Universidade Federal de Santa Maria – UFSM
Centro de Ciências Naturais e Exatas - CCNE
Curso de Especialização em Estatística e Modelagem Quantitativa - CEEMQ

REGRESSORAS CATEGÓRICAS: VARIÁVEIS *DUMMY*

Tatiane Ribeiro e Rosiéli Ruviaro

Prof.: Drª. Ana Lúcia Souza Silva Mateus
Disciplina: Análise de Correlação e Regressão

Departamento de Estatística
Universidade Federal de Santa Maria

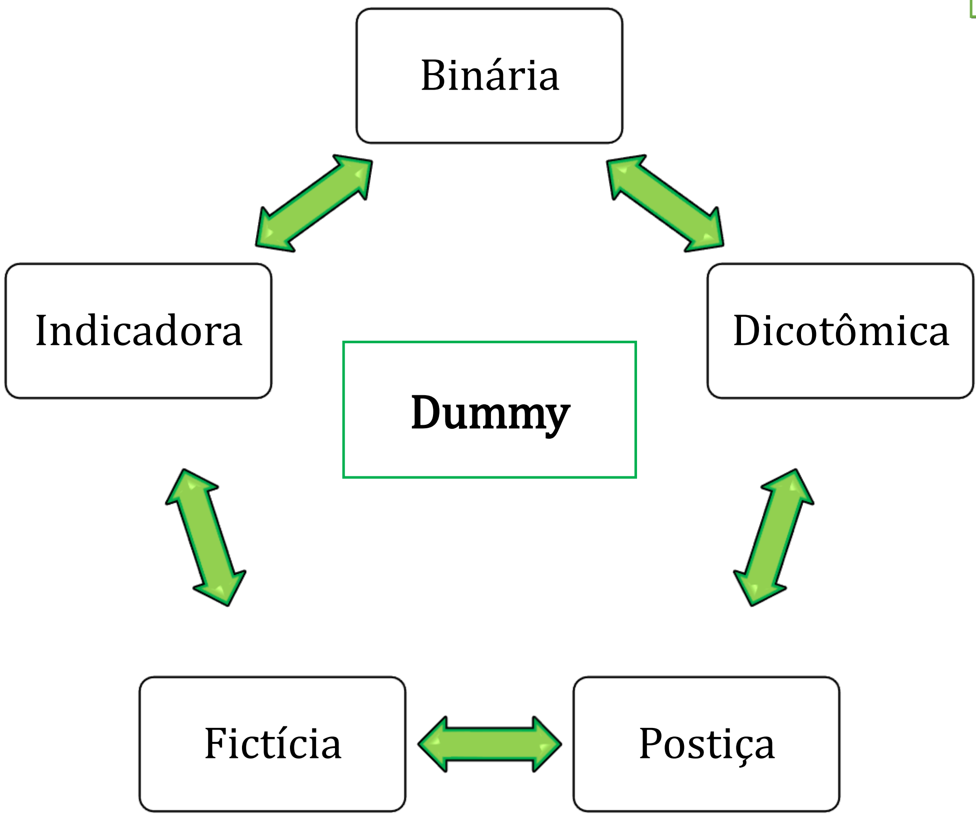
Santa Maria, 13 de dezembro de 2017.

SUMÁRIO

INTRODUZINDO...

Análise de Regressão: Regressoras usualmente **quantitativas.**

Porém, em diversos estudos, não é incomum haver regressoras **qualitativas.**



Variáveis **qualitativas** apresentam apenas **níveis** ou **categorias**. Por exemplo, as variáveis:

- estação do ano;
- turno de trabalho;
- sexo;

e outras.

Variáveis Dummy, portanto, possibilitam quantificar o efeito das categorias de uma regressora qualitativa na resposta Y.

DEFININDO ...

→ Uma das inúmeras formas de identificar quantitativamente as categorias de uma variável qualitativa é a utilização das variáveis indicadoras, que tomam valores 0 ou 1.

→ Esses valores indicam ausência ou presença de determinada característica, respectivamente.

→ Assim, uma variável *dummy*, D , pode ser descrita da seguinte forma:

$$D = \begin{cases} \mathbf{0}, \text{ se a característica não estiver presente} \\ \mathbf{1}, \text{ se a característica estiver presente} \end{cases}$$

Praticando...

→ Podemos usar uma variável em um modelo de regressão quando desejamos incluir certo tipo de variáveis no modelo. Suponha que desejamos estudar os fatores que podem influenciar as notas dos alunos do ensino fundamental.

Qual das variáveis a seguir pode ser representada por uma *dummy?*

(A) Idade do aluno.

