Inferência em Regressão Múltipla

Econometria

Alexandre Gori Maia

Ementa:

- Variância dos estimadores;
- Teste *t* para os coeficientes;
- Combinação linear dos parâmetros;
- Intervalo de Confiança para os valores previstos;

Bibliografia Básica:

Maia, Alexandre Gori (2017). *Econometria: Conceitos e Aplicações*. Cap. 8.

Variância dos Estimadores

Variância dos estimadores de MQO:

Seja o modelo de RLM:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{e}$$

Caso os pressupostos do Teorema de Gauss-Markov sejam válidos, o método de MQO oferecerá estimadores não viesados para os coeficientes do modelo e para suas respectivas variâncias. As variâncias dos estimadores e seus respetivos estimadores serão dados por:

$$Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 \quad \Rightarrow \quad S^2_{\hat{\boldsymbol{\beta}}} = (\mathbf{X}^T\mathbf{X})^{-1}\hat{\sigma}^2$$

Onde σ^2 é a variânica dos erros ou variância da regressão e $\hat{\sigma}^2$ seu respectivo estimador, dado por:

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{e}}^T\hat{\mathbf{e}}}{n-(k+1)} = \frac{\mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y}}{n-(k+1)}$$

Teste *t* para Coeficientes

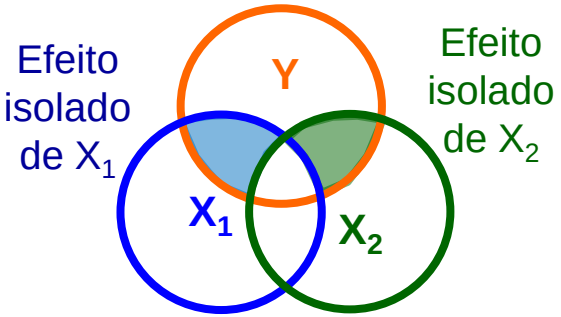
Seja o modelo de RLM:

$$Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + e_i$$

Para saber se X_j possui de fato relação isolada com Y, podemos realizar um teste de hipóteses para o coeficiente β_j :

$\mathbf{H_0}$: $\beta_j = 0$ (*Variável X_j não tem relação isolada com Y*)

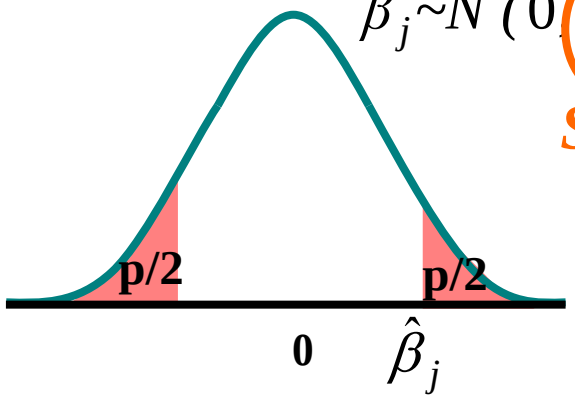
$\mathbf{H_1}$: $\beta_j \neq 0$ (*Variável X_j não tem relação isolada com Y*)



Sob as premissas do MCRL, o estimador de MQO para β_j , além de não viesado e de variância mínima, apresentará distribuição normal. Assim, caso H_0 seja válido, sua distribuição será dada por:

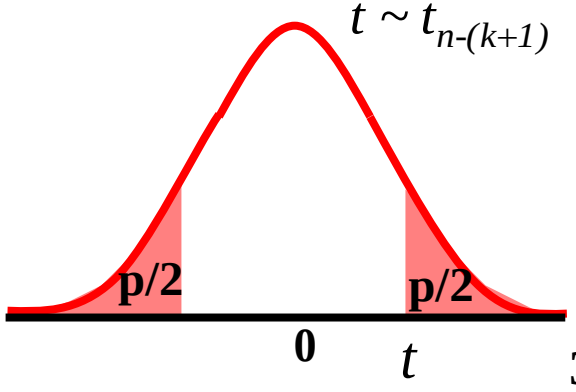
$$\hat{\beta}_j \sim N(0, \sigma^2_{\hat{\beta}_j})$$

