ANÁLISE DE DADOS MULTIVARIADOS I - REGRESSÃO

(AULA 05)

Novembro e dezembro de 2018

Reinaldo Soares de Camargo

Modelos de Regressão

- **Estimação dos coeficientes desconhecidos β_0 , β_1 via método de mínimos quadrados ordinário possui forma fechada a partir da amostra observada.**
  - No caso de uma variável explicativa (regressão linear simples),

$$\hat{y}_i = \beta_0 + \beta_1 x_1 + \varepsilon_i$$

  - podemos estimar β1 a expressão

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

  - O coeficientes **β_0 é estimado por**

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

  - O nome “mínimos quadrados ordinários” é utilizado porque as estimativas do intercepto ($\hat{\beta}_0$) e da inclinação ($\hat{\beta}_1$) minimizam a soma dos resíduos quadrados..

Modelos de Regressão

Dados de peso(x) altura(y) da turma de regressão da ENAP (aula2)

i	X	Y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	80	175	6,05	5,3	32,065	36,60	28,09
2	74	172	0,05	2,3	0,115	0,00	5,29
3	113	176	39,05	6,3	246,015	1.524,90	39,69
4	56	155	-17,95	-14,7	263,865	322,20	216,09
5	78	170	4,05	0,3	1,215	16,40	0,09
6	78	178	4,05	8,3	33,615	16,40	68,89
7	60	158	-13,95	-11,7	163,215	194,60	136,89
8	65	160	-8,95	-9,7	86,815	80,10	94,09
9	66	160	-7,95	-9,7	77,115	63,20	94,09
10	87	175	13,05	5,3	69,165	170,30	28,09
11	56	167	-17,95	-2,7	48,465	322,20	7,29
12	91	178	17,05	8,3	141,515	290,70	68,89
13	77	177	3,05	7,3	22,265	9,30	53,29
14	85	185	11,05	15,3	169,065	122,10	234,09
15	70	162	-3,95	-7,7	30,415	15,60	59,29
16	85	171	11,05	1,3	14,365	122,10	1,69
17	56	157	-17,95	-12,7	227,965	322,20	161,29
18	73	175	-0,95	5,3	-5,035	0,90	28,09
19	78	180	4,05	10,3	41,715	16,40	106,09
20	51	163	-22,95	-6,7	153,765	526,70	44,89
Soma	1479	3394	- 0,00	0,00	1.817,70	4.172,95	1.476,20
Média	74	169,7	- 0,00	0,00	90,89	208,65	73,81

$$y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \frac{1817,70}{4172,\mathbf{95}} = 0,43559$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_0 = 169,70 - 0,43559(74)$$

$$\hat{\beta}_0 = 137,48804$$

Equação da regressão

$$\hat{y}_i = 137,48804 + 0,43559 x_1$$

Modelos de Regressão

Execute o programa: **ajuste5.R** (até linha 68), disponível na aula 05 do git para acompanhar os cálculos.

```
> #----beta1--
> beta1 = sum(x_m_x*y_m_y)/sum(x_mx_2)
> beta1
[1] 0.4355911
>
>
> #----beta0--
> beta0 = m_y - beta1*m_x
> beta0
[1] 137.488
```

$$y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\hat{\beta}_1 = \frac{1817{,}70}{4172{,}\mathbf{95}} = 0{,}43559$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

$$\hat{\beta}_0 = 169{,}70 - 0{,}43559(74)$$

$$\hat{\beta}_0 = 137{,}48804$$

Resultados usando ml

```
> mod0$coefficients
(Intercept)        Peso
137.4880360   0.4355911
>
```

Equação da regressão

$$\hat{y}_i = 137{,}48804 + 0{,}43559 x_1$$

Modelos de Regressão

- Coeficiente de determinação (R^2) da regressão

  - Para um modelo de regressão qualquer estimado, é interessante termos uma medida do ajuste dessa regressão

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots + \beta_k x_{ki} + \ \epsilon_i$$

  - A medida de ajuste está intrinsicamente ligada à importância do termo ϵ_i. Esse termo corresponde à parcela da variável explicada y_i que não é explicada pelas variáveis independentes

  - A medida mais comumente utilizada para verificar o ajuste de uma regressão é chamada coeficiente de determinação, que é calculada pela expressão:

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

  - Pode-se mostrar que o coeficiente de determinação varia entre 0 e 1 (dado que o intercepto está incluído na regressão)

  - O coeficiente de determinação pode ser interpretado como o percentual da variação da variável predita que é explicado pela regressão

Modelos de Regressão – R^2

Dados de peso(x) altura(y) da turma de regressão da ENAP (aula2)