(B) Anos de estudo dos pais.

(C) Média de anos de experiência dos professores.

✓ Se o aluno tem computador em casa.

(E) Média das notas do aluno nas séries anteriores.

EXEMPLIFICANDO ...

Experimento → Resposta Y

- X_1: variável contínua
- X_2: variável qualitativa (com dois níveis - 1 e 2, ou categorias)

- Para incorporar a regressora X_2 no modelo de regressão, de modo que o efeito na esperança de Y de cada um de seus níveis possa ser quantificado, define-se a seguinte variável *dummy*:

$$X_2 = \begin{cases} \mathbf{0}, \text{ se nível } 1 \\ \mathbf{1}, \text{ se nível } 2 \end{cases}$$

→ **Em geral**, uma regressora qualitativa com $\boldsymbol{m}$ níveis ou categorias é representada por $\boldsymbol{m-1}$ variáveis indicadoras, cada qual tomando os valores 0 e 1.

EXEMPLO 1) Suponha que um engenheiro mecânico tem por interesse relacionar a vida efetiva de uma ferramenta de corte usada em um torno mecânico com a velocidade do torno em rpm (rotação por minuto) e com o tipo de ferramenta de corte utilizada (tipo A ou B). Os dados coletados estão apresentados a seguir:

Y vida efetiva (horas)	X_1 velocidade (rpm)	X_2 tipo ferramenta	Y vida efetiva (horas)	X_1 velocidade (rpm)	X_2 tipo ferramenta
18.73	610	A	30.16	670	B
14.52	950	A	27.09	770	B
17.43	720	A	25.40	880	B
14.54	840	A	26.05	1000	B
13.44	980	A	33.49	760	B
24.39	530	A	35.62	590	B
13.34	680	A	26.07	910	B
22.71	540	A	36.78	650	B
12.68	890	A	34.95	810	B
19.32	730	A	43.67	500	B

Fonte: Montgomery e Peck (1992)

Analisando e resolvendo ...

X_2 → Tipo de ferramenta ↔ Qualitativa com 2 níveis – A e B

Para quantificar efeito de seus níveis na esperança de Y e incorporá-la a um modelo de regressão, utiliza-se uma variável *dummy*, do que obtém-se:

$$X_2 = \begin{cases} \textbf{0}, \text{ se ferramenta do tipo A.} \\ \textbf{1}, \text{ se ferramenta do tipo B.} \end{cases}$$

Modelo:

$$E(Y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\begin{cases} x_2 = 0 \Rightarrow E(Y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 & \text{A} \\ x_2 = 1 \Rightarrow E(Y|\boldsymbol{x}) = (\beta_0 + \beta_2) + \beta_1 x_1 & \text{B} \end{cases}$$

Relação entre a vida efetiva da ferramenta e a velocidade do torno:

Tipo de ferramenta A ($x_2 = 0$): → Reta
- Intercepto: β_0
- Inclinação: β_1

Tipo de ferramenta B ($x_2 = 1$): → Reta
- Intercepto: $\beta_0 + \beta_2$
- Inclinação: β_1

Duas retas paralelas

→ Para x_1 **fixo**, o parâmetro β_2 expressa a mudança **na esperança do tempo de vida (Y)**, resultante da mudança da ferramenta do tipo **A** para a do tipo **B**.

→ **Intervalo de confiança** e **teste de hipóteses** para β_2 são obtidos de forma **análoga** aos obtidos para os parâmetros de um modelo de regressão com todas as regressoras quantitativas.

Mas e se a regressora qualitativa tiver mais de 2 níveis?

É possível generalizar o uso de uma variável *dummy* para incorporar uma variável regressora qualitativa com mais de 2 níveis. Considere-se, assim, o exemplo:

3 tipos de ferramenta

A
B
C

É necessário 2 variáveis indicadoras para incorporar os 3 níveis no modelo:

$$X_{21} = \begin{cases} \mathbf{1} \text{ se ferramenta do tipo A} \\ \mathbf{0} \text{ em caso contrário} \end{cases} \qquad X_{22} = \begin{cases} \mathbf{1} \text{ se ferramenta do tipo B} \\ \mathbf{0} \text{ em caso contrário,} \end{cases}$$

do que resultam as combinações binárias:

X_{21}	X_{22}	
1	0	Ferramenta tipo A
0	1	Ferramenta tipo B
0	0	Ferramenta tipo C.