$S^2_{\hat{\beta}_j}$



$$t = \frac{\hat{\beta}_j - 0}{S_{\hat{\beta}_j}}$$

$$t \sim t_{n-(k+1)}$$

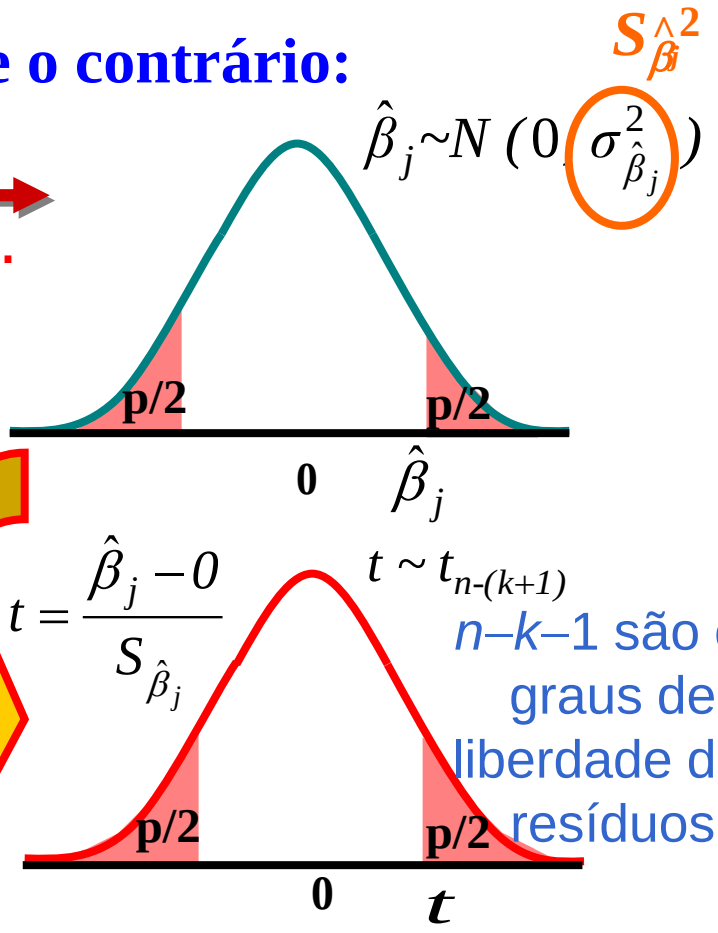


Teste *t* para Coeficientes

O parâmetro será nulo até que se prove o contrário:

$\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases}$

Se H_0 é verdadeiro, então...



n–k–1 são os graus de liberdade dos resíduos

Dados:

$\hat{\beta}_j$ = estimador de MQ

$S_{\hat{\beta}_j}$ = erro padrão do estimador

valor p = nível de significância observado

Conclusão:

Rejeito H_0 se valor *p* (prob de erro ao rejeitar H_0) for baixa. Nestas circunstâncias, posso afirmar que o estimador é significativo.

Teste *t* para Coeficientes - Exemplo

Sejam os dados da relação entre renda familiar (*Y*), anos de estudo (X_1) e idade (X_2) do responsável pela família:

$$Y_i = 1{,}9 + 1X_{1_i} + 0{,}06X_{2_i} + \hat{e}_i$$

A matriz de estimativas das variâncias e covariâncias dos coeficientes será:

$$\mathbf{S}^2_{\hat{\boldsymbol{\beta}}} = (\mathbf{X}^T\mathbf{X})^{-1}\hat{\sigma}^2 = \begin{pmatrix} 4 & 18 & 140 \\ 18 & 102 & 730 \\ 140 & 730 & 5400 \end{pmatrix}^{-1} \hat{\sigma}^2 = \begin{pmatrix} 8{,}95 & 2{,}5 & -0{,}57 \\ 2{,}5 & 1 & -0{,}2 \\ -0{,}57 & -0{,}2 & 0{,}042 \end{pmatrix} 0{,}2 = \begin{pmatrix} 1{,}79 & 0{,}5 & -0{,}114 \\ 0{,}5 & 0{,}2 & -0{,}04 \\ -0{,}114 & -0{,}04 & 0{,}0084 \end{pmatrix}$$

