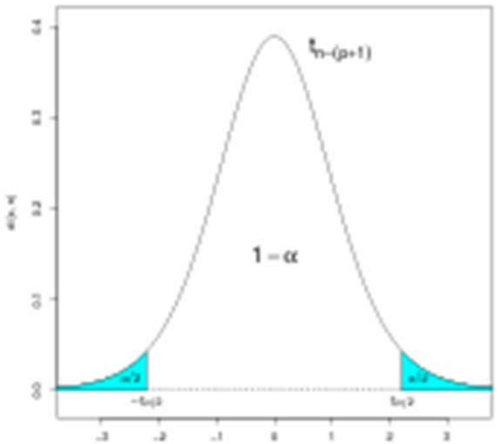
Testes de Hipóteses sobre os parâmetros

O resultado usado para construir ICs também permite Testes a Hipóteses sobre cada β_j. Admitindo a Hipótese Nula $H_0 : \beta_j = c$:

$$T = \frac{\hat{\beta}_j - \overbrace{\beta_{j_{|H_0}}}^{=c}}{\hat{\sigma}_{\hat{\beta}_j}} \frown t_{n-(p+1)} , \qquad \forall\, j = 0, 1, ..., p$$

Rejeita-se H_0 em favor da Hipótese Alternativa $H_1 : \beta_j \neq c$ se o valor calculado de T na amostra, T_{calc}, recair numa das caudas da distribuição.

Fixando o Nível de Significância α, tem-se a Região Crítica:



Testes de Hipóteses (bilateral) a $\hat{\beta}_j$

Testes de Hipóteses a β_j (Modelo de Regressão Linear Múltipla)

Hipóteses: $H_0 : \quad \beta_j = c \quad \text{vs.} \quad H_1 : \quad \beta_j \neq c$

Estatística do Teste: $T = \dfrac{\hat{\beta}_j - \overbrace{\beta_j|_{H_0}}^{=c}}{\hat{\sigma}_{\hat{\beta}_j}} \frown t_{n-(p+1)}$, se H_0 verdade.

Nível de significância do teste: α

Região Crítica (Região de Rejeição bilateral): Rejeitar H_0 se

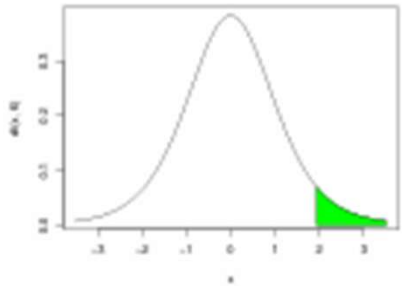
$$T_{calc} > t_{\frac{\alpha}{2}\,[n-(p+1)]} \quad \text{ou} \quad T_{calc} < -t_{\frac{\alpha}{2}\,[n-(p+1)]}$$

$$\iff \quad |T_{calc}| > t_{\frac{\alpha}{2}\,[n-(p+1)]}$$

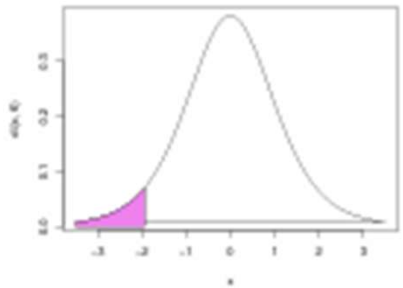
Testes de Hipóteses a $\hat{\beta}_j$ (unilaterais)

$$T = \frac{\hat{\beta}_j - \overbrace{\beta_{j|H_0}}^{=c}}{\hat{\sigma}_{\hat{\beta}_j}} \frown t_{n-(p+1)}$$

Com a Hipótese Alternativa $H_1 : \beta_j > c$, só valores grandes da estatística sugerem a rejeição de $H_0 : \beta_j \leq c$ (ou $H_0 : \beta_j = c$):