Consequentemente, o modelo de regressão será dado por:

$$E(Y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 \boldsymbol{x_{21}} + \beta_3 \boldsymbol{x_{22}}$$

X_{21}	X_{22}	
1	0	$E(Y \mid \mathbf{x}) = (\beta_0 + \beta_2) + \beta_1 x_1$
0	1	$E(Y \mid \mathbf{x}) = (\beta_0 + \beta_3) + \beta_1 x_1$
0	0	$E(Y \mid \mathbf{x}) = \beta_0 + \beta_1 x_1$

***

3 retas paralelas distintas

*** Nesse modelo, para x_1 fixo, o parâmetro β_2 expressa a mudança no tempo de vida médio resultante da mudança da ferramenta do tipo C para a do tipo A e, o parâmetro β_3, a mudança no tempo de vida médio resultante da mudança da ferramenta do tipo C para a do tipo B.

O que fazer se as retas não forem, de fato, paralelas?

[...] a inclusão da interação entre as regressoras permite que esse fato seja analisado. (GIOLO, 2007, p. 41)

MRLM que relaciona: tempo de vida das ferramentas X velocidade do torno e tipo de ferramenta (A e B)

e **produz**

Inclinações e interceptos diferentes.

Modelando essa situação tem-se:

$$E(Y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \boldsymbol{x_1 x_2}.$$

Observe que, como : $X_2 \begin{cases} \mathbf{0} \\ \text{ou} \\ \mathbf{1} \end{cases}$ tem-se:

$\boldsymbol{X_2 = 0}$

$$E(Y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \cancel{\beta_2 x_2}^{\,0} + \cancel{\beta_3 x_1 x_2}^{\,0}$$

$$\boxed{E(Y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1}$$

$\boldsymbol{X_2 = 1}$

$$E(Y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

$$E(Y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 \cdot 1 + \beta_3 x_1 \cdot 1$$

$$\boxed{E(Y|\boldsymbol{x}) = (\beta_0 + \boldsymbol{\beta_2}) + (\beta_1 + \boldsymbol{\beta_3})\boldsymbol{x_1}}$$

Retas concorrentes

→ Nesse caso, o efeito na esperança de Y resultante da mudança da ferramenta do tipo A para a do tipo B é de $\boldsymbol{\beta_2 + \beta_3 x_1}$ unidades.

Mas por que não alocar códigos quaisquer às categorias das variáveis qualitativas simplesmente?

Códigos definem uma métrica para os níveis da variável qualitativa, métrica esta que pode não ser razoável.

Alocar arbitrariamente os códigos 1, 2 e 3 para uma regressora, por exemplo, qualitativa com três níveis, implica em assumir que a resposta muda, em média, a mesma quantidade quando se passa de um nível para outro e, isso, pode **não** estar de acordo com a realidade.

Portanto, alocar códigos, igualmente ou não espaçados, aos níveis de uma regressora qualitativa equivale a assumir distâncias arbitrárias, porém definidas, entre os níveis.

Variáveis indicadoras, em contraste, não impõem qualquer métrica aos níveis da variável qualitativa.

Elas dependem dos dados para mostrar os efeitos diferenciais que ocorrem entre os níveis.

Exemplo: Suponha um experimento em que tenha Y, a variável resposta (quantitativa) e as regressoras X_1 (quantitativa) e X_2 (qualitativa com três níveis: A, B e C). Se for considerado os códigos 1, 2 e 3 aos níveis de X_2 tem-se que:

$$X_2 = \begin{cases} \mathbf{1} \text{ se nível A} \\ \mathbf{2} \text{ se nível B} \\ \mathbf{3} \text{ se nível C} \end{cases}$$

$$E(Y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$x_2 = 1 \ (A) \Rightarrow E(Y|\boldsymbol{x}) = (\beta_0 + \beta_2) + \beta_1 x_1$$
$$x_2 = 2 \ (B) \Rightarrow E(Y|\boldsymbol{x}) = (\beta_0 + 2\beta_2) + \beta_1 x_1$$
$$x_2 = 3 \ (C) \Rightarrow E(Y|\boldsymbol{x}) = (\beta_0 + 3\beta_2) + \beta_1 x_1$$

P/ x_1 fixo: A → B, a mudança na resposta média será de β_2 unidades. Analogamente para B → C. ∴ usar os códigos 1, 2 e 3, impõe que a resposta média muda a mesma quantidade quando se muda do nível A → B ou B → C. Isto pode **não estar de acordo com a realidade.**

Podem ser utilizados outros códigos para variáveis *dummy*, além do binário?

