combinação linear de variáveis aleatórias normais, segue que:

Proposição 2.1 (Distribuição do Estimador)**.** O estimador de mínimos quadrados $\hat{\boldsymbol{\beta}}$ segue uma distribuição normal multivariada:

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right) \tag{8}$$

Para fins de teste de hipótese, é conveniente particionar o vetor de parâmetros como:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_1 \end{pmatrix}, \quad \text{onde} \quad \boldsymbol{\beta}_1 = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \tag{9}$$

Analogamente, particionamos a matriz $\mathbf{X}$ como $\mathbf{X} = [\mathbf{1}_n \,|\, \mathbf{X}_1]$, onde $\mathbf{1}_n$ é o vetor coluna de n uns e $\mathbf{X}_1$ é a matriz $n \times p$ contendo as variáveis explicativas.

A distribuição marginal de $\hat{\boldsymbol{\beta}}_1$ é:

$$\hat{\boldsymbol{\beta}}_1 \sim N\left(\boldsymbol{\beta}_1, \sigma^2(\mathbf{X}_1^T\mathbf{M}_1\mathbf{X}_1)^{-1}\right) \tag{10}$$

onde $\mathbf{M}_1 = \mathbf{I}_n - \mathbf{1}_n(\mathbf{1}_n^T\mathbf{1}_n)^{-1}\mathbf{1}_n^T$ é a matriz de projeção ortogonal.

2.3 Estimador da Variância

O estimador não-viesado de σ^2 é dado por:

$$\hat{\sigma}^2 = \frac{SSE}{n-p-1} = \frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{n-p-1} \tag{11}$$

onde SSE denota a Soma dos Quadrados dos Erros (Sum of Squares Error):

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2 \tag{12}$$

e $\hat{Y}_i = \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$ são os valores preditos, com $\mathbf{x}_i^T$ sendo a i-ésima linha de $\mathbf{X}$.

Proposição 2.2 (Distribuição do Estimador da Variância)**.** Sob os pressupostos do modelo, temos:

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1} \tag{13}$$

e $\hat{\sigma}^2$ é independente de $\hat{\boldsymbol{\beta}}$.

3 Teste de Hipótese

3.1 Hipótese Nula

O teste de interesse consiste em verificar se as variáveis explicativas $X_1, X_2, \ldots, X_p$ têm efeito significativo sobre a variável resposta Y, após controlar pelo intercepto. Formalmente, testamos:

$$H_0 : \boldsymbol{\beta}_1 = \mathbf{0}_p \quad \text{versus} \quad H_1 : \boldsymbol{\beta}_1 \neq \mathbf{0}_p \tag{14}$$

onde $\mathbf{0}_p$ é o vetor nulo de dimensão p. Sob H_0, o modelo reduz-se a $Y_i = \beta_0 + \varepsilon_i$, indicando que nenhuma das variáveis explicativas contribui significativamente para explicar a variabilidade de Y.

3.2 Estatística F

Para testar a hipótese $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}_p$, utilizamos a estatística F definida como a razão entre a variância explicada pelo modelo completo e a variância residual:

$$F = \frac{MSR}{MSE} = \frac{SSR/p}{SSE/(n-p-1)} \tag{15}$$

onde:

- $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ é a Soma dos Quadrados da Regressão (Sum of Squares Regression);
- $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ é a Soma dos Quadrados dos Erros;
- $MSR = SSR/p$ é o Quadrado Médio da Regressão (Mean Square Regression);
- $MSE = SSE/(n-p-1)$ é o Quadrado Médio dos Erros (Mean Square Error);
- $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ é a média amostral da variável resposta.

Teorema 3.1 (Distribuição da Estatística F)**.** Sob a hipótese nula $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}_p$ e os pressupostos do modelo, a estatística F segue uma distribuição F de Fisher-Snedecor:

$$F \sim F_{p,n-p-1} \tag{16}$$

A interpretação da estatística F é intuitiva: valores grandes de F indicam que a variância explicada pelo modelo (MSR) é substancialmente maior que a variância residual (MSE), sugerindo que pelo menos uma das variáveis explicativas tem efeito significativo sobre Y.

3.3 Estatística t para Componentes Individuais

Além do teste global $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}_p$, podemos testar hipóteses sobre componentes individuais de $\boldsymbol{\beta}_1$. Para testar $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$, utilizamos a estatística t:

$$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}} \tag{17}$$

onde $SE(\hat{\beta}_j)$ é o erro padrão de $\hat{\beta}_j$ e $[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}$ denota o j-ésimo elemento da diagonal da matriz $(\mathbf{X}^T\mathbf{X})^{-1}$.