$S^2_{\hat{\beta}j}$

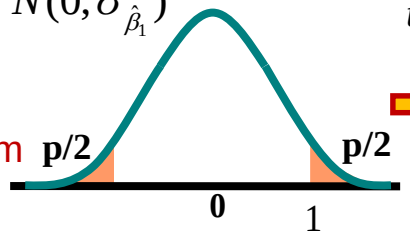
Realizando testes de hipóteses para os coeficientes angulares:

H_0: $\beta_1 = 0$

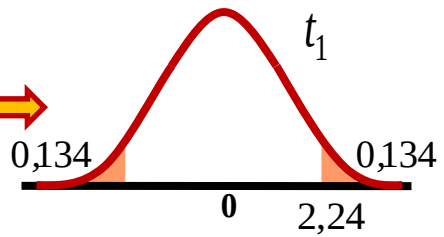
H_1: $\beta_1 \neq 0$

Se afirmarmos que anos de estudo tenham efeito isolado sobre a renda, estaremos sujeitos a um erro de 27%

$$\hat{\beta}_1 \sim N(0, \sigma^2_{\hat{\beta}_1})$$



$$t = \frac{1-0}{\sqrt{0{,}2}} = 2{,}24$$

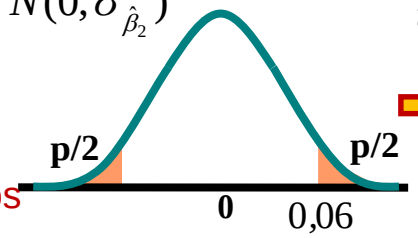


H_0: $\beta_2 = 0$

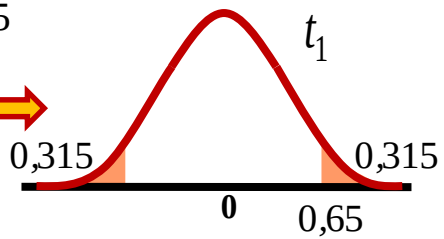
H_1: $\beta_2 \neq 0$

Se afirmarmos que a idade tenha efeito isolado sobre a renda, estaremos sujeitos a um erro de 63%

$$\hat{\beta}_2 \sim N(0, \sigma^2_{\hat{\beta}_2})$$



$$t = \frac{0{,}06-0}{\sqrt{0{,}0084}} = 0{,}65$$



Combinação Linear de Parâmetros

Combinação linear dos parâmetros:

Dado o modelo de RLM:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Sob as premissas do Teorema de Gauss-Markov, temos que:

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \quad \text{e} \quad Var(\hat{\boldsymbol{\beta}}) = E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T] = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$$

Uma combinação linear dos parâmetros, onde cada coeficiente β_j seja multiplicado por uma constante c_j , seria dado por:

$$c_0\alpha + c_1\beta_1 + \ldots + c_k\beta_k = \begin{pmatrix} c_0 & c_1 & \ldots & c_k \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \ldots \\ \beta_k \end{pmatrix} = \mathbf{c}^T\boldsymbol{\beta}$$

Não é difícil demonstrar que:

$$E(\mathbf{c}^T\hat{\boldsymbol{\beta}}) = \mathbf{c}^T E(\hat{\boldsymbol{\beta}}) = \mathbf{c}^T\boldsymbol{\beta}$$

ou seja, uma função linear dos estimadores de MQO é um estimador não viesado da função linear dos parâmetros

E:

$$Var(\mathbf{c}^T\hat{\boldsymbol{\beta}}) = E[(\mathbf{c}^T\hat{\boldsymbol{\beta}} - \mathbf{c}^T\boldsymbol{\beta})(\mathbf{c}^T\hat{\boldsymbol{\beta}} - \mathbf{c}^T\boldsymbol{\beta})^T] = \mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}\sigma^2$$

a variância da função linear será uma grandeza escalar.