Sim, é possível. Outros esquemas possíveis são apresentados a seguir:

1º esquema: $D = \begin{cases} 1 \\ -1 \end{cases}$

- Var. qual. c/ **2** níveis

$$X_2 = \begin{cases} 1 \text{ se nível A} \\ -1 \text{ se nível B} \end{cases}$$

- Var. qual. c/ **3** níveis

$$X_{21} = \begin{cases} 1 \text{ se A} \\ -1 \text{ se C} \\ 0 \text{ c. c.} \end{cases} \qquad X_{22} = \begin{cases} 1 \text{ se B} \\ -1 \text{ se C} \\ 0 \text{ c. c.} \end{cases}$$

Nível	X_{21}	X_{22}
A	1	0
B	0	1
C	−1	−1

2º esquema: $D = \begin{cases} 0 \\ 1 \end{cases}$

- Sem intercepto e $m = D$

$$E(Y|\boldsymbol{x}) = \beta_1 x_1 + \beta_2 x_{21} + \beta_3 x_{22}$$

$$X_{21} = \begin{cases} 1 \text{ se A} \\ 0 \text{ c.c.} \end{cases}$$

$$X_{22} = \begin{cases} 1 \text{ se B} \\ 0 \text{ c.c.} \end{cases}$$

Capítulo 9 – Modelos de Regressão com Variáveis binárias

2- Cuidados no uso de modelos com variáveis qualitativas

Suponha que desejamos inserir no modelo uma variável qualitativa com m categorias. É importante notar que o modelo deverá ser especificado da seguinte forma:

1) Modelo com termo constante e (m-1) variáveis dummy
2) Modelo SEM termo constante e m variáveis dummy

Por que? Poderia parecer natural, à primeira vista, escrever o modelo como:

$$Y_i = \alpha + \beta_1.D_{1i} + \beta_2.D_{2i} + \ldots + \beta_m.D_{mi} + \varepsilon_i \qquad (1)$$

onde D_j = 1 ou 0 se a observação pertence à j-ésima categoria da variável X, para j = 1, 2,..., n.

Mas, qual a matriz de design X para o modelo da equação (1) acima? VERIFIQUE que a primeira coluna de X é composta apenas de 1's. A segunda coluna contém 1's e 0's, assim como todas as demais colunas. Mas, a soma das colunas 2 até m+1 é igual à 1ª. coluna, pois somando todas as variáveis dummy em todos os seus níveis encontramos uma coluna de 1's.

Logo, o modelo representado por (1) NÃO PODE ser ajustado, pois sua matriz de design exibe colinearidade perfeita. As alternativas são as indicadas em 1) e 2) acima.

NOTA:

$$R^2_{(0)} \nLeftrightarrow R^2$$

Dispersão em torno de zero

Dispersão em torno da $\bar{y}$

Tendência de $R^2_{(0)} > R^2$

A consequência é uma grave confusão quando, erroneamente, modelos sem interceptos são comparados com aqueles com interceptos. Os modelos sem intercepto serão classificados, erroneamente, em um *rank* de diversos modelos sendo comparados como os melhores.

$$R^2_{(0)} = 1 - \left[\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \; / \; \sum_{i=1}^{n} (y_i - \bar{y})^2 \right]$$

$$\sum_{i=1}^{n} (y_1 - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i)^2 - \left[\left(\sum_{i=1}^{n} (y_i - x_i) \right)^2 \; / \; \sum_{i=1}^{n} (x_i)^2 \right]$$

Exemplo 9.1.

Neste exemplo a variável dependente é o percentual de votos nulos e brancos (soma dos dois percentuais) no 1º. Turno das eleições municipais para prefeito no município do Rio de Janeiro em 2008. A variável explicativa é a região da cidade em que está situada a seção eleitoral, dividida em 5 categorias: Centro, Sul, Norte, Oeste, Subúrbio.

Existem 10702 observações na amostra, cada uma corresponde a uma seção eleitoral, ou seja, uma urna de votação.

Inicialmente ajustamos o modelo:

$$Y_i = \alpha + \beta_1.CENTRO_i + \beta_2.NORTE_i + \beta_3.SUL_i + \beta_4.OESTE_i + \varepsilon_i$$

Ou seja, a categoria base é "SUBURBIO". Os resultados desta regressão estão abaixo:

Coefficients(a)

Model	Unstandardized Coefficients		t	Sig.
	B	Std. Error	B	Std. Error
(Constant)	13,405	,037	360,889	,000
indicador centro da cidade	,696	,125	5,565	,000
indicador zona norte	-2,423	,096	-25,262	,000
indicador zona sul	-3,529	,073	-48,395	,000
indicador zona oeste	-,190	,055	-3,479	,001

a Dependent Variable: soma dos percentuais de nulos e brancos

- No SUBURBIO (categoria base, variável dummy omitida), o percentual médio de votos brancos + nulos é 13,405%, que é superior à média geral do município (12,696%).

