

Universidade de Brasília
Departamento de Estatística
Disciplina: Regressão Linear Múltipla
Professora: Maria Teresa Leão

18/0111558 - João Kramer

Lista 2

1. Considere a função resposta: $E(Y) = 25 + 3X_1 + 4X_2 + 1,5X_1X_2$

a) Faça o gráfico de $E(Y) \times X_1$ quando $X_2 = 3$ e $X_2 = 6$.

b) Os efeitos de X_1 e X_2 são aditivos? Como você identificou isto no gráfico obtido no item a.

a)

$$X_2=3 \Rightarrow E(Y) = 33 + 7.5 \cdot X_1$$

$$X_2=6 \Rightarrow E(Y) = 49 + 12 \cdot X_1$$

b)

Os efeitos são aditivos porque quando aumenta X_2 , aumenta o intercepto o coeficiente de inclinação da reta que representa $E(Y)$.

2. Estabeleça a matriz \mathbf{X} e os vetores \mathbf{Y} e $\boldsymbol{\beta}$ para os seguintes modelos (assuma que $i = 1, 2, 3, 4$).

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

b) $\sqrt{Y_i} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$

a)

$$\mathbf{Y} = \text{transposta } [Y_1 \ Y_2 \ Y_3 \ Y_4] \Rightarrow 4 \times 1$$

$$\boldsymbol{\beta} = \text{transposta } [\beta_0 \ \beta_1 \ \beta_2] \Rightarrow 3 \times 1$$

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ 1 & X_{31} & X_{32} \\ 1 & X_{41} & X_{42} \end{bmatrix} \Rightarrow 4 \times 3$$

b)

$Y = \text{transposta} [\text{Sqrt}(Y_1) \text{ sqrt}(Y_2) \text{ sqrt}(Y_3) \text{ sqrt}(Y_4)] \Rightarrow 4 \times 1$

$\text{Beta} = \text{transposta} [B_0 \ B_1 \ B_2] \Rightarrow 3 \times 1$

$X = \begin{bmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ 1 & X_{31} & X_{32} \\ 1 & X_{41} & X_{42} \end{bmatrix} \Rightarrow 4 \times 3$

-
3. Por que não é significativo atribuir um sinal ao coeficiente de correlação múltipla, embora façamos isso para o coeficiente de correlação linear simples?

Porque uma variável explicativa pode estar correlacionada positivamente com uma e negativamente com outra, então nessa situação o sinal não tem sentido.

4. Exercícios 6.5 a 6.8 do livro-texto.

6.5. Brand preference. In a small-scale experimental study of the relation between degree of brand liking (Y) and moisture content (X_1) and sweetness (X_2) of the product, the following results were obtained from the experiment based on a completely randomized design (data are coded):

$i:$	1	2	3	...	14	15	16
$X_{i1}:$	4	4	4	...	10	10	10
$X_{i2}:$	2	4	2	..	4	2	4
$Y_i:$	64	73	61	...	95	94	100

```
libname reg '/home/u37563508/ARegLinear_20223_1';
options ls=80 ps=60 nodate;
```

```
proc import
  datafile="/home/u37563508/ARegLinear_20223_1/Dados_ex_6_5.xlsx"
  out=reg.lista2a DBMS=xlsx REPLACE;
  sheet="Sheet1";
  GETNAMES=YES;
run;
```

-
- a. Obtain the scatter plot matrix and the correlation matrix. What information do these diagnostic aids provide here?

```
proc corr data=reg.lista2;  
  var y x1 x2 ;  
run;
```

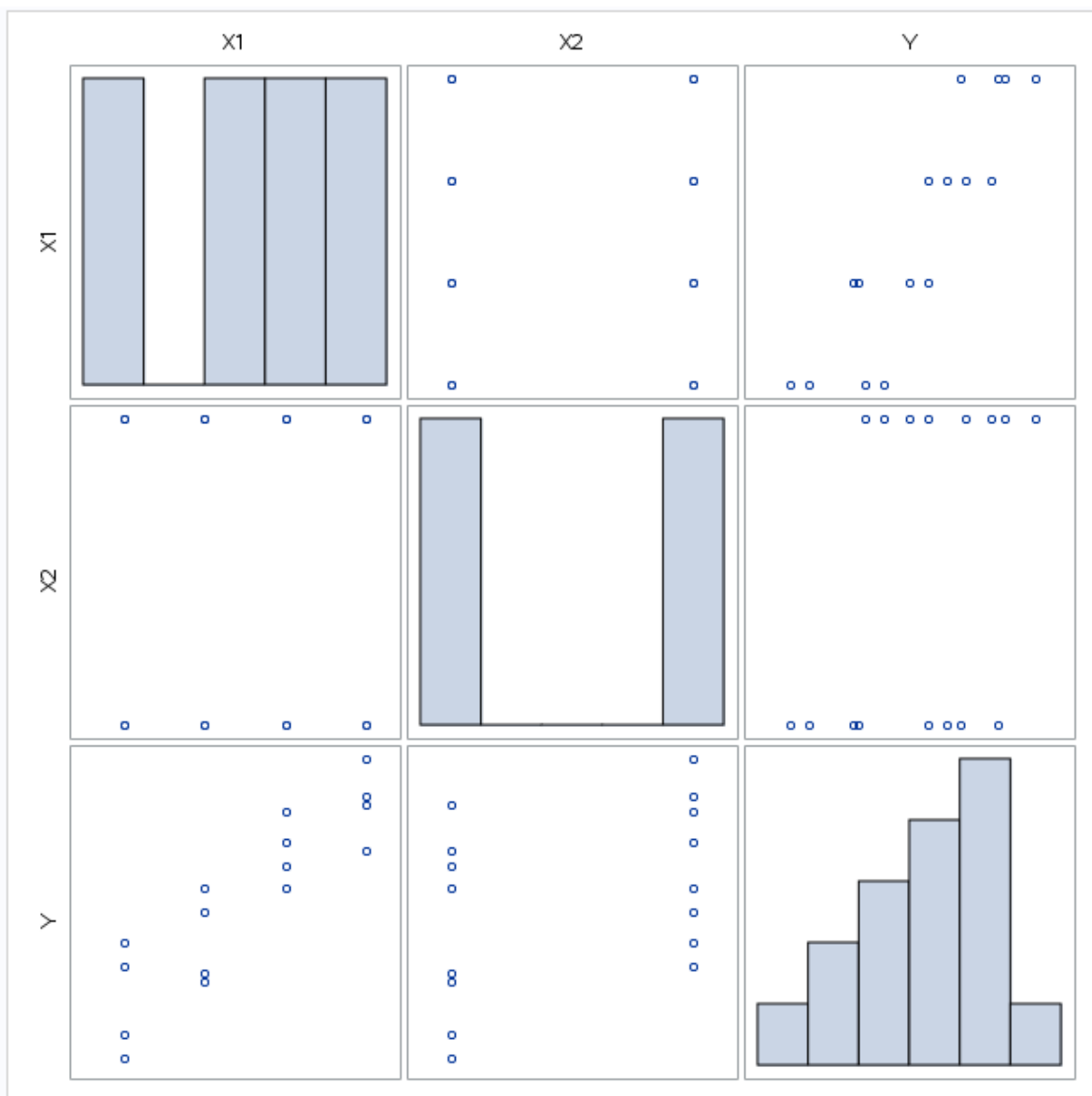
```
PROC SGSCATTER DATA=reg.lista2;  
  MATRIX x1 x2 y / DIAGONAL=(HISTOGRAM);  
RUN;
```