Proposição 3.1 (Distribuição da Estatística t)**.** Sob $H_0 : \beta_j = 0$ e os pressupostos do modelo:

$$t_j \sim t_{n-p-1} \tag{18}$$

3.4 Região de Rejeição

Para um nível de significância α fixado (tipicamente $\alpha = 0.05$ ou $\alpha = 0.01$), a região de rejeição para o teste F é:

$$\text{Rejeita } H_0 \text{ se } F > F_{p,n-p-1;\alpha} \tag{19}$$

onde $F_{p,n-p-1;\alpha}$ é o quantil $(1-\alpha)$ da distribuição F com p e $n-p-1$ graus de liberdade.

Alternativamente, podemos utilizar o p-valor:

$$p\text{-valor} = P(F_{p,n-p-1} > F_{\text{obs}}) \tag{20}$$

onde F_{obs} é o valor observado da estatística F. Rejeitamos H_0 se p-valor $< \alpha$.

Para os testes individuais usando a estatística t, a região de rejeição é:

$$\text{Rejeita } H_0 : \beta_j = 0 \text{ se } |t_j| > t_{n-p-1;\alpha/2} \tag{21}$$

4 Aplicações e Considerações Finais

4.1 Exemplo Numérico

Considere um modelo de regressão linear múltipla com $n = 30$ observações e $p = 3$ variáveis explicativas. Suponha que após ajustar o modelo, obtemos:

$$SSR = 450.2, \quad SSE = 89.5, \tag{22}$$

$$MSR = \frac{450.2}{3} = 150.07, \quad MSE = \frac{89.5}{26} = 3.44 \tag{23}$$

A estatística F observada é:

$$F_{\text{obs}} = \frac{150.07}{3.44} = 43.61 \tag{24}$$

Comparando com $F_{3,26;0.05} \approx 2.98$, temos $F_{\text{obs}} > F_{3,26;0.05}$, portanto rejeitamos $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}_3$ ao nível de significância de 5%. Concluímos que pelo menos uma das três variáveis explicativas tem efeito significativo sobre a variável resposta.

4.2 Relação com Análise de Variância (ANOVA)

O teste F está intimamente relacionado à Análise de Variância. A decomposição fundamental da variabilidade total é:

$$SST = SSR + SSE \tag{25}$$

onde $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ é a Soma Total dos Quadrados (Total Sum of Squares), representando a variabilidade total da variável resposta.

A tabela ANOVA resume essas informações:

Fonte	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	F
Regressão	SSR	p	$MSR = SSR/p$	$F = MSR/MS$
Erro	SSE	$n - p - 1$	$MSE = SSE/(n - p - 1)$	
Total	SST	$n - 1$		

Tabela 1: Tabela ANOVA para Regressão Linear Múltipla

O coeficiente de determinação múltiplo, $R^2 = SSR/SST$, mede a proporção da variabilidade total explicada pelo modelo. Valores próximos de 1 indicam bom ajuste do modelo aos dados.

4.3 Considerações Finais

O teste de hipótese $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}_p$ desempenha papel fundamental na análise de regressão, permitindo:

- **Seleção de variáveis**: Identificar quais variáveis explicativas são estatisticamente significativas para predizer a variável resposta.
- **Validação do modelo**: Verificar se o modelo proposto captura relações significativas entre as variáveis.

- **Comparação de modelos**: Avaliar se a inclusão de variáveis adicionais melhora significativamente o ajuste do modelo.

No entanto, é importante observar algumas limitações e cuidados:

1. O teste F global não identifica quais variáveis específicas são significativas; para isso, são necessários os testes t individuais.
2. A violação dos pressupostos (normalidade, homocedasticidade, independência) pode comprometer a validade dos testes.
3. A significância estatística não implica necessariamente significância prática; é importante considerar também o tamanho do efeito.
4. Em modelos com muitas variáveis, problemas de multicolinearidade podem afetar a precisão dos estimadores e a interpretação dos testes.

Em resumo, o teste de hipótese em regressão linear múltipla fornece ferramentas poderosas para inferência estatística, mas requer cuidadosa verificação de pressupostos e interpretação contextual dos resultados.

Referências Bibliográficas

- Casella, G. & Berger, R. L. (2002). *Statistical Inference*. 2nd ed. Duxbury Press.
- Montgomery, D. C., Peck, E. A. & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. 5th ed. Wiley.
- Seber, G. A. F. & Lee, A. J. (2012). *Linear Regression Analysis*. 2nd ed. Wiley.