- No CENTRO, o percentual médio de brancos + nulos é 13,405 + 0,696 = 14, 101%.

- Na zona NORTE, o percentual médio de brancos + nulos é 13,405 - 2,423= 10,982%, o SEGUNDO MENOR de todas as regiões da cidade.

- Na zona SUL, o percentual médio de brancos + nulos é 13,405 - 3,529= 9,876%, o MENOR de todas as regiões da cidade.

- Na zona OESTE, o percentual médio de brancos + nulos é 13,405 - 0,190= 13,215%, superior à média geral do município.

$$Y_i = \alpha + \beta_1.CENTRO_i + \beta_2.NORTE_i + \beta_3.SUL_i + \beta_4.OESTE_i + \varepsilon_i$$

perc_nulos_bra

Model	Unstandardized Coefficients		t	Sig.
	B	Std. Error	B	Std. Error
(Constant)	13,405	,037	360,889	,000
indicador centro da cidade	,696	,125	5,565	,000
indicador zona norte	-2,423	,096	-25,262	,000
indicador zona sul	-3,529	,073	-48,395	,000
indicador zona oeste	-,190	,055	-3,479	,001

a Dependent Variable: soma dos percentuais de nulos e brancos

Para comparação ajustamos o modelo que emprega todas as categorias da variável "região da cidade" e não contém termo constante.

$$Y_i = \beta_1.CENTRO_i + \beta_2.NORTE_i + \beta_3.SUL_i + \beta_4.OESTE_i + \beta_5.SUBURBIO_i + \varepsilon_i$$

Os resultados seguem abaixo.

ANOVA(c,d)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1743244,336	5	348648,867	58573,634	,000(a)
	Residual	63671,940	10697	5,952		
	Total	1806916,276(b)	10702			

Coefficients(a,b)

Model	Unstandardized Coefficients		t	Sig.
	B	Std. Error	B	Std. Error
indicador centro da cidade	14,102	,119	118,030	,000
indicador zona norte	10,982	,088	124,174	,000
indicador zona sul	9,877	,063	157,413	,000
indicador zona oeste	13,215	,040	329,390	,000
indicador suburbio	13,405	,037	360,889	,000

NOTA: Coef. da regressão estimada são os mesmo do modelo anterior.

ANÁLISE RESIDUAL PARA OBSERVAÇÕES INFLUENTES EM ML COM A VARIÁVEL *DUMMY*

[...] mesmos critérios empregados em regressões lineares para identificação de pontos influentes (*outlier*). (FIGUEIREDO, 2005, p. 16).

Elementos da diagonal principal da matriz H; → $\begin{bmatrix} h_{11} & - & - \\ - & h_{22} & - \\ - & - & h_{33} \end{bmatrix}$ → P. A. se $h_{ii} > 2 \cdot \frac{p}{n}$

$\boxed{\mathbf{H} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}}$

*
*
*

DFFITS; → $DFFITS_i = \left(\sqrt{\frac{h_{ii}}{1-h_{ii}}}\right) \cdot \left(\frac{e_i}{\sqrt{S_{(i)}^2}}\right)$ → $DIFFITS_i = \left(2\sqrt{\frac{p+1}{n}}\right)$

Cook's D.; → $D_i = \left(\frac{1}{p}\right) \cdot \left(\frac{h_{ii}}{1-h_{ii}}\right) \cdot \left(\frac{e_i^2}{S^2(1-h_{ii})}\right)$ → $D_i > 2 \cdot h_{ii_{(médio)}}$

SELEÇÃO DA EQUAÇÃO DE REGRESSÃO

- A avaliação da análise da variância consiste no primeiro passo no processo de seleção, juntamente com a observação do coeficiente de determinação para modelos não centrados na média (R^2) e do coeficiente de determinação ajustado ($R^2_{Ajustado}$), em que:

$$R^2 = \left[1 - \left(\frac{SQ_{Resíduo}}{SQ_{Total}}\right)\right]$$

preferível →

$$R^2_{Ajustado} = \left[1 - \frac{(1 - R^2) \cdot (n - 1)}{n - p - 1}\right]$$

- O erro padrão da estimativa ou erro padrão residual (S_{yx}) mede a dispersão média entre os valores observados e estimados ao longo da linha da regressão.

$$S_{yx} = \sqrt{\frac{SQ_{Res}}{(n - p - 1)}}$$

→ Critério PRESSp (soma de quadrados de predição): medida que avalia a qualidade dos valores ajustados por meio do modelo em questão;
→ é uma importante ferramenta para a seleção de modelos lineares.

$$PRESS_{p=} \sum_{i=1}^{n} \left(y_i - \hat{y}_{1(i)}\right)^2$$

∴ as melhores equações ajustadas são aquelas que apresentam valores pequenos de PRESSp (têm menor erro de predição).