i	X	Y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$\hat{y}$	$\hat{y} - \bar{y}$	$(\hat{y} - \bar{y})^2$
1	80	175	6,05	5,3	32,065	36,60	28,09	172,34	2,64	6,94
2	74	172	0,05	2,3	0,115	0,00	5,29	169,72	0,02	0,00
3	113	176	39,05	6,3	246,015	1.524,90	39,69	186,71	17,01	289,33
4	56	155	-17,95	-14,7	263,865	322,20	216,09	161,88	- 7,82	61,14
5	78	170	4,05	0,3	1,215	16,40	0,09	171,46	1,76	3,11
6	78	178	4,05	8,3	33,615	16,40	68,89	171,46	1,76	3,11
7	60	158	-13,95	-11,7	163,215	194,60	136,89	163,62	- 6,08	36,92
8	65	160	-8,95	-9,7	86,815	80,10	94,09	165,80	- 3,90	15,20
9	66	160	-7,95	-9,7	77,115	63,20	94,09	166,24	- 3,46	11,99
10	87	175	13,05	5,3	69,165	170,30	28,09	175,38	5,68	32,31
11	56	167	-17,95	-2,7	48,465	322,20	7,29	161,88	- 7,82	61,14
12	91	178	17,05	8,3	141,515	290,70	68,89	177,13	7,43	55,16
13	77	177	3,05	7,3	22,265	9,30	53,29	171,03	1,33	1,76
14	85	185	11,05	15,3	169,065	122,10	234,09	174,51	4,81	23,17
15	70	162	-3,95	-7,7	30,415	15,60	59,29	167,98	- 1,72	2,96
16	85	171	11,05	1,3	14,365	122,10	1,69	174,51	4,81	23,17
17	56	157	-17,95	-12,7	227,965	322,20	161,29	161,88	- 7,82	61,14
18	73	175	-0,95	5,3	-5,035	0,90	28,09	169,29	- 0,41	0,17
19	78	180	4,05	10,3	41,715	16,40	106,09	171,46	1,76	3,11
20	51	163	-22,95	-6,7	153,765	526,70	44,89	159,70	- 10,00	99,94
Soma	1479	3394,00	- 0,00	0,00	1.817,70	4.172,95	1.476,20	3.394,00	- 0,00	791,77
Média	74	169,70	- 0,00	0,00	90,89	208,65	73,81	169,70	- 0,00	39,59

$$R^2 = \left[1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right] = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad \boldsymbol{R^2} = \frac{791{,}77}{1476{,}20} = 0{,}5364$$

Modelos de Regressão

- Intepretação do coeficiente de determinação
    - Percentual da variação da variável dependente que pode ser explicado pela variação das variáveis independentes
- Cuidado: quando incluímos variáveis na equação, independente de essas fazerem sentido ou não, o R^2 sempre aumenta
- Alternativa para "avaliar" a inclusão da nova variável: **R^2 ajustado**

$$R^2_{ajustado} = 1 - \left[\frac{(1 - R^2)(n - 1)}{n - k - 1}\right]$$

- *n* é o número de observações na amostra
- *k* é o número de variáveis explicativas (sem considerar a constante)
- Quando incluímos variáveis 'desnecessárias' na regressão, o R^2 ajustado diminui

Modelos de Regressão – R^2 Ajustado

Dados de peso(x) altura(y) da turma de regressão da ENAP (aula2)

i	X	Y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$\hat{y}$	$\hat{y} - \bar{y}$	$(\hat{y} - \bar{y})^2$
1	80	175	6,05	5,3	32,065	36,60	28,09	172,34	2,64	6,94
2	74	172	0,05	2,3	0,115	0,00	5,29	169,72	0,02	0,00
3	113	176	39,05	6,3	246,015	1.524,90	39,69	186,71	17,01	289,33
4	56	155	-17,95	-14,7	263,865	322,20	216,09	161,88	- 7,82	61,14
5	78	170	4,05	0,3	1,215	16,40	0,09	171,46	1,76	3,11
6	78	178	4,05	8,3	33,615	16,40	68,89	171,46	1,76	3,11
7	60	158	-13,95	-11,7	163,215	194,60	136,89	163,62	- 6,08	36,92
8	65	160	-8,95	-9,7	86,815	80,10	94,09	165,80	- 3,90	15,20
9	66	160	-7,95	-9,7	77,115	63,20	94,09	166,24	- 3,46	11,99
10	87	175	13,05	5,3	69,165	170,30	28,09	175,38	5,68	32,31
11	56	167	-17,95	-2,7	48,465	322,20	7,29	161,88	- 7,82	61,14
12	91	178	17,05	8,3	141,515	290,70	68,89	177,13	7,43	55,16
13	77	177	3,05	7,3	22,265	9,30	53,29	171,03	1,33	1,76
14	85	185	11,05	15,3	169,065	122,10	234,09	174,51	4,81	23,17
15	70	162	-3,95	-7,7	30,415	15,60	59,29	167,98	- 1,72	2,96
16	85	171	11,05	1,3	14,365	122,10	1,69	174,51	4,81	23,17
17	56	157	-17,95	-12,7	227,965	322,20	161,29	161,88	- 7,82	61,14
18	73	175	-0,95	5,3	-5,035	0,90	28,09	169,29	- 0,41	0,17
19	78	180	4,05	10,3	41,715	16,40	106,09	171,46	1,76	3,11
20	51	163	-22,95	-6,7	153,765	526,70	44,89	159,70	- 10,00	99,94
Soma	1479	3394,00	- 0,00	0,00	1.817,70	4.172,95	1.476,20	3.394,00	- 0,00	791,77
Média	74	169,70	- 0,00	0,00	90,89	208,65	73,81	169,70	- 0,00	39,59

$$R^2_{ajustado} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right] = 1 - \left[\frac{(1-0{,}5364)(20-1)}{20-1-1}\right] = 0{,}5106$$