Com a Hipótese Alternativa $H_1 : \beta_j < c$, só valores pequenos de T_{calc} sugerem rejeitar $H_0 : \beta_j \geq c$ (ou $H_0 : \beta_j = c$):



Testes de Hipóteses sobre os parâmetros

Dado o Modelo de Regressão Linear Múltipla,

Testes de Hipóteses a β_j (Regressão Linear Múltipla)

Hipóteses: $H_0: \quad \beta_j \begin{array}{c} \geq \\ = \\ \leq \end{array} c \quad \text{vs.} \quad H_1: \quad \beta_j \begin{array}{c} < \\ \neq \\ > \end{array} c$

Estatística do Teste: $T = \dfrac{\hat{\beta}_j - \overbrace{\beta_j|_{H_0}}^{=c}}{\hat{\sigma}_{\hat{\beta}_j}} \frown t_{n-(p+1)}$, se H_0 verdade.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Rejeitar H_0 se

$T_{calc} < -t_{\alpha[n-(p+1)]}$ (Unilateral esquerdo)

$|T_{calc}| > t_{\alpha/2[n-(p+1)]}$ (Bilateral)

$T_{calc} > t_{\alpha[n-(p+1)]}$ (Unilateral direito)

Os *p-values*

Valores de prova (*p-value*)

O *p-value* é a probabilidade da estatística de teste tomar valores mais extremos que o valor calculado a partir da amostra, sob H_0

O cálculo do *p-value* é feito de forma diferente, consoante a natureza da Região Crítica (RC) (unilateral direita ou esquerda, ou bilateral).

Sendo $T \frown t_{n-(p+1)}$

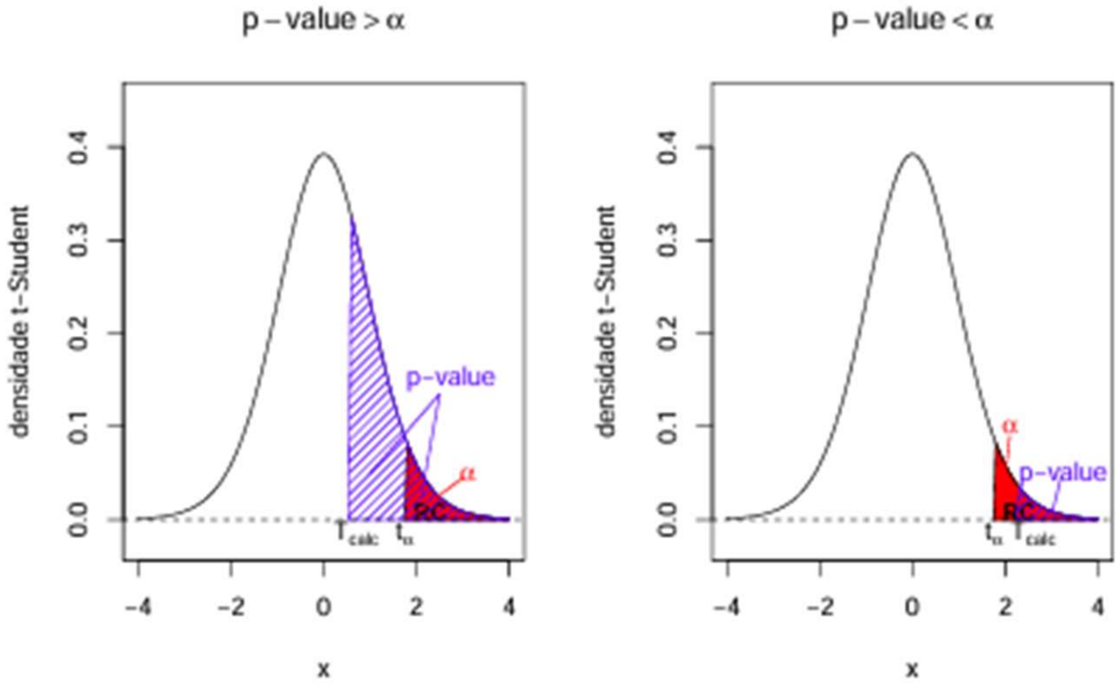
RC Unilateral direita: $p = P[T > T_{calc}]$

RC Unilateral esquerda: $p = P[T < T_{calc}]$

RC Bilateral: $p = 2 \times P[T > |T_{calc}|]$.