DIAGNÓSTICO DE NORMALIDADE PARA EQUAÇÕES COM A VARIÁVEL *DUMMY*

→ Verificar se os resíduos seguem, aproximadamente, uma distribuição normal para equações com a variável *Dummy* é necessário fragmentar a equação geral.

Resíduo padronizado

$$e_{padronizado} = \frac{Y_i - \hat{Y}}{\sqrt{\frac{SQ_{Res}}{(n-p-1)}}}$$

CONTINUAÇÃO RESOLUÇÃO EXEMPLO 1: Análise de resíduos e pontos influentes

1ª ANÁLISE) Variável *dummy:* $X_2 = \begin{cases} 0 \\ 1 \end{cases}$. Considera-se o modelo:

$$E(Y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2, \qquad sendo:$$

$$X_2 = \begin{cases} 0 \; se \; ferramenta \; A \\ 1 \; se \; ferramenta \; B \end{cases}$$

Análise de variância do modelo com a interação entre X_1 e X_2.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
X1	1	293.01	293.01	33.2545	2.889e-05	β_1 sign.
X2B	1	1125.03	1125.03	127.6847	4.891e-09	β_2 sign.
X1:X2	1	16.08	16.08	1.8248	0.1955	β_3 não sign.
Residuals	16	140.98	8.81			

Análise de variância do modelo SEM a interação entre X_1 e X_2.

```
          Df  Sum Sq Mean Sq F value    Pr(>F)
X1         1  293.01  293.01  31.716 2.990e-05
X2B        1 1125.03 1125.03 121.776 3.587e-09
Residuals 17  157.05    9.24
```

β_1 sign.

β_2 sign.

Estimativas dos coeficientes e outros resultados relevantes

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.98560    3.51038  10.536 7.16e-09
X1          -0.02661    0.00452  -5.887 1.79e-05
X2B         15.00425    1.35967  11.035 3.59e-09