Modelos de Regressão – R^2 e R^2 Ajustado

$$\boldsymbol{R^2} = \frac{791{,}77}{1476{,}20} = 0{,}5364$$

```
> r2 = sum(yh_ym^2)/sum(y_m_y^2)
> r2
[1] 0.5363596
```

$$R^2_{ajustado} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right] = 1 - \left[\frac{(1-0{,}5364)(20-1)}{20-1-1}\right] = 0{,}5106$$

```
> r2_ajustado =  1- (1-r2)*(n-1)/(n-k-1)
> r2_ajustado
[1] 0.5106018
```

Modelos de Regressão – Erro padrão dos resíduos

i	X	Genero	Y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$\hat{y}$	$\hat{y} - \bar{y}$	$(\hat{y} - \bar{y})^2$	$y - \hat{y}$	$(y - \hat{y})^2$
1	80	M	175	6,05	5,3	32,065	36,60	28,09	172,34	2,64	6,94	2,66	7,10
2	74	M	172	0,05	2,3	0,115	0,00	5,29	169,72	0,02	0,00	2,28	5,19
3	113	M	176	39,05	6,3	246,015	1.524,90	39,69	186,71	17,01	289,33	- 10,71	114,70
4	56	F	155	-17,95	-14,7	263,865	322,20	216,09	161,88	- 7,82	61,14	- 6,88	47,35
5	78	M	170	4,05	0,3	1,215	16,40	0,09	171,46	1,76	3,11	- 1,46	2,14
6	78	M	178	4,05	8,3	33,615	16,40	68,89	171,46	1,76	3,11	6,54	42,72
7	60	F	158	-13,95	-11,7	163,215	194,60	136,89	163,62	- 6,08	36,92	- 5,62	31,62
8	65	F	160	-8,95	-9,7	86,815	80,10	94,09	165,80	- 3,90	15,20	- 5,80	33,66
9	66	F	160	-7,95	-9,7	77,115	63,20	94,09	166,24	- 3,46	11,99	- 6,24	38,90
10	87	M	175	13,05	5,3	69,165	170,30	28,09	175,38	5,68	32,31	- 0,38	0,15
11	56	F	167	-17,95	-2,7	48,465	322,20	7,29	161,88	- 7,82	61,14	5,12	26,20
12	91	M	178	17,05	8,3	141,515	290,70	68,89	177,13	7,43	55,16	0,87	0,76
13	77	M	177	3,05	7,3	22,265	9,30	53,29	171,03	1,33	1,76	5,97	35,66
14	85	M	185	11,05	15,3	169,065	122,10	234,09	174,51	4,81	23,17	10,49	109,97
15	70	F	162	-3,95	-7,7	30,415	15,60	59,29	167,98	- 1,72	2,96	- 5,98	35,75
16	85	M	171	11,05	1,3	14,365	122,10	1,69	174,51	4,81	23,17	- 3,51	12,34
17	56	F	157	-17,95	-12,7	227,965	322,20	161,29	161,88	- 7,82	61,14	- 4,88	23,82
18	73	M	175	-0,95	5,3	-5,035	0,90	28,09	169,29	- 0,41	0,17	5,71	32,65
19	78	M	180	4,05	10,3	41,715	16,40	106,09	171,46	1,76	3,11	8,54	72,86
20	51	F	163	-22,95	-6,7	153,765	526,70	44,89	159,70	- 10,00	99,94	[illegible]	[illegible]
Soma	1479		3394,00	- 0,00	0,00	1.817,70	4.172,95	1.476,20	3.394,00	- 0,00	791,77	0,00	684,43
Média	74		169,70	- 0,00	0,00	90,89	208,65	73,81	169,70	- 0,00	39,59	[illegible]	[illegible]

```
Residual standard error: 6.166 on 18 degrees of freedom
Multiple R-squared:  0.5364,    Adjusted R-squared:  0.5106
F-statistic: 20.82 on 1 and 18 DF,  p-value: 0.000241
```

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y - \hat{Y}_i\right)^2}{n-2}} = \sqrt{\frac{684,43}{18}} = 6,166$$

Modelos de Regressão – Erro padrão do β_0, β_1

$$\widehat{\sigma i}_{(\beta_1)}=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/(n-2)}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}=\sqrt{\frac{684{,}43/18}{4172{,}95}}=0{,}9546$$

```
> err_pad_beta1 = sqrt(sum(y_yh^2)/(n-2)/sum(x_m_x^2))
> err_pad_beta1
[1] 0.09545648
```

$$\widehat{\sigma i}_{(\beta_0)}=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/(n-2)\sum_{i=1}^{n}(x_i)^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}}=\sqrt{\frac{38{,}02*113545{,}00}{20*4172{,}955}}=7{,}1924$$

```
> err_pad_beta0 = sqrt((sum(y_yh^2)/(n-2))*sum(x^2) / (n*sum(x_m_x^2)))
> err_pad_beta0
[1] 7.19241
```

Modelos de Regressão – t_valor e p_valor

$$\text{t–valor}_{(\beta_0)} = \frac{\widehat{b_0}}{\widehat{\sigma}_{i(\beta_0)}}$$

```
> t_value_b0 = beta0/err_pad_beta0
> t_value_b0
[1] 19.11571
```

$$\text{t–valor}_{(\beta_1)} = \frac{\widehat{b_1}}{\widehat{\sigma}_{i(\beta_1)}}$$

```
> t_value_b1 = beta1/err_pad_beta1
> t_value_b1
[1] 4.563243
```