A relação de *p-values* e níveis de significância

- $p-value > \alpha \Rightarrow$ não rejeição de H_0 ao nível α;
- $p-value < \alpha \Rightarrow$ rejeição de H_0 ao nível α;



Em geral: p-value muito pequeno implica rejeição H_0.

Ainda o exemplo dos lírios

RLM

```
proc reg data=iris;
   model PetalWidth = SepalLength SepalWidth PetalLength/clb;
run;
```

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-0.24031	0.17837	-1.35	0.1800	-0.59283	0.11221
SepalLength	1	-0.20727	0.04751	-4.36	<.0001	-0.30115	-0.11338
SepalWidth	1	0.22283	0.04894	4.55	<.0001	0.12611	0.31955
PetalLength	1	0.52408	0.02449	21.40	<.0001	0.47568	0.57249

Teste H_0: β_o=0 vs. H_1: $\beta_o \neq$0

Teste H_0: β_1=0 vs. H_1: $\beta_1 \neq$0

Teste H_0: β_2=0 vs. H_1: $\beta_2 \neq$0

Teste H_0: β_3=0 vs. H_1: $\beta_3 \neq$0

Exemplo: $T_{Calc} = \frac{b_3 - \beta_3 | H_0}{\hat{\sigma}_{\hat{\beta}_3}} = \frac{0.52408}{0.02449} = 21.40$

O valor de prova (*p-value*) indica uma claríssima rejeição da hipótese nula para um nível de significância usual

Nota: por exemplo, para o teste H_0: β_3=0.5 vs. H_1: $\beta_3 \neq$0.5

$$T_{Calc} = \frac{b_3 - \beta_3 | H_0}{\hat{\sigma}_{\hat{\beta}_3}} = \frac{0.52408 - 0.5}{0.02449} = 0.983258473$$

$$|T_{Calc}| < 1.96, para\ \alpha = 0.05, não\ se\ rejeita\ a\ hipótese\ nula$$

$$t_{\frac{0.05}{2}(146)} = t_{0.025(146)} \approx 1.96$$

(O valor de prova (*p-value*) da tabela não é válido neste caso)

Combinações lineares dos parâmetros

Seja $\vec{\mathbf{a}} = (a_0, a_1, ..., a_p)^t$ um vector não aleatório em $\mathbb{R}^{p+1}$. O produto interno $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$ define uma combinação linear dos parâmetros do modelo:

$$\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} \;=\; a_0\beta_0 + a_1\beta_1 + a_2\beta_2 + ... + a_p\beta_p \,.$$

Casos particulares importantes são se:

- $\vec{\mathbf{a}}$ tem um único elemento não-nulo, $a_{j+1} = 1$: $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = \beta_j$.
- $\vec{\mathbf{a}}$ só tem dois elementos não-nulos, $a_{i+1} = 1$ e $a_{j+1} = \pm 1$: $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = \beta_i \pm \beta_j$.
- $\vec{\mathbf{a}} = (1, x_1, x_2, ..., x_p)$: $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$ é o valor esperado de Y associado aos valores indicados das variáveis preditoras:

$$\begin{aligned}
\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \\
&= E[Y \,|\, X_1 = x_1, X_2 = x_2, ..., X_p = x_p] \\
&= \mu_{Y|\vec{\mathbf{x}}}
\end{aligned}$$