A combinação linear dos parâmetros pode ser utilizada em teste de hipóteses para combinações dos parâmetros ou intervalos de confiança para os valores previstos.

Teste *t* para Combinação Linear

Seja o ajuste de MQ:

$$Y_i = \hat{\alpha} + \hat{\beta}_1 X_{1_i} + \hat{\beta}_2 X_{2_i} + \ldots + \hat{\beta}_k X_{k_i} + \hat{e}_i$$

Suponha as seguintes hipóteses para os parâmetros β_1 e β_2 do modelo:

H_0: $\beta_1 = \beta_2$

H_1: $\beta_1 \neq \beta_2$

Testar a hipótese nula que $\beta_1 = \beta_2$ é o mesmo que testar a combinação:

$$(0)\alpha + (1)\beta_1 + (-1)\beta_2 \ldots + (0)\beta_k = 0$$

ou, matricialmente: $\begin{pmatrix} 0 & 1 & -1 & 0 & \ldots & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \ldots \\ \beta_k \end{pmatrix} = \mathbf{c}^T \boldsymbol{\beta} = 0$

A estatística de teste será: $\begin{pmatrix} 0 & 1 & -1 & 0 & \ldots & 0 \end{pmatrix} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \ldots \\ \hat{\beta}_k \end{pmatrix} = \mathbf{c}^T \hat{\boldsymbol{\beta}}$

Como combinação linear de v.a.s normais é também uma normal, teremos: $\mathbf{c}^T \hat{\boldsymbol{\beta}} \sim N(\mathbf{c}^T \boldsymbol{\beta}, \sigma^2_{\mathbf{c}^T \hat{\boldsymbol{\beta}}})$

Onde: $\sigma^2_{\mathbf{c}^T \hat{\boldsymbol{\beta}}} = \mathbf{c}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{c} \sigma^2$

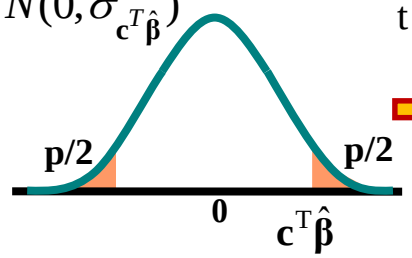
Finalmente, para resolver o teste de hipóteses:

H_0: $\beta_1 = \beta_2$

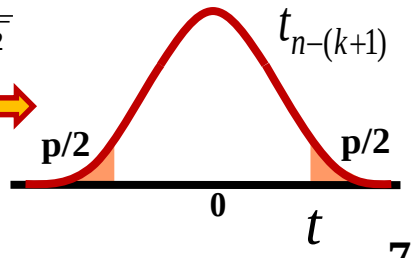
H_1: $\beta_1 \neq \beta_2$

Rejeito H_0 quando o valor p, probabilidade de erro ao rejeitar H_0, for suficientemente pequeno.

$$\mathbf{c}^T \hat{\boldsymbol{\beta}} \sim N(0, \sigma^2_{\mathbf{c}^T \hat{\boldsymbol{\beta}}})$$



$$t = \frac{\mathbf{c}^T \hat{\boldsymbol{\beta}} - 0}{\sqrt{\mathbf{c}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{c} \hat{\sigma}^2}}$$



Teste *t* para Combinação Linear - Exemplo

Sejam os dados da relação entre renda familiar (*Y*), anos de estudo (X_1) e idade (X_2) do responsável pela família:

$$Y_i = 1{,}9 + 1X_{1_i} + 0{,}06X_{2_i} + \hat{e}_i$$

Há evidências de que o efeito dos anos de estudo seja maior que o da idade?