Modelos de Regressão – Intervalo de confiança

$$\widehat{\beta}_i \pm t_{\frac{\alpha}{2},n-1} \frac{\widehat{\sigma}_{i(\beta_0)}}{\sqrt{n}}$$

```
> alpha <- 0.05 # intevalo de confiança de 95%
>
> half.width <- qt(1-alpha/2, n-1)*err_pad_beta0/sqrt(n)
> half.width
[1] 3.366151
>
> c(beta0 - half.width, beta0 + half.width)
[1] 134.1219 140.8542
> beta0
[1] 137.488
```

```
> half.width <- qt(1-alpha/2, n-1)*err_pad_beta1/sqrt(n)
> half.width
[1] 0.04467501
>
> c(beta1 - half.width, beta1 + half.width)
[1] 0.3909161 0.4802661
> beta1
[1] 0.4355911
```

Modelos de Regressão na Forma Matricial

Modelos de Regressão em Notação Matricial

- Considere agora o modelo geral de regressão:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots + \beta_k x_{ki} + \epsilon_i$$

- Na equação acima, o termo ϵ_i corresponde ao erro da regressão, e incorpora todos os demais fatores, não presentes na equação linear, que explicam a variável y_i
- Até agora, utilizamos notações mais simplificadas, para fins de apresentação dos conceitos e utilização dos modelos de regressão
- Na maioria dos livros de regressão, e em vários trabalhos publicados, utiliza-se a notação matricial
- A notação matricial, além de simplificar a apresentação dos resultados, também indica expressões que podem ser utilizadas no software R
- O R tem todo um arcabouço para somas, multiplicação, inversão, subtração, transposição etc. de matrizes
- Conforme veremos mais adiante, fórmulas para o estimador de mínimos quadrados ordinário são bem simples, quando utilizamos a notação matricial
- Nesta seção, apresentaremos alguns conceitos básicos da notação matricial, que são muito utilizados na literatura

Modelos de Regressão – Forma Matricial

Modelo de regressão linear simples (apenas um preditor)

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

$$\epsilon_i \sim^{iid} N(0, \sigma^2)$$

Amostra com *n* observações

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \epsilon_2 \\ \vdots & \vdots \quad \vdots \\ Y_n &= \beta_0 + \beta_1 X_n + \epsilon_n \end{aligned}$$

Modelos de Regressão – Forma Matricial

Escrevendo na forma de matrizes:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Design matriz = $\mathbf{X}_{n\times 2} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$

No caso de regressão linear múltipla, com várias variáveis preditoras, as fórmulas são totalmente similares, aumentando-se apenas o número de colunas da matriz de desenho

Modelos de Regressão – Forma Matricial

Vetor de parâmetros = $$\beta_{2\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Vetor de resíduos = $$\epsilon_{n\times 1} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Vetor de respostas = $$\mathbf{Y}_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Modelos de Regressão – Forma Matricial

Equação matricial:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \epsilon \\ \mathbf{Y}_{n\times 1} &= \mathbf{X}_{n\times 2}\beta_{2\times 1} + \epsilon_{n\times 1} \end{aligned}$$

Os resíduos são normais, independentes e identicamente distribuídos (possuem correlação igual a zero). A matriz de covariância do vetor de resíduos é dada por:

$$\sigma^{\mathbf{2}}\{\epsilon\}_{n\times n} = Cov \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \sigma^2 \mathbf{I}_{n\times n} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

Portanto, os resíduos possuem distribuição normal multivariada, com médias zero e matriz de covariância dada pela matriz anterior

Modelos de Regressão – Forma Matricial

Equação matricial:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \epsilon \\ \mathbf{Y}_{n\times 1} &= \mathbf{X}_{n\times 2}\beta_{2\times 1} + \epsilon_{n\times 1} \end{aligned}$$

A matriz de covariância do vetor de respostas resíduos é dada por:

$$\sigma^{\mathbf{2}}\{\mathbf{Y}\}_{n\times n} = Cov \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \sigma^2 \mathbf{I}_{n\times n}$$

Portanto, a matriz de respostas possui distribuição normal multivariada, com médias dadas pelo vetor $\mathbf{X} \times \beta$, e matriz de covariância dada pela matriz anterior

Modelos de Regressão – Forma Matricial

Para um vetor de parâmetros β, os resíduos podem ser escritos como:

$$\epsilon = \mathbf{Y} - \mathbf{X}\beta$$

A soma dos quadrados dos resíduos é dada pelo produto:

$$\sum \epsilon_i^2 = [\epsilon_1 \, \epsilon_2 \, \cdots \epsilon_n] \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \epsilon' \epsilon$$

Pode se mostrar que o estimador de mínimos quadrados ordinários para o vetor de parâmetros β é dado por:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Modelos de Regressão – Forma Matricial

Para uma regressão linear simples, pode-se mostrar que a expressão

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

se reduz a:

$$\begin{aligned} b_1 &= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \equiv \frac{SS_{XY}}{SS_X} \equiv \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ b_0 &= \bar{Y} - b_1\bar{X} \end{aligned}$$

Os valores preditos com a regressão são escritos como:

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \begin{bmatrix} b_0 + b_1 X_1 \\ b_0 + b_1 X_2 \\ \vdots \\ b_0 + b_1 X_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \mathbf{Xb}$$