Inferência sobre combinações lineares dos β_js

Estima-se $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$ com a mesma combinação linear dos estimadores:

$$\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}} \;=\; a_0\hat{\beta}_0 + a_1\hat{\beta}_1 + a_2\hat{\beta}_2 + ... + a_p\hat{\beta}_p \;.$$

Sabemos determinar a distribuição de probabilidades de $\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}$:

- Sabemos que $\hat{\vec{\boldsymbol{\beta}}} \frown \mathcal{N}_{p+1}\left(\vec{\boldsymbol{\beta}}\,,\,\sigma^2\,(\mathbf{X}^t\mathbf{X})^{-1}\right)$
- Logo, $\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}} \frown \mathcal{N}_1(\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}\,,\,\sigma^2\,\vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}\,)$
- Ou seja, $\vec{\mathbf{Z}} \;=\; \frac{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}-\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}}{\sqrt{\sigma^2\,\vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}} \frown \mathcal{N}(0,1)$;
- Por um raciocínio análogo ao usado nos β_j individuais, tem-se:

$$\frac{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}-\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}}{\sqrt{QMRE\cdot\vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}} \frown t_{n-(p+1)} \;.$$

Quantidades centrais para a inferência sobre $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$

Teorema (Distribuições para combinações lineares dos βs)

Dado o Modelo de Regressão Linear Múltipla, tem-se

$$\frac{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}} - \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}}{\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}}} \frown t_{n-(p+1)} ,$$

com $\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}} = \sqrt{QMRE \cdot \vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}$.

Este Teorema dá-nos os resultados que servem de base à construção de intervalos de confiança e testes de hipóteses para quaisquer combinações lineares dos parâmetros β_j do modelo.

Intervalo de confiança para $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$

Intervalo de Confiança a $(1-\alpha)\times 100\%$ para $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$

Dado o Modelo de Regressão Linear Múltipla e uma amostra, o intervalo a $(1-\alpha)\times 100\%$ de confiança para uma combinação linear dos parâmetros, $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} = a_0\beta_0 + a_1\beta_1 + ... + a_p\beta_p$, é:

$$\left]\quad \vec{\mathbf{a}}^t\vec{\mathbf{b}} - t_{\frac{\alpha}{2}[n-(p+1)]}\cdot\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}} \quad , \quad \vec{\mathbf{a}}^t\vec{\mathbf{b}} + t_{\frac{\alpha}{2}[n-(p+1)]}\cdot\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}} \quad\right[,$$

com $\vec{\mathbf{a}}^t\vec{\mathbf{b}} = a_0b_0 + a_1b_1 + ... + a_pb_p$ e $\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}} = \sqrt{QMRE\cdot\vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}$.

Fórmulas para a estimação de $\beta_i \pm \beta_j$

A fórmula geral $\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}} = \sqrt{QMRE \cdot \vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}$ admite uma expressão alternativa no caso particular duma soma ou diferença de dois βs.

Pela fórmula geral da variância duma soma ou diferença de v.a.s,

$$\begin{aligned} V[\hat{\beta}_i \pm \hat{\beta}_j] &= V[\hat{\beta}_i] + V[\hat{\beta}_j] \pm 2\,Cov[\hat{\beta}_i, \hat{\beta}_j] \,. \\ \Leftrightarrow \quad \sigma^2_{\hat{\beta}_i \pm \hat{\beta}_j} &= \sigma^2 \cdot \left[(\mathbf{X}^t\mathbf{X})^{-1}_{[i+1,i+1]} + (\mathbf{X}^t\mathbf{X})^{-1}_{[j+1,j+1]} \pm 2\,(\mathbf{X}^t\mathbf{X})^{-1}_{[i+1,j+1]} \right] \,. \end{aligned}$$