Residual standard error: 3.039 on 17 degrees of freedom
Multiple R-Squared: 0.9003,     Adjusted R-squared: 0.8886
F-statistic: 76.75 on 2 and 17 DF,  p-value: 3.086e-09
```

β_0 sign.

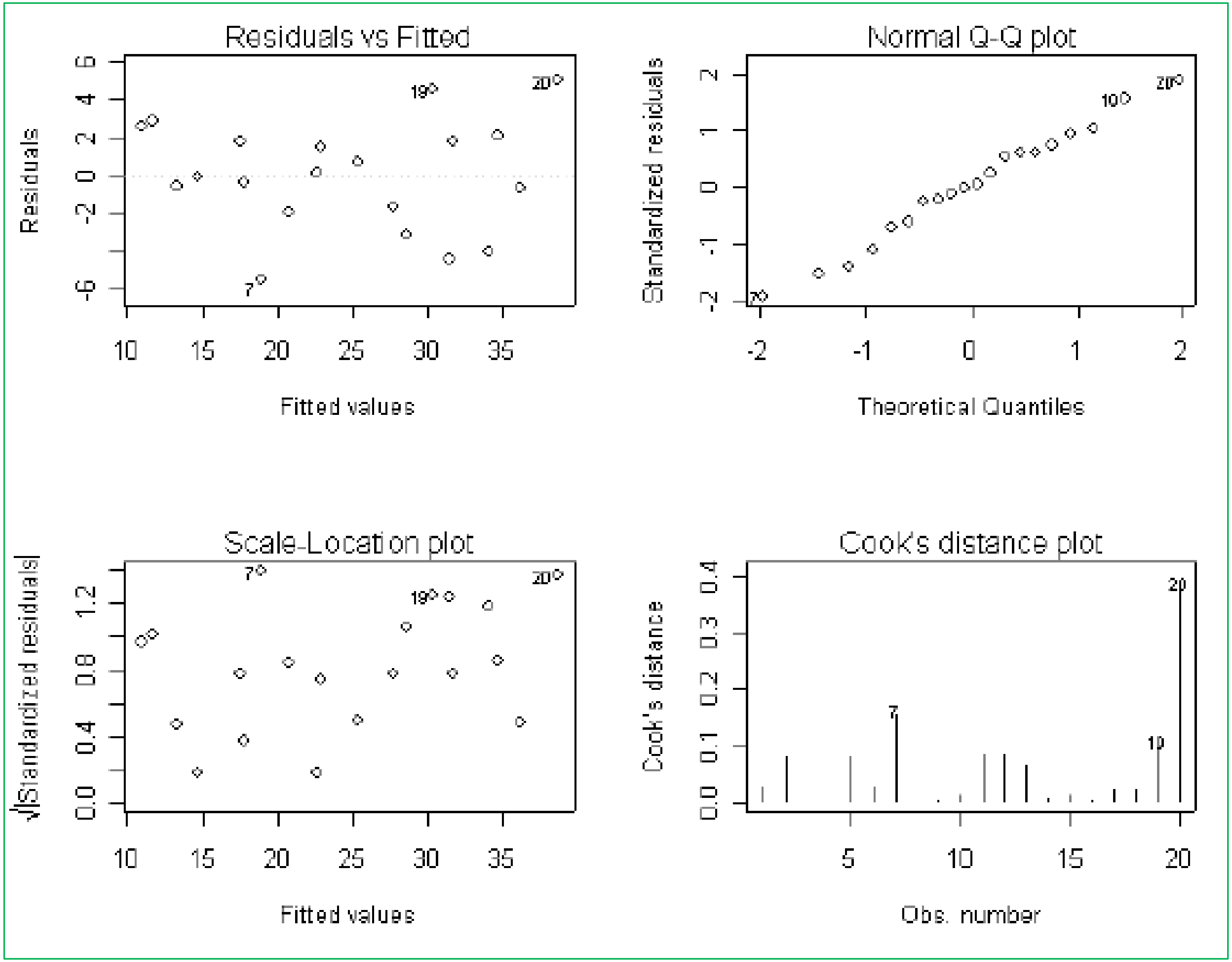
β_1 sign.

β_2 sign.

ANÁLISE DE VARIÂNCIA SATISFATÓRIA

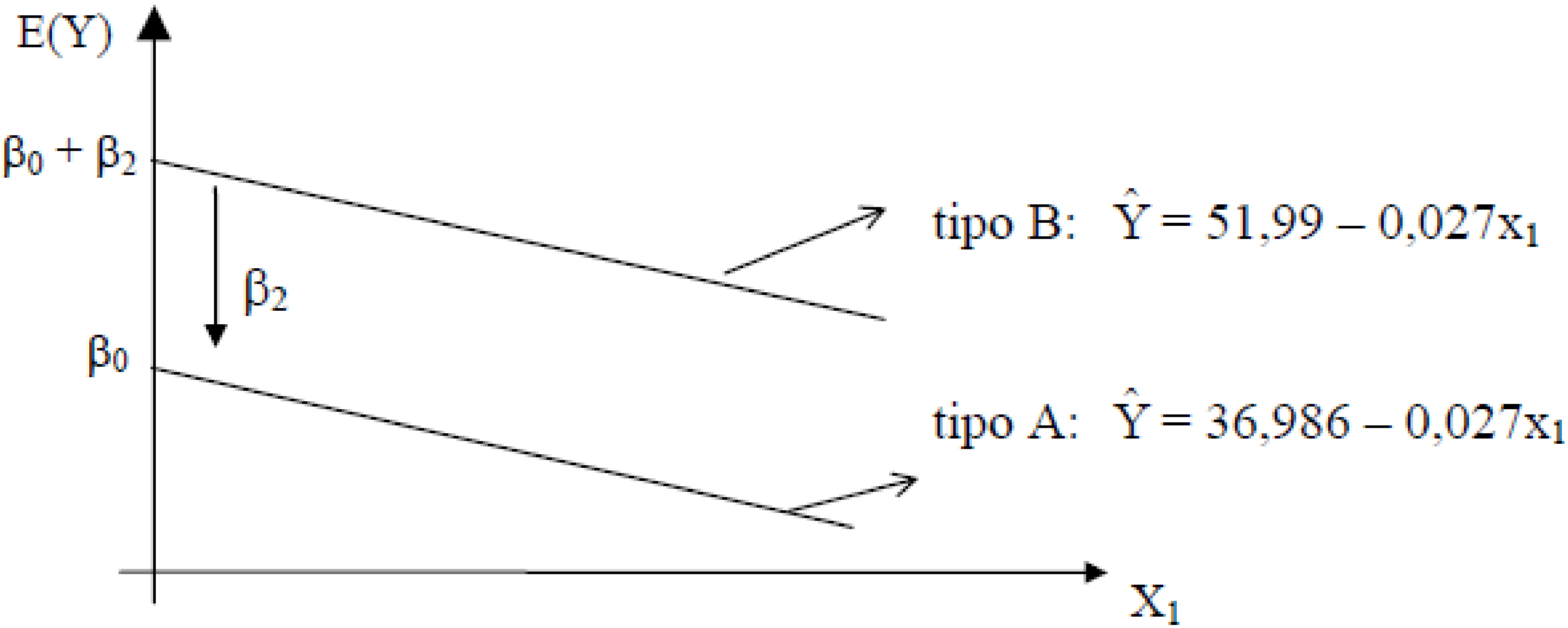
```r
rm(list=ls())