Modelos de Regressão – Forma Matricial

Exemplo do peso da turma de regressão da ENAP em função do peso e do genero

$$X = \begin{vmatrix} \text{Intercepto} & \text{Peso} & \text{Masculino} \\ 1 & 80 & 1 \\ 1 & 74 & 1 \\ 1 & 113 & 1 \\ 1 & 56 & 0 \\ 1 & 78 & 1 \\ 1 & 78 & 1 \\ 1 & 60 & 0 \\ 1 & 65 & 0 \\ 1 & 66 & 0 \\ 1 & 87 & 1 \\ 1 & 56 & 0 \\ 1 & 91 & 1 \\ 1 & 77 & 1 \\ 1 & 85 & 1 \\ 1 & 70 & 0 \\ 1 & 85 & 1 \\ 1 & 56 & 0 \\ 1 & 73 & 1 \\ 1 & 78 & 1 \\ 1 & 51 & 0 \end{vmatrix} \quad Y = \begin{vmatrix} \text{Altura} \\ 175 \\ 172 \\ 176 \\ 155 \\ 170 \\ 178 \\ 158 \\ 160 \\ 160 \\ 175 \\ 167 \\ 178 \\ 177 \\ 185 \\ 162 \\ 171 \\ 157 \\ 175 \\ 180 \\ 163 \end{vmatrix}$$

$$X'X = \begin{vmatrix} \text{Intercepto} & \text{Peso} & \text{Masculino} \\ 2{,}406 & -0{,}038 & 0{,}7588 \\ -0{,}038 & 0{,}0006 & -0{,}1473 \\ 0{,}7258 & -0{,}0147 & 0{,}5508 \end{vmatrix}$$

$$(X'X)^{-1} = \begin{vmatrix} 2{,}842 & - & 0{,}005 & & 153{,}741 \\ - & 0{,}005 & & 0{,}000 & - \\ 153{,}741 & - & 2{,}830 & & 9.206.713 \end{vmatrix} \quad 2{,}830$$

$$X'Y = \begin{vmatrix} 3.394 \\ 252.804 \\ 2.112 \end{vmatrix}$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \equiv (X'X)^{-1}X'Y = \begin{vmatrix} 157{,}9690 \\ 0{,}0380 \\ 14{,}8661 \end{vmatrix}$$

```
> mod$coefficients
 (Intercept)         Peso      GeneroM
157.96899256   0.03801679  14.86610962
```

Modelos de Regressão – Forma Matricial

- De onde vem a coluna de erros padrões dos parâmetros da regressão?

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.8004 -2.4620 -0.5442  1.8289  8.9335

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 157.96899    6.33119  24.951 7.84e-15 ***
Peso          0.03802    0.10274   0.370 0.715936
GeneroM      14.86611    3.02935   4.907 0.000133 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.082 on 17 degrees of freedom
Multiple R-squared:  0.8081,    Adjusted R-squared:  0.7856
F-statistic:  35.8 on 2 and 17 DF,  p-value: 8.041e-07
```

- A partir dessa coluna de erros padrões, encontram-se:
    - intervalos de confiança
    - estatísticas testes
    - p-valores

Modelos de Regressão – Forma Matricial

Matriz de variância-covariância para os estimadores dos coeficientes da regressão:

$$\hat{\Sigma} = \begin{bmatrix} \hat{\mathrm{Var}}(\hat{\beta}_0) & \hat{\mathrm{Cov}}(\hat{\beta}_0, \hat{\beta}_1) & \hat{\mathrm{Cov}}(\hat{\beta}_0, \hat{\beta}_2) & \hat{\mathrm{Cov}}(\hat{\beta}_0, \hat{\beta}_3) \\ \hat{\mathrm{Cov}}(\hat{\beta}_1, \hat{\beta}_0) & \hat{\mathrm{Var}}(\hat{\beta}_1) & \hat{\mathrm{Cov}}(\hat{\beta}_1, \hat{\beta}_2) & \hat{\mathrm{Cov}}(\hat{\beta}_1, \hat{\beta}_3) \\ \hat{\mathrm{Cov}}(\hat{\beta}_2, \hat{\beta}_0) & \hat{\mathrm{Cov}}(\hat{\beta}_2, \hat{\beta}_1) & \hat{\mathrm{Var}}(\hat{\beta}_2) & \hat{\mathrm{Cov}}(\hat{\beta}_2, \hat{\beta}_3) \\ \hat{\mathrm{Cov}}(\hat{\beta}_3, \hat{\beta}_0) & \hat{\mathrm{Cov}}(\hat{\beta}_3, \hat{\beta}_1) & \hat{\mathrm{Cov}}(\hat{\beta}_3, \hat{\beta}_2) & \hat{\mathrm{Var}}(\hat{\beta}_3) \end{bmatrix}$$

Na diagonal principal, temos as variâncias das estimativas para cada coeficiente.
Fora da diagonal principal, temos as covariâncias entre as estimativas.

Fórmula matricial para a matriz de variância-covariância:

$$\hat{\Sigma} = \hat{\sigma}^2(\mathrm{X}'\mathrm{X})^{-1}$$

Na qual $\hat{\sigma}^2$ é a variância estimada para os erros da regressão (com *k* variáveis explicativas):

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}[y_i - \hat{y}_i]^2}{n-k-1}$$

Modelos de Regressão – Forma Matricial

- Exemplo no programa “ajuste5.R”