Logo, o erro padrão de $\hat{\beta}_i \pm \hat{\beta}_j$ é:

$$\hat{\sigma}_{\hat{\beta}_i \pm \hat{\beta}_j} = \sqrt{QMRE \cdot \left[(\mathbf{X}^t\mathbf{X})^{-1}_{[i+1,i+1]} + (\mathbf{X}^t\mathbf{X})^{-1}_{[j+1,j+1]} \pm 2\,(\mathbf{X}^t\mathbf{X})^{-1}_{[i+1,j+1]} \right]} \,.$$

ICs para combinações lineares

Numa RLM, o IC duma combinação linear genérica $\vec{\mathbf{a}}^t\vec{\beta}$, precisa da matriz das (co)variâncias estimadas dos estimadores $\hat{\vec{\beta}}$,

$$\widehat{V[\hat{\vec{\beta}}]} = QMRE \cdot (\mathbf{X}^t\mathbf{X})^{-1} \, ,$$

que é gerada

```
proc reg data=iris;
   model PetalWidth = SepalLength SepalWidth PetalLength/clb covb xpx;
run;
```

A matriz das (co)variâncias estimadas no exemplo RLM dos lírios é:

Covariance of Estimates				
Variable	Intercept	SepalLength	SepalWidth	PetalLength
Intercept	0.0318157664	-0.005075942	-0.002486105	0.0015144174
SepalLength	-0.005075942	0.0022568367	-0.001344002	-0.001065046
SepalWidth	-0.002486105	-0.001344002	0.0023949317	0.000802941
PetalLength	0.0015144174	-0.001065046	0.000802941	0.0005998259

ICs para combinações lineares

O erro padrão estimado de $\hat{\beta}_1 + \hat{\beta}_2$

$$\hat{\sigma}_{\hat{\beta}_1+\hat{\beta}_2} = \sqrt{\hat{V}[\hat{\beta}_1] + \hat{V}[\hat{\beta}_2] + 2\widehat{Cov}[\hat{\beta}_1, \hat{\beta}_2]}$$

$$\hat{\sigma}_{\hat{\beta}_1+\hat{\beta}_2} = \sqrt{0.0022568367 + 0.0023949317 + 2(-0.001344002)} = 0.04431439$$

Covariance of Estimates				
Variable	Intercept	SepalLength	SepalWidth	PetalLength
Intercept	0.0318157664	-0.005075942	-0.002486105	0.0015144174
SepalLength	-0.005075942	0.0022568367	-0.001344002	-0.001065046
SepalWidth	-0.002486105	-0.001344002	0.0023949317	0.000802941
PetalLength	0.0015144174	-0.001065046	0.000802941	0.0005998259

0

Testes a combinações lineares dos parâmetros

Dado o Modelo de Regressão Linear Múltipla,

Testes de Hipóteses relativos a $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$

Hipóteses: $H_0 : \quad \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} \;\begin{matrix} \geq \\ = \\ \leq \end{matrix}\; c \quad \text{vs.} \quad H_1 : \quad \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}} \;\begin{matrix} < \\ \neq \\ > \end{matrix}\; c$

Estatística do Teste: $T = \dfrac{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}} - \overbrace{\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}|_{H_0}}^{=c}}{\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}}} \frown t_{n-(p+1)}$, se H_0 verdade

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Rejeitar H_0 se

$T_{calc} < -t_{\alpha[n-(p+1)]}$ (Unilateral esquerdo)

$|T_{calc}| > t_{\alpha/2[n-(p+1)]}$ (Bilateral)

$T_{calc} > t_{\alpha[n-(p+1)]}$ (Unilateral direito)

Intervalos de confiança para $E[Y|X_1=x_1,...,X_p=x_p]$

Como caso particular do resultado anterior, tem-se:

IC para o valor esperado de Y, dados os preditores

Dado o Modelo RLM e uma amostra com os valores $\vec{\mathbf{x}}=(x_1,x_2,...,x_p)^t$ das variáveis preditoras, o valor esperado de Y,

$$\mu_{Y|\vec{\mathbf{x}}} = E[Y|X_1=x_1,...,X_p=x_p] = \beta_0+\beta_1 x_1+...+\beta_p x_p \ ,$$

é estimado por $\hat{\mu}_{Y|\vec{\mathbf{x}}} = b_0+b_1\,x_1+...+b_p\,x_p$.