The CORR Procedure

3 Variables: Y X1 X2

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Y	16	81.75000	11.45135	1308	61.00000	100.00000	Y
X1	16	7.00000	2.30940	112.00000	4.00000	10.00000	X1
X2	16	3.00000	1.03280	48.00000	2.00000	4.00000	X2

Pearson Correlation Coefficients, N = 16 Prob > r under H0: Rho=0			
	Y	X1	X2
Y	1.00000	0.89239	0.39458
Y		<.0001	0.1304
X1	0.89239	1.00000	0.00000
X1	<.0001		1.0000
X2	0.39458	0.00000	1.00000
X2	0.1304	1.0000	



b. Fit regression model (6.1) to the data. State the estimated regression function. How is b_1 interpreted here?

```
proc reg data=reg.lista2a;
  model y=x1 x2 /covb;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: Y Y

Number of Observations Read	17
Number of Observations Used	16
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1872.70000	936.35000	129.08	<.0001
Error	13	94.30000	7.25385		
Corrected Total	15	1967.00000			

Root MSE	2.69330	R-Square	0.9521
Dependent Mean	81.75000	Adj R-Sq	0.9447
Coeff Var	3.29455		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	37.65000	2.99610	12.57	<.0001
X1	X1	1	4.42500	0.30112	14.70	<.0001
X2	X2	1	4.37500	0.67332	6.50	<.0001

Covariance of Estimates				
Variable	Label	Intercept	X1	X2
Intercept	Intercept	8.9766346154	-0.634711538	-1.360096154
X1	X1	-0.634711538	0.0906730769	0
X2	X2	-1.360096154	0	0.4533653846

$$Y_i = 37.65 + 4.425 * X1 + 4.375$$

B0 ⇒ preferência pela marca independente do valor do teor de umidade ou doçura

B1 ⇒ mantendo constante todas as variáveis quanto a preferência pela marca aumenta se aumentamos o teor de umidade em uma unidade

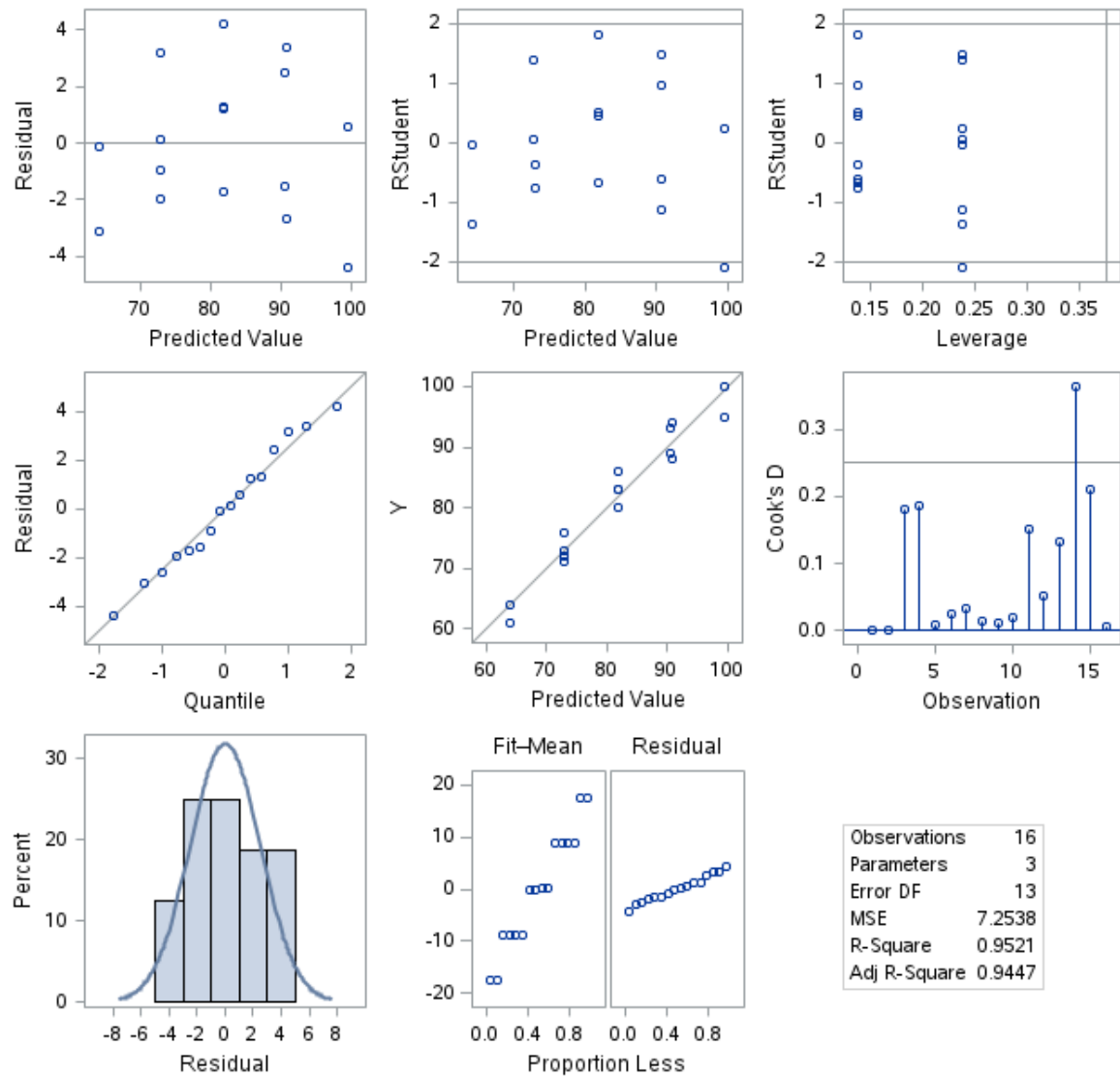
B2 ⇒ mantendo constante todas as variáveis quanto a preferência pela marca aumenta se aumentamos a doçura em uma unidade

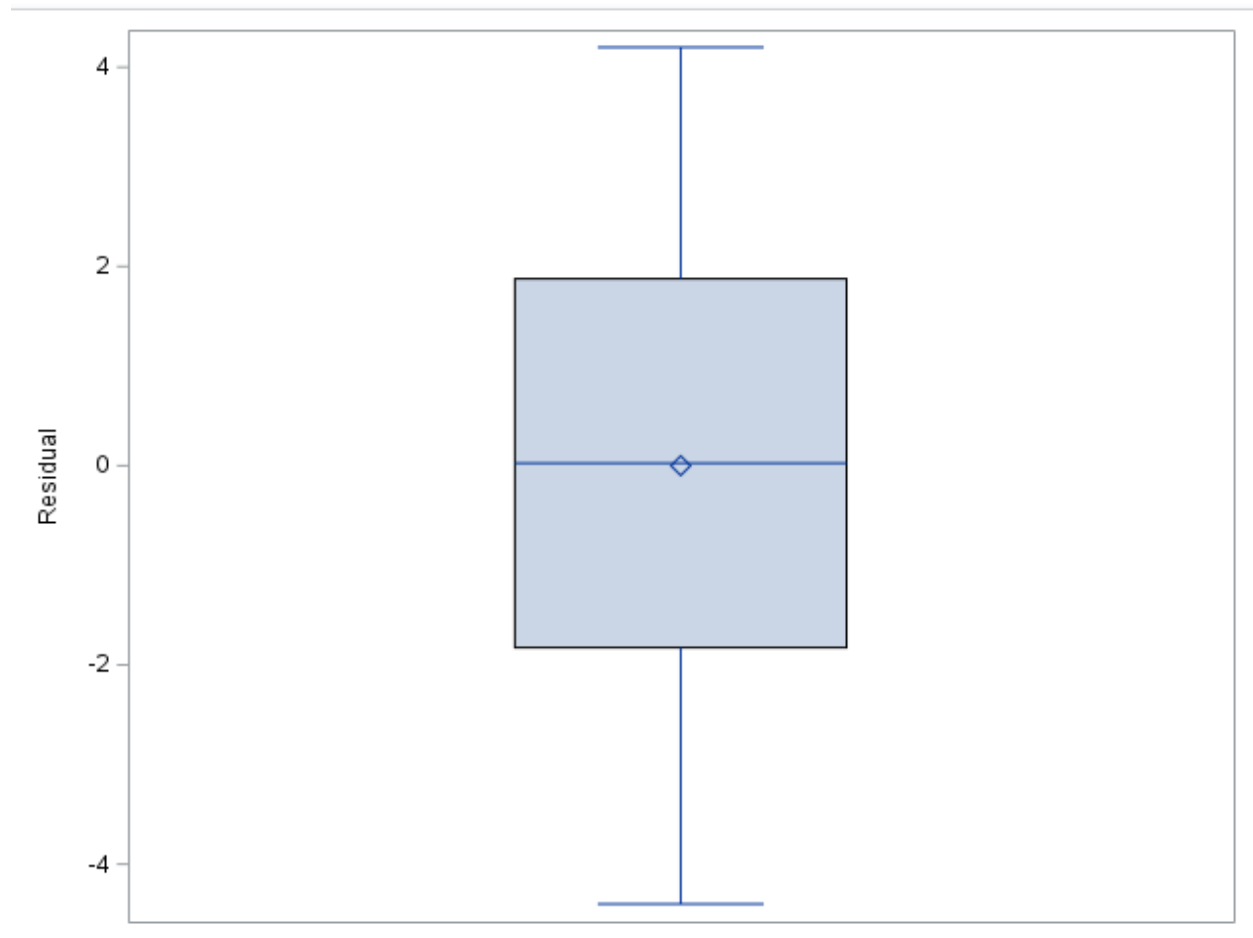
c. Obtain the residuals and prepare a box plot of the residuals. What information does this plot provide?