```
#-------------------------------------------------------------------------
#---- exemplos de expressões matriciais em R para modelos de regressão
#-------------------------------------------------------------------------

mod <- lm(Altura ~ Peso + Genero , data = dados)
summary(mod)


X1 <- model.matrix(mod) #---- design matrix para o modelo de regressão
head(X1)
tail(X1)
df.X1 <- as.data.frame(X1)  #---- transformando em data.frame para visualização
mais fácil
View(df.X1)
```

Modelos de Regressão – Forma Matricial

- Exemplo no programa “ajuste5.R”

```
#--- desvio padrão e variância dos resíduos da regressão - cálculo manual

n <- nrow(X1)      #--- número de observações
k <- ncol(X1) - 1  #--- número de var explicativas
n;k

mod2.residuos <- mod$residuals
head(mod2.residuos)
tail(mod2.residuos)
hist(mod2.residuos, col = 'red', breaks = 5)

mod2.residuos.var <- (t(mod2.residuos) %*% mod2.residuos) / (n-k-1)
mod2.residuos.var
mod2.residuos.desvpad <- sqrt(mod2.residuos.var)
mod2.residuos.desvpad
```

- Resultado no R:

```
> mod2.residuos.desvpad
         [,1]
[1,] 4.081659
```

Modelos de Regressão – Forma Matricial

- Comparação com o resultado direto no sumário da regressão:

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.8004 -2.4620 -0.5442  1.8289  8.9335

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 157.96899    6.33119  24.951 7.84e-15 ***
Peso          0.03802    0.10274   0.370 0.715936
GeneroM      14.86611    3.02935   4.907 0.000133 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.082 on 17 degrees of freedom
Multiple R-squared:  0.8081,    Adjusted R-squared:  0.7856
F-statistic:  35.8 on 2 and 17 DF,  p-value: 8.041e-07
```

Modelos de Regressão – Forma Matricial

- Exemplo no programa “ajuste5.R”
- Expressão para a matriz de variância-covariância: $\hat{\Sigma} = \hat{\sigma}^2(\mathrm{X}'\mathrm{X})^{-1}$

```
#--- matriz de variância-covariância e erros padrões dos coeficientes

mod2.residuos.var <- as.numeric(mod2.residuos.var)
mod2.residuos.var
sqrt(mod2.residuos.var)

cov1 <- mod2.residuos.var * (solve(t(X1) %*% X1))
cov1
diag(cov1)
erropadrao1 <- sqrt(diag(cov1))
erropadrao1
```

- Resultado no R:

```
> erropadrao1
(Intercept)        Peso     GeneroM
  6.3311880   0.1027422   3.0293524
> 
```

Modelos de Regressão – Forma Matricial

- Comparação com o resultado direto no sumário da regressão:

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.8004 -2.4620 -0.5442  1.8289  8.9335

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 157.96899    6.33119  24.951 7.84e-15 ***
Peso          0.03802    0.10274   0.370 0.715936
GeneroM      14.86611    3.02935   4.907 0.000133 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.082 on 17 degrees of freedom
Multiple R-squared:  0.8081,    Adjusted R-squared:  0.7856
F-statistic:  35.8 on 2 and 17 DF,  p-value: 8.041e-07
```

Modelos de Regressão – Forma Matricial

- Exemplo no programa “ajuste5.R”
- Expressão para os coeficientes estimados: $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

```
#--- coeficientes estimados, estatística teste e pvalores


Y1 <- dados$Altura
beta1 <- (solve(t(X1) %*% X1)) %*% (t(X1) %*% Y1) #--- coeficientes
beta1

estatistica_t1 <- beta1 / erropadrao1 #--- estatística teste t
estatistica_t1

pvalor1 <- 2*(1 - pt(abs(estatistica_t1), n-k-1)) #--- p-valores (com t-Student)
pvalor1

resultados1 <- cbind(beta1, erropadrao1, estatistica_t1, pvalor1) #--- juntando tudo
resultados1
```

Modelos de Regressão – Forma Matricial

- Resultado no R:

```
> resultados1
                          erropadrao1
(Intercept) 157.96899256   6.3311880 24.9509243 7.993606e-15
Peso          0.03801679   0.1027422  0.3700212 7.159362e-01
GeneroM      14.86610962   3.0293524  4.9073556 1.330771e-04
```

- Output tradicional usando *summary* da regressão (para comparação):

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 157.96899    6.33119  24.951 7.84e-15 ***
Peso          0.03802    0.10274   0.370 0.715936
GeneroM      14.86611    3.02935   4.907 0.000133 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.082 on 17 degrees of freedom
Multiple R-squared:  0.8081,  Adjusted R-squared:  0.7856
F-statistic:  35.8 on 2 and 17 DF,  p-value: 8.041e-07
```

Testes de Hipóteses para Modelos de Regressão Linear

Testes de Hipóteses

- Testes de hipótese são utilizados para testar se 'suspeitas' sobre os parâmetros do modelo de regressão são verdadeiras ou não

- O teste de hipótese mais comum em modelos de regressão é a respeito do valor de um determinado coeficiente específico