Um intervalo a $(1-\alpha)\times 100\%$ de confiança para $\mu_{Y|\vec{\mathbf{x}}}$ é dado por:

$$\Big]\ \hat{\mu}_{Y|\vec{\mathbf{x}}} - t_{\frac{\alpha}{2}\,[n-(p+1)]}\cdot\hat{\sigma}_{\hat{\mu}_{Y|\vec{\mathbf{x}}}} \ ,\ \hat{\mu}_{Y|\vec{\mathbf{x}}} + t_{\frac{\alpha}{2}\,[n-(p+1)]}\cdot\hat{\sigma}_{\hat{\mu}_{Y|\vec{\mathbf{x}}}}\ \Big[\ ,$$

com $\hat{\sigma}_{\hat{\mu}_{Y|\vec{\mathbf{x}}}} = \sqrt{QMRE\cdot\vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}$, onde $\vec{\mathbf{a}}=(1,x_1,x_2,...,x_p)$.

Se $p = 1$, RLS

Fórmulas para uma regressão linear simples

Numa regressão linear simpes, a fórmula da variância de $\hat{\mu}_{Y|x}$ é:

$$\sigma^2_{\hat{\mu}_{Y|x}} = V[\hat{\mu}_{Y|x}] = \sigma^2 \cdot \left[\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\cdot s_x^2}\right]$$

$$\Longrightarrow \quad \hat{\sigma}^2_{\hat{\mu}_{Y|x}} = QMRE \cdot \left[\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\cdot s_x^2}\right]$$

O intervalo de confiança para $\mu_{Y|x}$ na RLS é:

$$\left] \ (b_0 + b_1\, x) - t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\mu}_{Y|x}} \ , \ (b_0 + b_1\, x) + t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\mu}_{Y|x}} \ \right[.$$

A variabilidade duma observação individual de Y

Consideraram-se intervalos de confiança para o valor esperado de Y,

$$\mu_{Y|\vec{\mathbf{x}}} = E[Y|x_1{=}x_1,x_2{=}x_2,...,x_p{=}x_p] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \ ,$$

usam a variabilidade associada ao estimador $\hat{\mu}_{Y|\vec{\mathbf{x}}}$:

$$\sigma^2_{\hat{\mu}_{Y|\vec{\mathbf{x}}}} = V[\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p] = \sigma^2 \cdot \vec{\mathbf{a}}^t (\mathbf{X}^t\mathbf{X})^{-1} \vec{\mathbf{a}},$$

com $\vec{\mathbf{a}} = (1, x_1, x_2, ..., x_p)$.

Uma observação individual de Y tem uma variabilidade adicional, pois:

$$Y \quad = \quad \mu_{Y|\vec{\mathbf{x}}} + \varepsilon \quad = \quad \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon \ .$$

A flutuação aleatória de observações individuais em torno do hiperplano é $V[\varepsilon] = \sigma^2$. Será necessário somar a variância associada à estimação do hiperplano e a variância das observações individuais:

$$\sigma^2_{Indiv} \ = \ V[\hat{\mu}_{Y|\vec{\mathbf{x}}}] + V[\varepsilon] \ = \ \sigma^2 \cdot \vec{\mathbf{a}}^t (\mathbf{X}^t\mathbf{X})^{-1} \vec{\mathbf{a}} + \sigma^2 \ = \ \sigma^2 \cdot \left[\vec{\mathbf{a}}^t (\mathbf{X}^t\mathbf{X})^{-1} \vec{\mathbf{a}} + 1 \right] .$$

Intervalos de predição para Y

Podem obter-se intervalos de predição para uma observação individual de Y, associada aos valores $X_1 = x_1, ..., X_p = x_p$ das variáveis preditoras.