```
proc reg data=reg.lista2a;  
  model y=x1 x2;  
  output out=reg.residuos_lista2a p=predicted r=residual;  
run;
```

```
proc sgplot data=reg.residuos_lista2a;  
  vbox residual;  
run;
```

Fit Diagnostics for Y





Os resíduos estão distribuídos normalmente. Existe um outlier em relação ao eixo y e apenas um observação que parece estar influenciando as outras (Cook's D)

d. Plot the residuals against \hat{Y} , X_1 , X_2 , and $X_1 X_2$ on separate graphs. Also prepare a normal probability plot. Interpret the plots and summarize your findings.

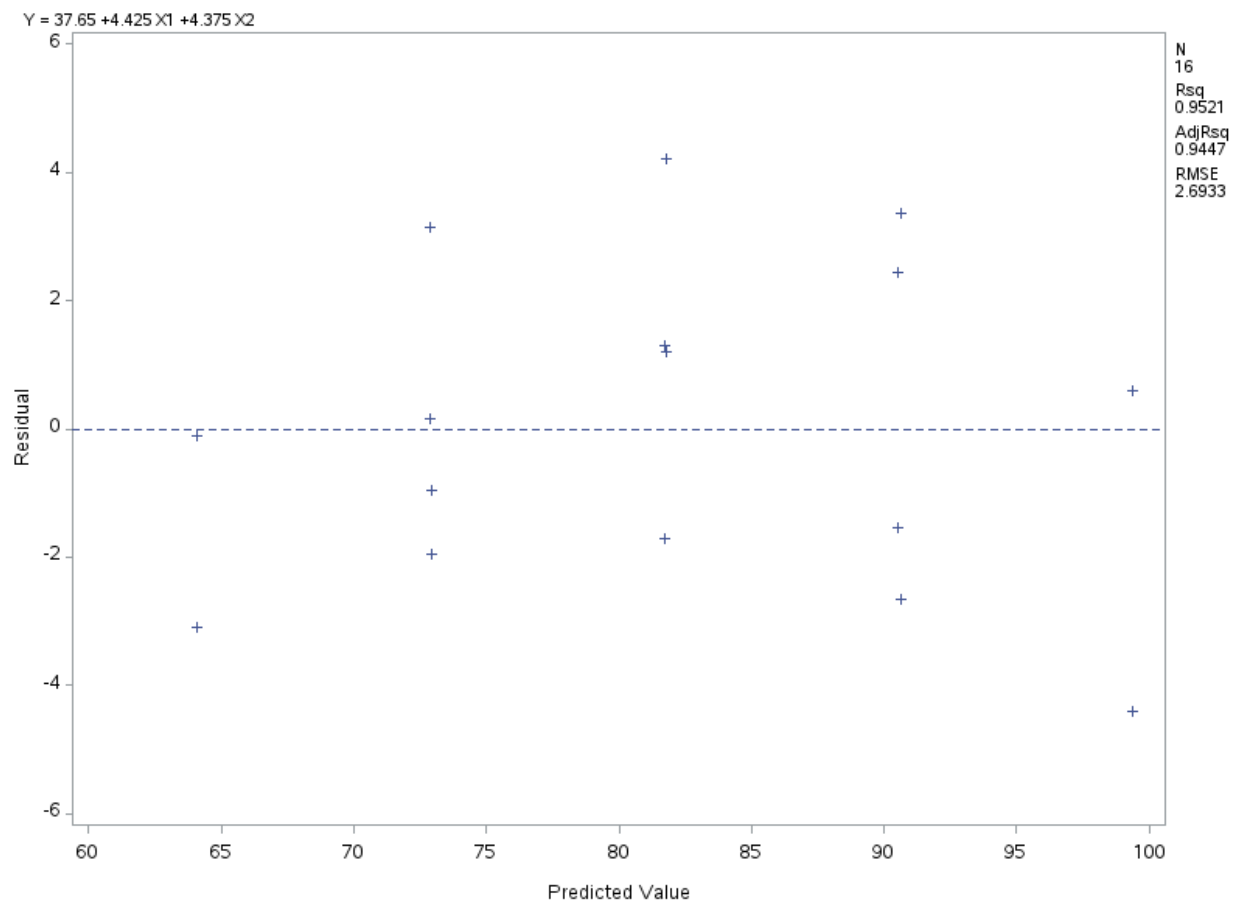
Faltou resíduos contra $x_1 \cdot x_2$ e o de probabilidade normal.

```
proc reg data=reg.lista2a;
  model y=x1 x2;
  output out=reg.residuos_lista2a p=predicted r=residual;
run;
```

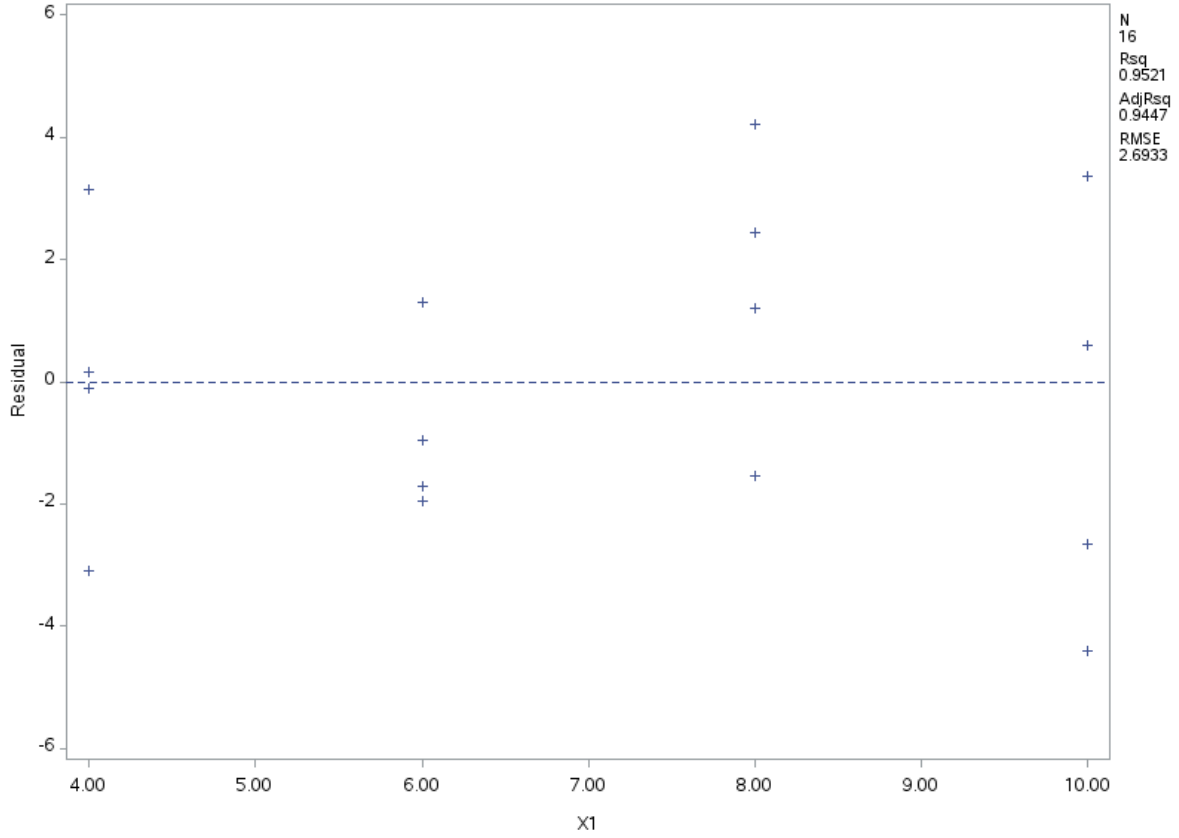
```
proc sgplot data=reg.residuos_lista2a;
  vbox residual;
run;
```