H₀: $\beta_1 = \beta_2$

H₁: $\beta_1 > \beta_2$

Testar a hipótese nula que $\beta_1 = \beta_2$ é o mesmo que testar a combinação:

$$(0)\alpha + (1)\beta_1 + (-1)\beta_2 = 0$$

ou, matricialmente: $\begin{pmatrix} 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} = \mathbf{c}^T\boldsymbol{\beta} = 0$

E a estatística de teste será: $\mathbf{c}^T\hat{\boldsymbol{\beta}} = \begin{pmatrix} 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} 1{,}9 \\ 1 \\ 0{,}06 \end{pmatrix} = 0{,}94$

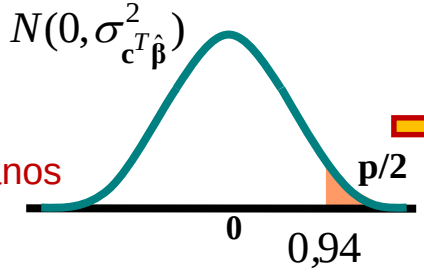
Que estará normalmente distribuída e a estimativa de sua variância será:

$$S^2_{\mathbf{c}^T\hat{\boldsymbol{\beta}}} = \mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}\hat{\sigma}^2 = \begin{pmatrix} 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} 8{,}95 & 2{,}5 & -0{,}57 \\ 2{,}5 & 1 & -0{,}2 \\ -0{,}57 & -0{,}2 & 0{,}042 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} 0{,}2 = 0{,}2884$$

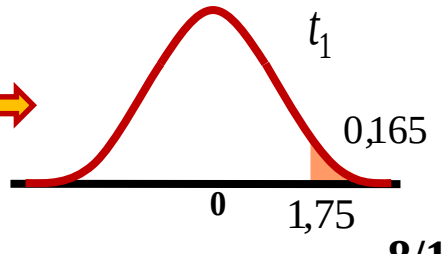
H₀: $\beta_1 = \beta_2$

H₁: $\beta_1 > \beta_2$

$$\mathbf{c}^T\hat{\boldsymbol{\beta}} \sim N(0, \sigma^2_{\mathbf{c}^T\hat{\boldsymbol{\beta}}})$$



$$t = \frac{0{,}94 - 0}{\sqrt{0{,}2884}}$$



Se afirmarmos que o efeito isolado dos anos de escolaridade é superior ao da idade estaremos sujeitos a um erro de 16,5%

Intervalo de Confiança

Seja o modelo de RLM:

$$Y_i = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + ... + \beta_k X_{k_i} + e_i$$

A esperança condicional de *Y* também é uma combinação linear dos parâmetros:

$$E(Y_i / X_{1_i}, ..., X_{k_i}) = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + ... + \beta_k X_{k_i}$$

ou, matricialmente:

$$E(Y_i / X_{1_i}, ..., X_{k_i}) = \mathbf{x}^T\boldsymbol{\beta} = \begin{pmatrix} 1 & X_{1_i} & X_{2_i} & \dots & X_{k_i} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \dots \\ \beta_k \end{pmatrix}$$

E seu ajuste de MQ:

$$Y_i = \hat{\alpha} + \hat{\beta}_1 X_{1_i} + \hat{\beta}_2 X_{2_i} + ... + \hat{\beta}_k X_{k_i} + e_i$$

Assim como o valor previsto da amostra é uma combinação linear dos estimadores:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_{1_i} + \hat{\beta}_2 X_{2_i} + ... + \hat{\beta}_k X_{k_i}$$

ou, matricialmente:

$$\hat{Y}_i = \mathbf{x}^T\hat{\boldsymbol{\beta}} = \begin{pmatrix} 1 & X_{1_i} & X_{2_i} & \dots & X_{k_i} \end{pmatrix} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \dots \\ \hat{\beta}_k \end{pmatrix}$$