dados<- data.frame(
  y=c(18.73,14.52,17.43,14.54,13.44,24.39,13.34,22.71,12.68,19.32,30.16,27.09,25.40,26.05,33.49,35.62,26.07,36.7
  x1=c(610,950,720,840,980,530,680,540,890,730,670,770,880,1000,760,590,910,650,810,500),
  x2=c(0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1)
)

attach(dados)

m1<-lm(y~x1+x2)
m1
anova(m1)
summary(m1)
```

```
    Min      1Q  Median      3Q     Max
-5.5527 -1.7868 -0.0016  1.8395  4.9838

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 36.98560    3.51038  10.536 7.16e-09 ***
x1          -0.02661    0.00452  -5.887 1.79e-05 ***
x2          15.00425    1.35967  11.035 3.59e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.039 on 17 degrees of freedom
Multiple R-squared:  0.9003,	Adjusted R-squared:  0.8886 
F-statistic: 76.75 on 2 and 17 DF,  p-value: 3.086e-09

> su
```



ANÁLISE DE RESÍDUOS SATISFATÓRIA

Modelo ajustado: $\hat{Y} = 36{,}986 - 0{,}027x_1 + 15{,}004x_2$



2ª ANÁLISE) Variável *dummy:* $D = X_2 = \begin{cases} 1 \\ -1 \end{cases} \rightarrow X_2 = \begin{cases} 1\ se\ ferramenta\ A \\ -1\ se\ ferramenta\ B \end{cases}$

Interação não foi significativa e o resultado sem interação foi análogo à 1ª.

3ª ANÁLISE) Variável *dummy:* $D{=}X_2 = \begin{cases} 0 \\ 1 \end{cases}$; Modelo sem intercepto.

$$X_{21} = \begin{cases} 1\ se\ ferramenta\ A \\ 0\ c.c. \end{cases} \qquad X_{22} = \begin{cases} 1\ se\ ferramenta\ B \\ 0\ c.c. \end{cases}$$

Interação não foi significativa e o resultado sem interação foi melhor.

$$R^2_{(3^a\ análise)} > R^2_{(1^a\ 2^a\ análise)}$$

$$\hat{Y} = -0.027x_1 + 36{,}985\ x_{21} + 51{,}989\ x_{22}$$

Note, para uma velocidade fixa x_1 do torno, que a diferença (51,989 – 36,985) = 15,004, representa, como nas análises anteriores, o quanto a esperança de Y muda ao se mudar da ferramenta A para a B.

CONCLUINDO ...

É preferível ajustar um único modelo pelas seguintes razões:

→ O analista tem somente uma equação final (mais prático);

→ O ajuste de um único modelo produz uma estimativa comum da variância dos erros, se tem mais graus de liberdade do que se teria no ajuste de dois ou mais modelos de regressão lineares separados.

→ O procedimento de construção de modelos completos pode envolver duas ou mais variáveis de interesse;

→ As estatísticas de análise de resíduos e normalidade seguem **os mesmos** procedimentos matemáticos que os modelos múltiplos.

REFERÊNCIAS

FIGUEIREDO E. O., **Método da Variável Fictícia para Ajuste de Modelos Volumétricos Estáveis e Compatíveis em Povoamentos Florestais**. Rio Branco - AC, 2005.

GIOLO S. R., **Análise de Regressão Linear.** Universidade Federal Do Paraná Departamento De Estatística. Curitiba, Paraná, 2007.

MODELO LINEAR V. Aula 10. Disponível em: <http://hedibert.org/wp content/uploads/2014/03/Econometria201401-Aula09-ARLM-VI-Dummy.pdf>. Acesso em: 09. dez 2017.

BARROS, Mônica. **Econometria – Semestre 2010.01.** Resumo Gujarati, USP, 2010.

Universidade Federal de Santa Maria – UFSM
Centro de Ciências Naturais e Exatas - CCNE
Curso de Especialização em Estatística e Modelagem Quantitativa - CEEMQ

REGRESSORAS CATEGÓRICAS: VARIÁVEIS *DUMMY*

Tatiane Ribeiro e Rosiéli Ruviaro

Prof.: Drª. Ana Lúcia Souza Silva Mateus
Disciplina: Análise de Correlação e Regressão

Santa Maria, 13 de dezembro de 2017.