```
proc sgplot data=reg.residuos_lista2a;
  scatter y=residual x=x1x2;
run;
```

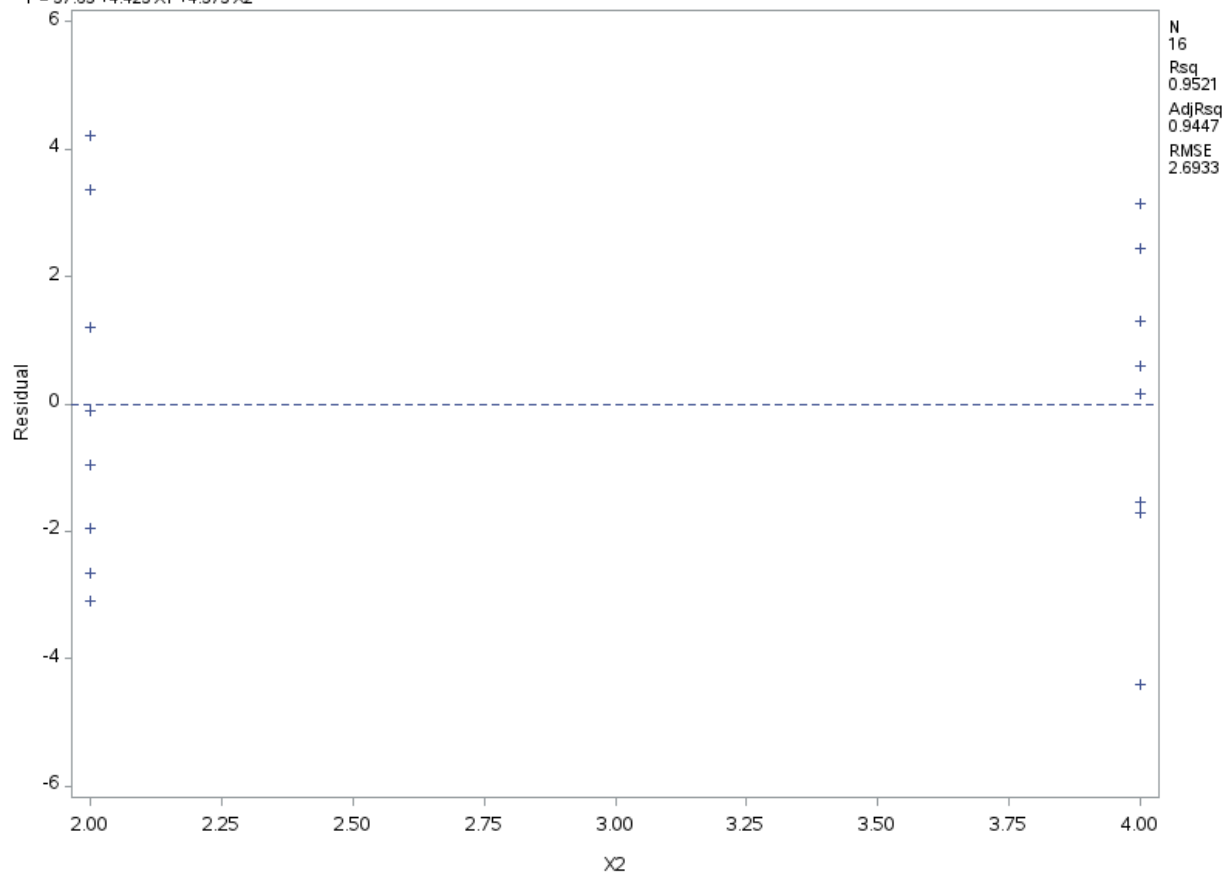
```
proc univariate data=reg.residuos_lista2a normal;
  var residual;
  qqplot residual / normal(mu=est sigma=est);
run;
```

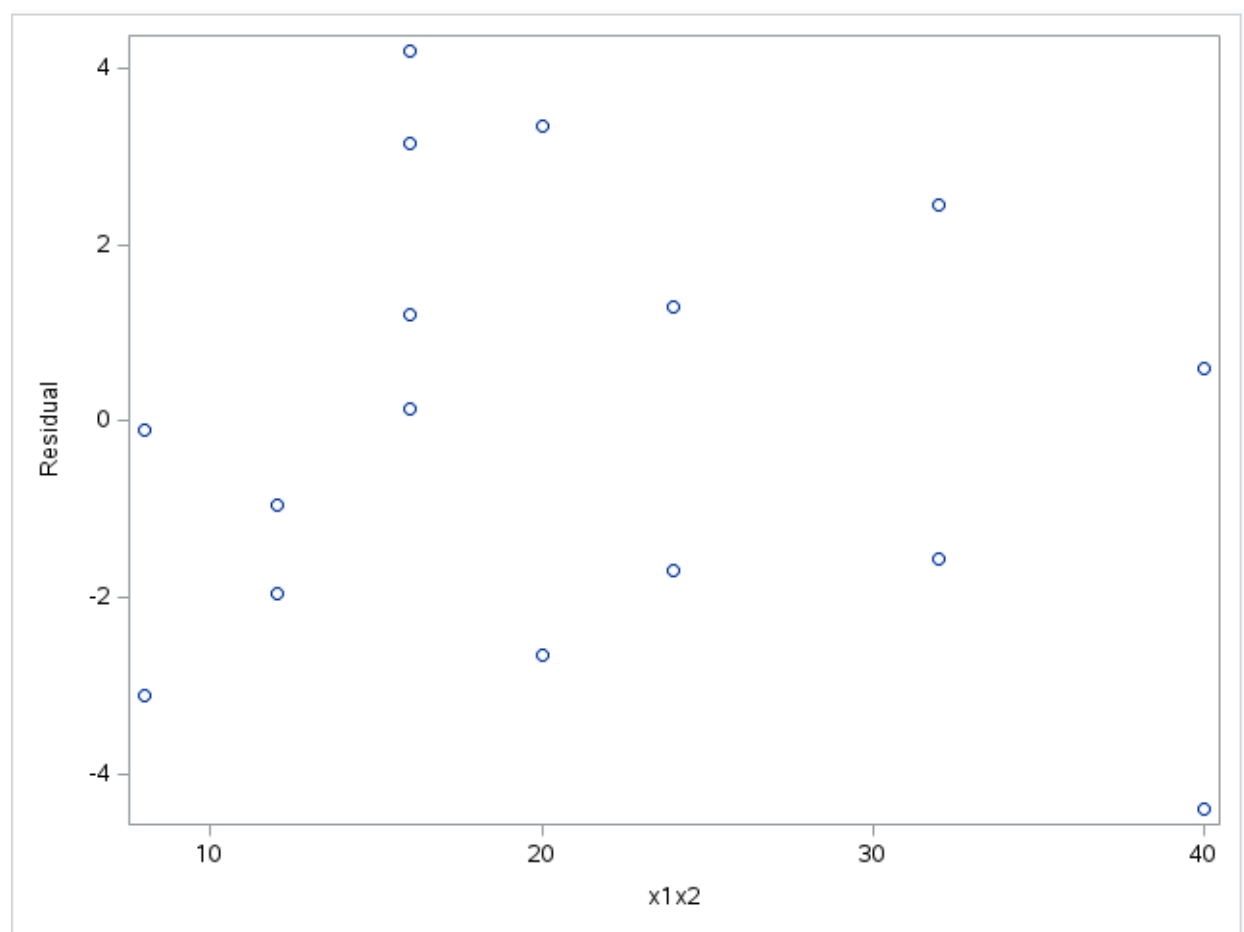


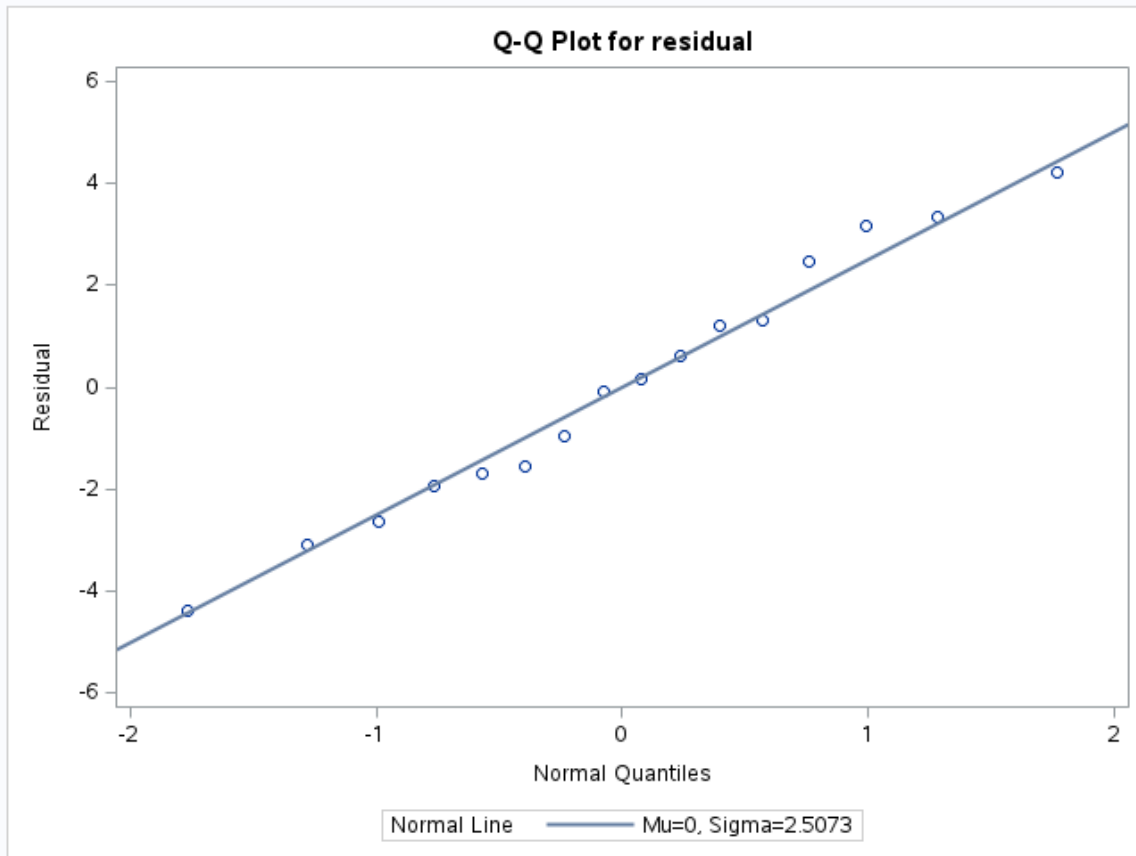
$$Y = 37.65 + 4.425 X_1 + 4.375 X_2$$



$$Y = 37.65 + 4.425 X_1 + 4.375 X_2$$







e. Conduct the Breusch-Pagan test for constancy of the error variance, assuming $\log \sigma_i^2 = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2}$; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

O Teste de Breusch-Pagan é um teste estatístico que verifica a presença de heteroscedasticidade em um modelo de regressão. Heteroscedasticidade significa que a variância dos erros ou resíduos de um modelo de regressão não é constante. Em um modelo ideal, gostaríamos que os resíduos tivessem variância constante, uma condição conhecida como homoscedasticidade. Se a variância dos resíduos muda em função das variáveis independentes, isso pode causar problemas, como estimativas de coeficientes não confiáveis e testes de hipóteses inválidos.

A hipótese nula do teste de Breusch-Pagan é que os erros (ou resíduos) do modelo de regressão têm variância constante, independentemente do valor das variáveis independentes. Isso é conhecido como homoscedasticidade.

Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
Y	White's Test	2.82	4	0.5880	Cross of all vars