Onde: $\mathbf{x}^T\hat{\boldsymbol{\beta}} \sim N(\mathbf{x}^T\boldsymbol{\beta}, \sigma^2_{\mathbf{x}^T\hat{\boldsymbol{\beta}}})$

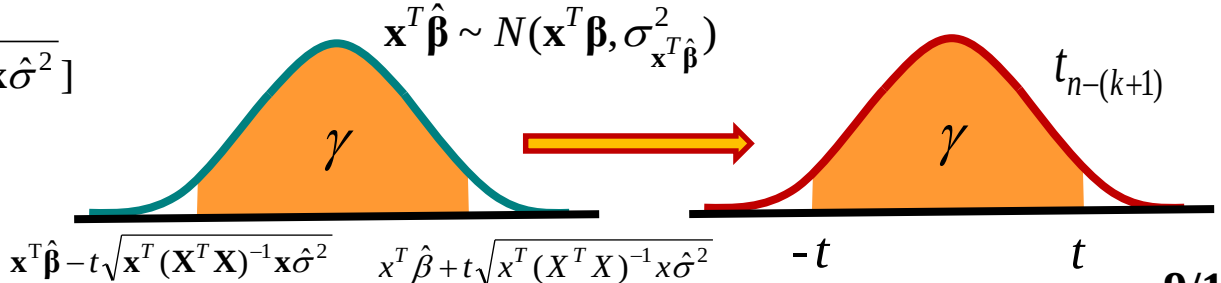
E: $\sigma^2_{\mathbf{x}^T\hat{\boldsymbol{\beta}}} = \mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}\sigma^2$

Podemos, por exemplo, realizar intervalos de confiança para previsões:

Para uma confiança igual a γ:

$$IC[E(Y_i); \gamma] = [\mathbf{x}^T\hat{\boldsymbol{\beta}} \pm t\sqrt{\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}\hat{\sigma}^2}]$$

Intervalo que, em reptidas amostras, conterá a real esperança condicional de Y em γ das situações.



Intervalo de Confiança para Previsão

Sejam os dados da relação entre renda familiar (*Y*), anos de estudo (X_1) e idade (X_2) do responsável pela família:

$$Y_i = 1{,}9 + 1X_{1_i} + 0{,}0{,}6X_{2_i} + \hat{e}_i$$

Qual seria a estimativa por intervalo para a renda esperada de uma família com responsável com nível superior completo (X_1=15) e 30 anos de idade(X_2=30)?

A renda prevista pelo ajuste de MQ seria:

$$\hat{Y}_i = 1{,}9 + 1(15) + 0{,}06(30) = 18{,}7$$

ou seja, 18,7 Salários Mínimos

Ou, matricialmente:

$$\hat{Y}_i = x^T\hat{\beta} = \begin{pmatrix}1 & 15 & 30\end{pmatrix}\begin{pmatrix}1{,}9\\ 1\\ 0{,}06\end{pmatrix} = 18{,}7$$

onde: $\hat{Y}_i \sim N(E(Y_i), x^T(X^TX)^{-1}x\sigma^2)$

A estimativa da variância do valor previsto será:

$$S^2_{x^T\hat{\beta}} = x^T(X^TX)^{-1}x\hat{\sigma}^2 = \begin{pmatrix}1 & 15 & 30\end{pmatrix}\begin{pmatrix}8{,}95 & 2{,}5 & -0{,}57\\ 2{,}5 & 1 & -0{,}2\\ -0{,}57 & -0{,}2 & 0{,}042\end{pmatrix}\begin{pmatrix}1\\ 15\\ 30\end{pmatrix}0{,}2 = 26{,}51$$

Para uma confiança igual a *95%*:

$$IC[E(Y_i);0{,}95] = [6{,}983 \pm 0{,}452]$$

Estimativa do intervalo de 95% de confiança para a real renda esperada de uma família com tais características.