- Por exemplo, considere o modelo de regressão abaixo

$$[Salário]_i = \beta_0 + \beta_1[Experiência]_i + \delta_1[DummyMulher]_i + \gamma_1[DummyMulher]_i \times [Experiência]_i + \epsilon_i$$

- Conforme vimos anteriormente, a variável $[DummyMulher]_i$ indica se o trabalhador é do sexo feminino ou não

- O coeficiente δ_1 indica a diferença de salários entre homens e mulheres, 'controlando-se' para as demais variáveis

- Gostaríamos de testar se existe ou não discriminação no mercado de trabalho. Para isso, podemos testar se o coeficiente δ_1 é nulo ou diferente de zero

Testes de Hipóteses

- Para testar a presença ou não de discriminação, podemos então proceder com um teste de hipóteses

- Os testes de hipóteses possuem quatro elementos básicos:

  - **Hipóteses nula e alternativa**

  - **Estatística teste**

  - **Distribuição da estatística teste**

  - **Regressão de rejeição da hipótese nula**

- A hipótese nula no exemplo em questão é dada por: $H_0: \gamma_1 = 0$

- A hipótese alternativa é dada por: $H_A: \gamma_1 \neq 0$

Testes de Hipóteses

- A estatística teste utilizada nesse caso, tem expressão

$$t_{stat} = \frac{[\text{Estimativa do coeficiente}]}{[\text{Erro padrão do coeficiente}]} = \frac{\hat{\gamma}_1}{s.e._{\hat{\gamma}_1}}$$

- Se quiséssemos testar um valor mais geral (não zero) para o coeficiente, teríamos:

$$H_0: \gamma_1 = a$$
$$H_A: \gamma_1 \neq a$$

- A estatística teste seria dada então por:

$$t_{stat} = \frac{[\text{Estimativa do coeficiente}] - a}{[\text{Erro padrão do coeficiente}]} = \frac{\hat{\gamma}_1 - a}{s.e._{\hat{\gamma}_1}}$$

- Em geral, os sumários da regressão linear sempre trazem o erro padrão, e a estatística teste, para testar a hipótese nula de que cada coeficiente individualmente é igual a zero (ou seja, a = 0)

Testes de Hipóteses

```
Call:
lm(formula = dados3$mort_infantil ~ dados3$renda_per_capita +
    dados3$indice_gini + dados3$salario_medio_mensal + dados3$perc_criancas_extrem_pobres +
    dados3$perc_criancas_pobres + dados3$perc_pessoas_dom_agua_estogo_inadequados +
    dados3$perc_pessoas_dom_paredes_inadequadas + dados3$perc_pop_dom_com_coleta_lixo)

Residuals:
     Min       1Q   Median       3Q      Max
-16.5530  -2.4952  -0.3666   1.9344  20.8067

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  1.936e+01  8.196e-01  23.627  < 2e-16 ***
dados3$renda_per_capita                     -1.278e-03  5.784e-04  -2.209  0.02721 *
dados3$indice_gini                          -1.430e+01  1.247e+00 -11.470  < 2e-16 ***
dados3$salario_medio_mensal                 -1.775e-01  9.515e-02  -1.866  0.06212 .
dados3$perc_criancas_extrem_pobres           3.854e-02  1.216e-02   3.169  0.00154 **
dados3$perc_criancas_pobres                  2.159e-01  1.148e-02  18.812  < 2e-16 ***
dados3$perc_pessoas_dom_agua_estogo_inadequados  5.055e-02  6.021e-03   8.397  < 2e-16 ***
dados3$perc_pessoas_dom_paredes_inadequadas  4.297e-02  7.924e-03   5.423 6.12e-08 ***
dados3$perc_pop_dom_com_coleta_lixo         -7.045e-03  6.520e-03  -1.080  0.27999
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.038 on 5555 degrees of freedom
Multiple R-squared:  0.6804,    Adjusted R-squared:  0.6799
F-statistic:  1478 on 8 and 5555 DF,  p-value: < 2.2e-16
```

[Erro padrão do coeficiente]

Testes de Hipóteses

```
Call:
lm(formula = dados3$mort_infantil ~ dados3$renda_per_capita +
    dados3$indice_gini + dados3$salario_medio_mensal + dados3$perc_criancas_extrem_pobres +
    dados3$perc_criancas_pobres + dados3$perc_pessoas_dom_agua_estogo_inadequados +
    dados3$perc_pessoas_dom_paredes_inadequadas + dados3$perc_pop_dom_com_coleta_lixo)

Residuals:
     Min       1Q   Median       3Q      Max
-16.5530  -2.4952  -0.3666   1.9344  20.8067

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  1.936e+01  8.196e-01  23.627  < 2e-16 ***
dados3$renda_per_capita                     -1.278e-03  5.784e-04  -2.209  0.02721 *
dados3$indice_gini                          -1.430e+01  1.247e+00 -11.470  < 2e-16 ***
dados3$salario_medio_mensal                 -1.775e-01  9.515e-02  -1.866  0.06212 .
dados3$perc_criancas_extrem_pobres           3.854e-02  1.216e-02   3.169  0.00154 **
dados3$perc_criancas_pobres                  2.159e-01  1.148e-02  18.812  < 2e-16 ***
dados3$perc_pessoas_dom_agua_estogo_inadequados  5.055e-02  6.021e-03   8.397  < 2e-16 ***
dados3$perc_pessoas_dom_paredes_inadequadas   4.297e-02  7.924e-03   5.423 6.12e-08 ***
dados3$perc_pop_dom_com_coleta_lixo         -7.045e-03  6.520e-03  -1.080  0.27999
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.038 on 5555 degrees of freedom
Multiple R-squared:  0.6804,    Adjusted R-squared:  0.6799
F-statistic:  1478 on 8 and 5555 DF,  p-value: < 2.2e-16
```

t_{stat}

Testes de Hipóteses

- Quando o número de observações na amostra é muito alto, a estatística teste tem distribuição aproximadamente normal padronizada (média zero e desvio-padrão igual a 1)