Nestes intervalos, a estimativa da variância duma observação individual de Y é a estimativa de σ^2_{indiv}, resultante de substituir σ^2 pelo $QMRE$ amostral:

Intervalos de predição para observações individuais

$$\left] \quad \hat{\mu}_{Y|\vec{\mathbf{x}}} - t_{\frac{\alpha}{2}[n-(p+1)]} \cdot \hat{\sigma}_{indiv} \quad , \quad \hat{\mu}_{Y|\vec{\mathbf{x}}} + t_{\frac{\alpha}{2}[n-(p+1)]} \cdot \hat{\sigma}_{indiv} \quad \right[$$

onde

$$\hat{\mu}_{Y|X} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p$$

e

$$\hat{\sigma}_{indiv} = \sqrt{QMRE\left[1+\vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}\right]} \quad \text{com} \quad \vec{\mathbf{a}}=(1,x_1,x_2,...,x_p).$$

Se $p = 1$, RLS

Fórmulas para a regressão linear simples

Na regressão linear simples usa-se a fórmula

$$\sigma^2_{Indiv} = \underbrace{\sigma^2 \cdot \left[\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\cdot s_x^2}\right]}_{=V[\hat{\mu}_{Y|\vec{x}}]} + \underbrace{\sigma^2}_{=V[\varepsilon]} = \sigma^2 \cdot \left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\cdot s_x^2}\right] .$$

Logo,

RLS: Intervalo de predição para observação individual de Y

$$\left] \;\hat{\mu}_{Y|x} - t_{\alpha/2(n-2)} \cdot \hat{\sigma}_{Indiv} \quad , \quad \hat{\mu}_{Y|x} + t_{\alpha/2(n-2)} \cdot \hat{\sigma}_{Indiv}\; \right[.$$

com $\hat{\mu}_{Y|x} = b_0 + b_1 x$ e $\hat{\sigma}_{Indiv} = \sqrt{QMRE \cdot \left[1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)\cdot s_x^2}\right]}$.

Quer numa regressão linear simples, quer numa múltipla, estes intervalos são necessariamente de maior amplitude que os intervalos de confiança para $\mu_{Y|\vec{x}}$ (para igual nível de confiança $(1-\alpha)\times 100\%$).

Testando a qualidade do ajustamento global

Numa Regressão Linear, o modelo é inútil se fôr indistinguível do modelo nulo, i.e., do modelo de equação $Y_i = \beta_0 + \varepsilon_i$. O modelo nulo pode ser visto como um submodelo de qualquer modelo linear, em que todas as variáveis preditoras têm coeficiente nulo: $\beta_j = 0$, $\forall j > 0$.

O teste de ajustamento global visa testar se um dado modelo linear é significativamente diferente do modelo nulo.

As hipóteses em confronto são:

$$H_0 : \ \beta_1 = \beta_2 = ... = \beta_p = 0$$
$$[\text{MODELO COMPLETO} \equiv \text{MODELO NULO}]$$
$$\text{vs.}$$
$$H_1 : \ \exists j = 1, ..., p \quad \text{t.q.} \quad \beta_j \neq 0$$
$$[\text{MODELO COMPLETO} \not\equiv \text{MODELO NULO}]$$

NOTA: repare que β_0 não intervém nas hipóteses.

O teste de ajustamento global (cont.)

Definindo:

- O Quadrado Médio da Regressão como $QMR = \frac{SQR}{p}$.
- O Quadrado Médio Residual como $QMRE = \frac{SQRE}{n-(p+1)}$.

Sob a Hipótese Nula do teste de ajustamento global:

$$F = \frac{QMR}{QMRE} \frown F_{[\,p\,,\,n-(p+1)\,]}\,.$$

Esta é a estatística F do teste de ajustamento global.

