Estimação pontual da média da variável resposta

Dados os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ dos parâmetros da função de regressão:

$$E(Y) = \beta_0 + \beta_1 X$$

A função de regressão estimada é dada por:

$$\hat{Y} = E(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 X,$$

em que $\hat{Y}$ é o valor estimado da média da distribuição de Y, dado a informação da variável explicativa X.

Estimação pontual da média da variável resposta

Pode ser mostrado como uma extensão do **Teorema de Gauss Markov** que:

- $\hat{Y}$ é um estimador não viesado de $E(Y)$,
- com variância mínima na classe de estimadores lineares não viesados.

Para as observações da amostra, o valor estimado para i-ésima observação é dado por:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,$$

em que $i = 1, \ldots, n$.

Resíduo do Modelo

- O i-ésimo **resíduo** é a diferença entre o valor observado Y_i e seu correspondente valor estimado $\hat{Y}_i$.
- O resíduo é denotado por e_i e é definido por:

$$e_i = Y_i - \hat{Y}_i.$$

  O resíduo é representado por:

$$e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X) = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X.$$

- Os resíduos são **importantes** para estudar se o modelo de regressão é apropriado para analisar os dados em estudo.

Atenção!!!

É importante distinguir entre:

- Erro do modelo:

$$\varepsilon_i = Y_i - E(Y_i),$$

e

- Resíduo:

$$e_i = Y_i - \hat{Y}_i.$$

Propriedades da Reta de Regressão estimada

1. A soma dos resíduos do modelo de regressão é **ZERO**:

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = \sum_{i=1}^{n} e_i = 0$$

2. A soma do quadrado do resíduo, $\sum_{i=1}^{n} e_i^2$, é um mínimo.
   $\Longrightarrow$ Pressuposto a ser satisifeito na dedução dos estimadores de mínimos quadrados.

Propriedades da Reta de Regressão estimada

3. A soma dos valores observados Y_i, é igual a soma dos valores estimados $\hat{Y}_i$:

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$$

4. A soma ponderada dos resíduos é **ZERO** quando o resíduo da i-ésima observação é ponderado pelo nível da variável explicativa da i-ésima observação:

$$\sum_{i=1}^{n} X_i e_i = 0$$

Propriedades da Reta de Regressão estimada

5. A soma ponderada dos resíduos é **ZERO** quando o resíduo da i-ésima observação é ponderado pelo valor estimado da variável resposta da i-ésima observação:

$$\sum_{i=1}^{n} \hat{Y}_i e_i = 0$$

6. A reta de regressão sempre passa através do ponto $(\bar{X}, \bar{Y})$.

Estimação da Variância do Erro

A variância σ^2 do erro ε_i do modelo de regressão precisa ser estimada por alguns motivos:

- $\text{Var}(Y_i) = \text{Var}(\varepsilon_i) = \sigma^2$ - estimar σ^2 para obter uma indicação da variabilidade da distribuição de probabilidade de Y,
- Para construir Intervalos de Confiança e Testes de Hipóteses dos parâmetros β_0 e β_1,
- Para fazer predição de Y.

Revisando

Na amostragem de uma população, a variância σ^2, é estimada pela variância da amostra s^2:

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

em que:

- $Y_i - \bar{Y}$ - são os desvios de Y_i em torno da média estimada,
- $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ - Soma de quadrados,
- $n - 1$ graus de liberdade.

Estimação da Variância do Erro

Ao considerar o modelo de regressão linear simples, o desvio de uma observação Y_i deve ser calculado em torno da sua própria média estimada $\mathrm{E}(\hat{Y}_i) = \hat{Y}_i$. Esses desvios são os resíduos:

$$e_i = Y_i - \hat{Y}_i.$$

Dessa forma, o estimador da variância σ^2 do erro ε_i é:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\mathrm{SQRes}}{n-2} = \mathrm{MSRes}, \qquad (1)$$

- $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ - Soma de quadrados do resíduo - SQRes,
- $n-2$ graus de liberdade,
- MSRes - quadrado médio do resíduo
- é um estimador não viesado de σ^2 para o modelo de regressão.