```
proc model data=reg.residuos_lista2a;
  parms b0 b1 b2;
  y = b0 + b1*x1 + b2*x2;
  fit y / white;
  test white; /* este é o teste de Breusch-Pagan */
run;
```

f. Conduct a formal test for lack of fit of the first-order regression function; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

```
proc reg data=reg.lista2a ;
  model y=x1 x2 /r p lackfit covb;
  plot r.*p. r.*(x1 x2);
  plot r.*nqq.;
run;
```

O teste de falta de ajustamento é usado para verificar se um modelo de regressão é apropriado para os dados. Especificamente, ele testa se há um padrão nos resíduos que não é capturado pelo modelo. Por exemplo, se houver um relacionamento não linear entre as variáveis independentes e a variável dependente que não é capturado por um modelo de regressão linear, isso pode resultar em uma falta de ajustamento.

A hipótese nula do teste de falta de ajustamento é que o modelo está adequadamente especificado, ou seja, que o modelo captura corretamente a relação entre as variáveis independentes e a variável dependente.

H0) A média de Y **pode** ser descrita em função dos coeficientes

H1) A média de Y **não pode** ser descrita em função dos coeficientes

The REG Procedure
Model: MODEL1
Dependent Variable: Y Y

Number of Observations Read	17
Number of Observations Used	16
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1872.70000	936.35000	129.08	<.0001
Error	13	94.30000	7.25385		
Lack of Fit	5	37.30000	7.46000	1.05	0.4530
Pure Error	8	57.00000	7.12500		
Corrected Total	15	1967.00000			

No caso em questão não podemos rejeitar H_0 .

6.6. Refer to **Brand preference** Problem 6.5. Assume that regression model (6.1) with independent normal error terms is appropriate.

a. Test whether there is a regression relation, using $\alpha = .01$. State the alternatives, decision rule, and conclusion. What does your test imply about β_1 and β_2 ?

b. What is the P -value of the test in part (a)?

```
proc reg data=reg.lista2a ;
  model y=x1 x2 x1x2 /r p lackfit covb;
  model logy=x1 x2 /r p lackfit covb;
run;
```


The REG Procedure
Model: MODEL1
Dependent Variable: Y Y

Number of Observations Read	17
Number of Observations Used	16
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1892.70000	630.90000	101.90	<.0001
Error	12	74.30000	6.19167		
Lack of Fit	4	17.30000	4.32500	0.61	0.6690
Pure Error	8	57.00000	7.12500		
Corrected Total	15	1967.00000			

Root MSE	2.48831	R-Square	0.9622
Dependent Mean	81.75000	Adj R-Sq	0.9528
Coeff Var	3.04380		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	27.15000	6.46481	4.20	0.0012
X1	X1	1	5.92500	0.87975	6.73	<.0001
X2	X2	1	7.87500	2.04435	3.85	0.0023
x1x2		1	-0.50000	0.27820	-1.80	0.0975

Covariance of Estimates					
Variable	Label	Intercept	X1	X2	x1x2
Intercept	Intercept	41.79375	-5.417708333	-12.538125	1.6253125
X1	X1	-5.417708333	0.7739583333	1.6253125	-0.2321875
X2	X2	-12.538125	1.6253125	4.179375	-0.541770833
x1x2		1.6253125	-0.2321875	-0.541770833	0.0773958333

The REG Procedure
Model: MODEL2
Dependent Variable: logy

Number of Observations Read	17
Number of Observations Used	16
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.29481	0.14740	101.18	<.0001
Error	13	0.01894	0.00146		
Lack of Fit	5	0.01111	0.00222	2.27	0.1446
Pure Error	8	0.00783	0.00097829		
Corrected Total	15	0.31375			

Root MSE	0.03817	R-Square	0.9396
Dependent Mean	4.39408	Adj R-Sq	0.9303
Coeff Var	0.86865		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	3.83865	0.04246	90.40	<.0001
X1	X1	1	0.05525	0.00427	12.95	<.0001
X2	X2	1	0.05621	0.00954	5.89	<.0001

Covariance of Estimates					
Variable	Label	Intercept	X1	X2	
Intercept	Intercept	0.0018029062	-0.000127478	-0.000273168	
X1	X1	-0.000127478	0.0000182112	0	
X2	X2	-0.000273168	0	0.0000910559	


- c. Estimate β_1 and β_2 jointly by the Bonferroni procedure, using a 99 percent family confidence coefficient. Interpret your results.

O método de Bonferroni é comumente usado para fazer ajustes em testes de hipóteses múltiplas, para controlar a taxa de erro do tipo I em toda a família de hipóteses. O método de Bonferroni não é usado para estimar os coeficientes de regressão (betas), mas sim para ajustar os valores-p associados aos coeficientes em um modelo de regressão quando múltiplos testes de hipóteses estão sendo realizados.

O ajuste de Bonferroni corrige o valor-p multiplicando o valor-p original pelo número de testes de hipóteses. Isso torna o critério para a rejeição da hipótese nula mais rigoroso, o que ajuda a controlar a taxa de erros do Tipo I ao realizar múltiplos testes de hipóteses. No entanto, também aumenta a taxa de erros do Tipo II, ou seja, aumenta a chance de não rejeitar a hipótese nula quando ela é falsa.

1. Execute o modelo de regressão e obtenha os valores-p para beta1 e beta2:

```
sas
proc reg data=your_dataset;
    model y = x1 x2;
run;
quit;
```

 Copy code

1. Agora, observe os valores-p associados aos coeficientes de regressão x1 e x2 (vamos chamá-los de p1 e p2).
2. Para aplicar o procedimento de Bonferroni, divida o nível de significância desejado (1% neste caso) pelo número de comparações. Neste caso, temos duas comparações (beta1 e beta2), então nosso novo nível de significância será $0.01/2 = 0.005$.
3. Compare os valores-p obtidos no passo 1 com o novo nível de significância (0.005). Se p1 ou p2 for menor que 0.005, rejeitamos a hipótese nula correspondente de que o coeficiente de regressão é igual a zero.

Então, usando o procedimento de Bonferroni, estamos testando cada coeficiente ao nível de 99.5% de confiança para garantir que, globalmente, ainda temos 99% de confiança em todos os nossos testes.

6.7. Refer to **Brand preference** Problem 6.5.

- Calculate the coefficient of multiple determination R^2 . How is it interpreted here?
- Calculate the coefficient of simple determination R^2 between Y_i and \hat{Y}_i . Does it equal the coefficient of multiple determination in part (a)?

Não sei essa

O R^2 ajustado é uma versão modificada do R^2 que foi ajustada de acordo com o número de preditores no modelo. O R^2 sempre aumentará quando adicionamos mais variáveis independentes ao modelo, mesmo que essas variáveis não melhorem o modelo. O R^2 ajustado, no entanto, aumentará apenas se a nova variável melhorar o modelo mais do que seria esperado por acaso. Ele diminuirá quando uma variável não informativa é adicionada.

A fórmula para o R^2 ajustado é: $1 - [(1 - R^2) * (n - 1) / (n - p - 1)]$

Pegar o R^2 (SSR/SST) do modelo completo e comparar com o R^2 do modelo com apenas X_1 . O R^2 do modelo completo sempre será maior que os R^2 dos modelos com menos variável é por isso que é melhor usar o R^2 ajustado.

6.8. Refer to **Brand preference** Problem 6.5. Assume that regression model (6.1) with independent normal error terms is appropriate.

- Obtain an interval estimate of $E\{Y_h\}$ when $X_{h1} = 5$ and $X_{h2} = 4$. Use a 99 percent confidence coefficient. Interpret your interval estimate.
- Obtain a prediction interval for a new observation $Y_{h(new)}$ when $X_{h1} = 5$ and $X_{h2} = 4$. Use a 99 percent confidence coefficient.

- $IC[E(Y_o)] = [\text{Estimativa } Y_o \pm \text{Strudend}(1-\alpha/2) * \text{Erro Padrão } (Y_o)]$
- $IC[Y_o] = [\text{Estimativa } Y_o \pm \text{Strudend}(1-\alpha/2) * \text{Erro Padrão } (\text{Predição})]$

```
proc iml;
reset print;
```

```
Y = { 64, 73, 61,76,72,80,71,83,83,89,86,93,88,95,94,100};
```

```
X = { 1 4 2, 1 4 2, 1 4 2, 1 4 4, 1 6 2, 1 6 4, 1 6 2, 1 6 4, 1 8 2, 1 8 4,
      1 8 2, 1 8 4, 1 10 2, 1 10 4, 1 10 2, 1 10 4};
```

```
n=nrow(X);
p=ncol(X);
glres = n - p;
```

```

xlinhay= T(X)*Y;
xlinhax = T(X)*X;
Ixlindhax=inv(xlinhax);
H = X * Ixlindhax * T(X);
um = J(n,1);
umlinha = T(um);

```

```

Beta = Ixlindhax * xlinhay;

```

```

SSR = T(Beta)*xlinhay - (T(Y)* um * umlinha* Y)/n; /* (F(xi) - YMédia)^2 */
SSE = T(Y)*Y - T(Beta)*xlinhay; /* (F(xi) - Yi)^2 */
SSTO = T(Y)*Y - (T(Y)*um*umlinha*Y)/n; /* (Yi - YMédia)^2 */

```

```

MSR = SSR/(p-1);
MSE= SSE/(n-p);
MSTO= SSTO/(n-1);

```

```

/* Estimar a matriz de covariâncias do Beta^*/
s2beta = MSE * Ixlindhax;

```

```

/*
    IC usando variáveis explicativas 5 e 4
    Existem duas formas de calcular a variância de Yo
*/
Xo = {1 ,5, 4};

```

```

Yo = T(Xo) * Beta;

```

```

VarYo_v1 = T(Xo) * s2beta * Xo;

```

```

VarYo_v2 = MSE * T(Xo) * Ixlindhax * Xo;

```

```

alpha=0.01;
t0005=tinv((1-alpha/2),n-p);

```

```

LIYoMedia = Yo - t0005*sqrt(VarYo_v2);
LSYoMedia = Yo + t0005*sqrt(VarYo_v2);

```

```

s2Predicao = MSE + VarYo_v1;

```

```

LIYoPontual = Yo - t0005*sqrt(s2Predicao);
LIYoPontual = Yo + t0005*sqrt(s2Predicao);

```


Xo 3 rows 1 col (numeric)

1
5
4

Yo 1 row 1 col (numeric)

77.979167

VarYo_v1 1 row 1 col (numeric)

2.6543784

VarYo_v2 1 row 1 col (numeric)

2.6543784

alpha 1 row 1 col (numeric)

0.01

t0005 1 row 1 col (numeric)

3.0122758

LIYoMedia 1 row 1 col (numeric)

73.071488

LSYoMedia 1 row 1 col (numeric)

82.886846

s2Predicao 1 row 1 col (numeric)

14.78868

LIYoPontual 1 row 1 col (numeric)

66.395143

LIYoPontual 1 row 1 col (numeric)

89.56319