Expressão alternativa para a estatística do teste F

A estatística do teste F de ajustamento global do modelo numa Regressão Linear Múltipla pode ser escrita na forma alternativa:

$$F = \frac{n-(p+1)}{p} \cdot \frac{R^2}{1-R^2} .$$

A estatística F é uma função crescente do coeficiente de determinação amostral R^2, o que justifica a natureza unilateral direita da região crítica.

As hipóteses do teste também se podem escrever como

$$H_0 : \mathscr{R}^2 = 0 \qquad \text{vs.} \qquad H_1 : \mathscr{R}^2 > 0 .$$

A hipótese $H_0 : \mathscr{R}^2 = 0$ indica ausência de relação linear entre Y e o conjunto dos preditores. Corresponde a um ajustamento "péssimo" do modelo. A sua rejeição não garante um bom ajustamento.

O Teste F de ajustamento global do Modelo

Teste F de ajustamento global do modelo RLM

Hipóteses: $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$

vs.

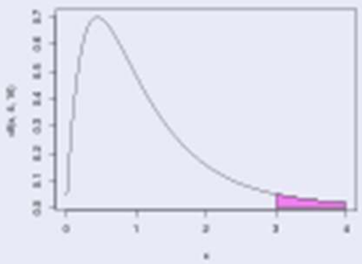
$H_1 : \exists\, j = 1, ..., p \text{ tal que } \beta_j \neq 0.$

Estatística do Teste: $F = \frac{QMR}{QMRE} \frown F_{[p,n-(p+1)]}$ se H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se $F_{calc} > f_{\alpha[p,n-(p+1)]}$



Outra formulação do teste F de ajustamento global

Teste F de ajustamento global do modelo RLM (alternativa)

Hipóteses: $H_0 : \mathscr{R}^2 = 0$ vs. $H_1 : \mathscr{R}^2 > 0$.

Estatística do Teste: $F = \frac{n-(p+1)}{p} \cdot \frac{R^2}{1-R^2} \frown F_{[p,n-(p+1)]}$ se H_0.

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se $F_{calc} > f_{\alpha(p,n-(p+1))}$

A hipótese nula $H_0 : \mathscr{R}^2 = 0$ afirma que, na população, o coeficiente de determinação é nulo.

Ainda o exemplo dos lírios

Informação Teste F de ajustamento Global

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	81.18964	27.06321	734.39	<.0001
Error	146	5.38030	0.03685		
Corrected Total	149	86.56993			

Root MSE	0.19197	R-Square	0.9379
Dependent Mean	1.19933	Adj R-Sq	0.9366
Coeff Var	16.00615		

O R^2 modificado (adjusted R^2)

O Coeficiente de Determinação usual define-se como:

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQRE}{SQT}$$

O R^2 modificado, sendo $QMT = \frac{SQT}{n-1} = s_y^2$, é:

$$R^2_{mod} = 1 - \frac{QMRE}{QMT} = 1 - \frac{SQRE}{SQT} \cdot \frac{n-1}{n-(p+1)} = 1 - (1 - R^2) \cdot \frac{n-1}{n-(p+1)}.$$

Para qualquer modelo linear (com preditores), tem-se: $R^2_{mod} < R^2$.
Se $n \gg p+1$ (muito mais observações que parâmetros), $R^2 \approx R^2_{mod}$.
Se n é pouco maior que p, $R^2_{mod} \ll R^2$ (excepto se $R^2 \approx 1$).

$\frac{QMRE}{QMT} = \frac{\hat{\sigma}^2}{s_y^2}$ é a proporção da variabilidade total de Y que permanece inexplicada após a introdução dos preditores. Logo, R^2_{mod} é o ganho na explicação de s_y^2 associado ao modelo.

Root MSE	0.19197	R-Square	0.9379
Dependent Mean	1.19933	Adj R-Sq	0.9366
Coeff Var	16.00615		