Estimação da Variância do Erro

Consequentemente, um estimador para σ é:

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\text{MSRes}}.$$

Variância de $\hat{\beta}_1$ e $\hat{\beta}_0$

Pode-se encontrar a variância de $\hat{\beta}_1$ e $\hat{\beta}_0$, respectivamente, por:

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sigma^2}{SQ_{XX}},$$

e

$$Var(\hat{\beta}_0) = \sigma^2.\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] = \sigma^2.\left[\frac{1}{n} + \frac{\bar{X}^2}{SQ_{XX}}\right].$$

Estimador da Variância dos parâmetros

Ao considerar o estimador da variância σ^2 definido em (1), temos:

- Estimador da variância de $\hat{\beta}_1$:

$$\widehat{\text{Var}}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\hat{\sigma}^2}{SQ_{XX}},$$

- Estimador da variância de $\hat{\beta}_0$:

$$\widehat{\text{Var}}(\hat{\beta}_0) = \hat{\sigma}^2 . \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] = \hat{\sigma}^2 . \left[\frac{1}{n} + \frac{\bar{X}^2}{SQ_{XX}}\right].$$

Distribuição amostral

Ao considerar as hipóteses do modelo:

- E(ε_i)=0
- Var(ε_i)= σ^2
- Cov($\varepsilon_i, \varepsilon_j$) para todo $i, j; i \neq j$
- ε_i são independentes e identicamente distribuidos: N(0,σ^2)

A distribuição de probabilidade de $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ é: N($\beta_0 + \beta_1 X_i, \sigma^2$).

Distribuição amostral de β_1

Como $\hat{\beta}_1$ é uma combinação linear das observações Y_i (variáveis normais independentes):

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right).$$

A estatística padronizada:

$$\frac{(\hat{\beta}_1 - \beta_1)}{\sqrt{\mathrm{Var}(\hat{\beta}_1)}} = \frac{(\hat{\beta}_1 - \beta_1)}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}},$$

tem distribuição Normal padrão, N(0,1).

Distribuição amostral de β_1

Como é preciso estimar $\text{Var}(\hat{\beta}_1)$ por $\widehat{\text{Var}}(\hat{\beta}_1)$, a estatística:

$$\frac{(\hat{\beta}_1 - \beta_1)}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_1)}} = \frac{(\hat{\beta}_1 - \beta_1)}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}},$$

tem distribuição t-Student(n-2) para o modelo definido.

Intervalo de Confiança para β_1

Ao considerar o resultado do slide anterior, escrevemos a seguinte probabilidade:

$$P\left(t_{(\alpha/2;n-2)} \leq \frac{(\hat{\beta}_1 - \beta_1)}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_1)}} \leq t_{(1-\alpha/2;n-2)} \right) = 1 - \alpha$$

Devido a simetria da distribuição t-Student, temos que:

$$t_{(\alpha/2,n-2)} = -t_{(1-\alpha/2;n-2)},$$

logo:

$$P\left(-\, t_{(1-\alpha/2;n-2)} \leq \frac{(\hat{\beta}_1 - \beta_1)}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_1)}} \leq t_{(1-\alpha/2;n-2)} \right) = 1 - \alpha$$

Intervalo de Confiança para β_1

Dessa forma, um Intervalo de Confiança $(1-\alpha)$ para o parâmetro β_1 do modelo de regressão linear simples é dado por:

$$IC(\beta_1,(1-\alpha)) = \left(\hat{\beta}_1 - t_{(1-\alpha/2;n-2)}\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_1)}; \hat{\beta}_1 + t_{(1-\alpha/2;n-2)}\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_1)}\right),$$

ao substituir o intervalo pelo estimador da variância de $\hat{\beta}_1$, o intervalo é definido por:

$$IC(\beta_1,(1-\alpha)) = \left(\hat{\beta}_1 - t_{(1-\alpha/2;n-2)}\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}; \hat{\beta}_1 + t_{(1-\alpha/2;n-2)}\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}\right)$$

Teste de hipótese para β_1

- Suponha que desejamos testar a hipótese de que o coeficiente angular é igual a uma constante, β_{10}. As hipóteses são:

$$H_0 : \beta_1 = \beta_{10}$$
$$H_1 : \beta_1 \neq \beta_{10}$$

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

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\hat{\text{Var}}(\hat{\beta}_1)}},$$

e tem distribuição t-Student com (n-2) graus de liberdade.

Teste de hipótese para β_1

- Um particular teste de hipótese para o coeficiente angular:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Ou seja,

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

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

$$t = \frac{\hat{\beta}_1 - 0}{\sqrt{\hat{\text{Var}}(\hat{\beta}_1)}},$$

e tem distribuição t-Student com (n-2) graus de liberdade.

Teste de hipótese para β_1

- Fixando o nível de siginificância, α, define-se a região crítica ou de rejeição do teste,
- Seja t o valor da estatística do teste para a amostra obtida:
  - Se $t \in (-t_{1-\alpha/2}, t_{1-\alpha/2})$, então não há evidências para rejeitar H_0,
  - Se t não pertence $(-t_{1-\alpha/2}, t_{1-\alpha/2})$, então há evidências para rejeitar H_0

Distribuição amostral de β_0

Como $\hat{\beta}_0$ é uma combinação linear das observações Y_i (variáveis normais independentes):

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2.\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]\right).$$

A estatística padronizada:

$$\frac{(\hat{\beta}_0 - \beta_0)}{\sqrt{\text{Var}(\hat{\beta}_0)}} = \frac{(\hat{\beta}_0 - \beta_0)}{\sqrt{\sigma^2.\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]}},$$

tem distribuição Normal padrão, N(0,1).

Distribuição amostral de β_0

Como é preciso estimar $\text{Var}(\hat{\beta}_0)$ por $\widehat{\text{Var}}(\hat{\beta}_0)$, a estatística:

$$\frac{(\hat{\beta}_0 - \beta_0)}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_0)}} = \frac{(\hat{\beta}_0 - \beta_0)}{\sqrt{\hat{\sigma}^2 . \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]}},$$

tem distribuição t-Student(n-2) para o modelo definido.

Intervalo de Confiança para β_0

Ao considerar o resultado do slide anterior, escrevemos a seguinte probabilidade:

$$P\left(t_{(\alpha/2;n-2)} \leq \frac{(\hat{\beta}_0 - \beta_0)}{\sqrt{\hat{\text{Var}}(\hat{\beta}_0)}} \leq t_{(1-\alpha/2;n-2)}\right) = 1 - \alpha$$

Devido a simetria da distribuição t-Student, temos que:

$$t_{(\alpha/2,n-2)} = -t_{(1-\alpha/2;n-2)},$$

logo:

$$P\left(-t_{(1-\alpha/2;n-2)} \leq \frac{(\hat{\beta}_0 - \beta_0)}{\sqrt{\hat{\text{Var}}(\hat{\beta}_0)}} \leq t_{(1-\alpha/2;n-2)}\right) = 1 - \alpha$$

Intervalo de Confiança para β_0

Dessa forma, um Intervalo de Confiança $(1-\alpha)$ para o parâmetro β_0 do modelo de regressão linear simples é dado por:

$$IC(\beta_0,(1-\alpha)) = \left(\hat{\beta}_0 - t_{(1-\alpha/2;n-2)}\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_0)}; \hat{\beta}_0 + t_{(1-\alpha/2;n-2)}\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_0)}\right),$$

ao substituir o intervalo pelo estimador da variância de $\hat{\beta}_0$, o intervalo é definido por:

$$IC(\beta_0,(1-\alpha)) = \left(\hat{\beta}_0 - t_{(1-\alpha/2;n-2)}\sqrt{\hat{\sigma}^2.\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right]}; \hat{\beta}_0 + t_{(1-\alpha/2;n-2)}\sqrt{\hat{\sigma}^2.\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i-\bar{X})}\right.}\right.$$

Teste de hipótese para β_0

- Suponha que desejamos testar a hipótese de que o intercepto é igual a uma constante, β_{00}. As hipóteses são:

$$H_0 : \beta_0 = \beta_{00}$$
$$H_1 : \beta_0 \neq \beta_{00}$$

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

$$t = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{\hat{\text{Var}}(\hat{\beta}_0)}},$$

e tem distribuição t-Student com (n-2) graus de liberdade.

Distribuição amostral de $\hat{Y}_0 = \mathrm{E}(\hat{Y}_0)$

- Para cada X_0, nível da variável explicativa, X, desejamos estimar a resposta média $E(Y_0) = E(Y_0|X = X_0) = \beta_0 + \beta_1 X_0$,
- X_0 pode ser:
  - um particular valor que ocorreu na amostra;
  - qualquer valor da variável explicativa, X, que esteja dentro do alcance do modelo.
- Um estimador pontual do valor esperado de Y já foi definido anteriormente por:

$$\hat{Y} = E(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 X$$

- Ao considerar um particular valor da variável explicativa, X_0, o estimador do valor esperado de Y_0 é:

$$\hat{Y}_0 = E(\hat{Y}_0) = \hat{\beta}_0 + \hat{\beta}_1 X_0 \tag{2}$$

Distribuição amostral de $\hat{Y}_0 = \mathrm{E}(\hat{Y}_0)$

- A distribuição amostral de $\hat{Y}_0$, refere-se a diferentes valores de $\hat{Y}_0$ que seriam obtidos se amostras repetidas são selecionadas, cada uma mantendo os níveis da variável regressora X constante, e calculando $\hat{Y}_0$ para cada amostra.

- Para o modelo de regressão linear simples, a distribuição amostral de $\hat{Y}_0$ é Normal com média e variância dadas por:

- $\mathrm{E}(\hat{Y}_0) = \mathrm{E}(\mathrm{E}(\hat{Y}_0)) = \mathrm{E}(Y_0)$

  e

- $\mathrm{Var}(\hat{Y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right].$

Distribuição amostral de $\hat{Y}_0 = \mathrm{E}(\hat{Y}_0)$

- Ao considerar o estimador da variância σ^2, a variância de $\mathrm{Var}(\hat{Y}_0)$ estimada é dada por:

$$\widehat{\mathrm{Var}}(\hat{Y}_0) = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right]$$

- Observações:
  - A variabilidade de $\hat{Y}_0$ é afetada pela distância entre X_0 e $\bar{X}$,
  - Quando $X_0 = 0$, a variância de $\hat{Y}_0$ reduz a variância de $\hat{\beta}_0$ e a estimativa da variância de $\hat{Y}_0$ reduz a $\widehat{\mathrm{Var}}(\hat{\beta}_0)$.,

Distribuição amostral de $\hat{Y}_0 = \mathrm{E}(\hat{Y}_0)$

- Ao considerar a distribuição de $\hat{Y}_0$, a estatística:

$$\frac{\hat{Y}_0 - \mathrm{E}(Y_0)}{\widehat{\mathrm{Var}}(\hat{Y})} \tag{3}$$

tem distribuição t-Student com $(n-2)$ graus de liberdade para o modelo de regressão linear simples.

Intervalo de Confiança

- Ao considerar a distribuição de $\hat{Y}_0$, um Intervalo de Confiança $(1-\alpha)$ para $\mathrm{E}(Y_0)$ é dado por:

$$IC(\mathrm{E}(Y_0); 1-\alpha) = \left(\hat{Y}_0 - t_{(1-\alpha/2;n-2)}\hat{\sigma}\sqrt{A}; \hat{Y}_0 + t_{(1-\alpha/2;n-2)}\hat{\sigma}\sqrt{A}\right)$$

em que $A = \left[\frac{1}{n} + \frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right]$.