- Portanto, podemos escrever

$$t = \frac{\hat{\gamma}_1 - a}{s.e._{\hat{\gamma}_1}} \approx \mathrm{N}(0,1)$$

- Em geral, podemos melhorar a aproximação da estatística teste, utilizando-se uma distribuição *t-Student*, com *(n-k-1)* graus de liberdade, onde *n* é o número de observações na amostra, *k* é o número de variáveis preditoras

- Quando o valor de *n* é muito alto, a aproximação normal e a aproximação via *t-Student* apresentam resultados praticamente idênticos

- É possível mostrar que a distribuição *t-Student* converge para uma distribuição normal padronizada, quando o número de graus de liberdade aumenta

Testes de Hipóteses

- Finalmente, temos que ter uma regra de rejeição para a hipótese nula
- Nessa regra, temos que estabelecer a probabilidade de erro tipo I
- Essa probabilidade, corresponde à chance de rejeitarmos a hipótese nula, quando de fato ela é verdadeira
- Em geral, estabelecemos a probabilidade de erro tipo I (também denominada α do teste) igual a 1%, 5% ou 10%
- A partir daí, temos que encontrar os valores críticos do teste de hipótese, com base na distribuição para a estatística teste (por exemplo, normal padronizada ou *t-Student*)
- Esses valores críticos vão depender também da característica do teste de hipótese: bicaudal, unicaudal à direita, ou unicaudal à esquerda

Testes de Hipóteses

- Teste bicaudal, temos as hipóteses nula e alternativa da forma:

$$H_0: \gamma_1 = a$$
$$H_A: \gamma_1 \neq a$$

- Teste unicaudal à direita, temos:

$$H_0: \gamma_1 \leq a$$
$$H_A: \gamma_1 > a$$

- Finalmente, teste unicaudal à esquerda, temos:

$$H_0: \gamma_1 \geq a$$
$$H_A: \gamma_1 < a$$

Testes de Hipóteses

Valores críticos para o teste **bicaudal**:

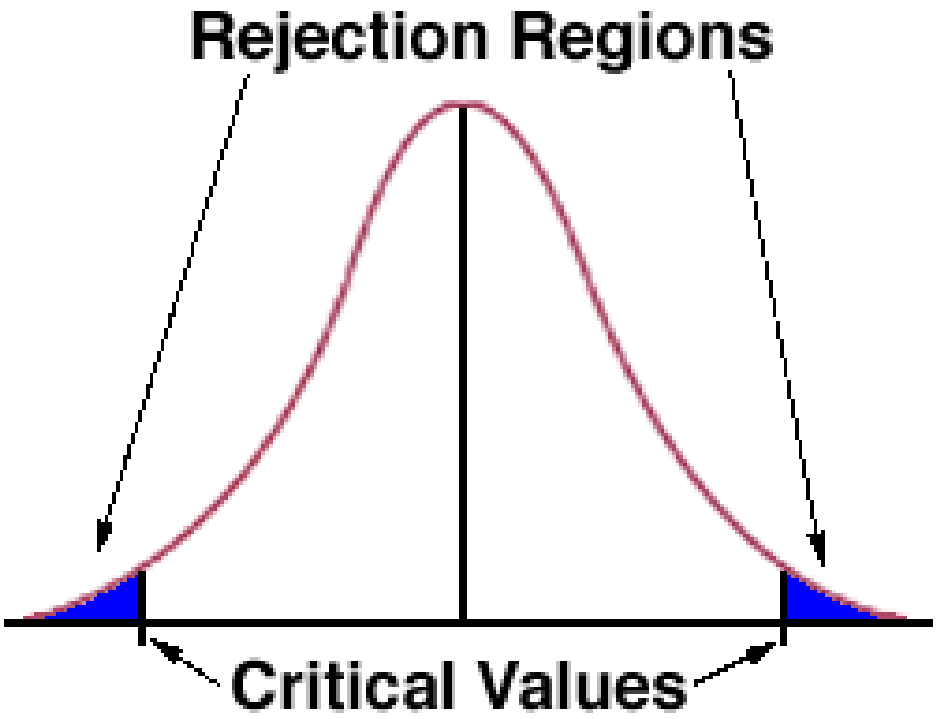
```
alpha <- 0.05;
q1 <- -qnorm(alpha/2);
q2 <- -qt(alpha/2, df=19);
q3 <- -qt(alpha/2, df=29);
q4 <- -qt(alpha/2, df=10000);
```

Obs. Note o sinal '-' nas fórmulas acima

Regra de rejeição da hipótese nula:

$|t_{stat}| > \text{valor crítico}$

Para um alpha = 5%, uma regra de bolsa é valor crítico aproximadamente igual a 2



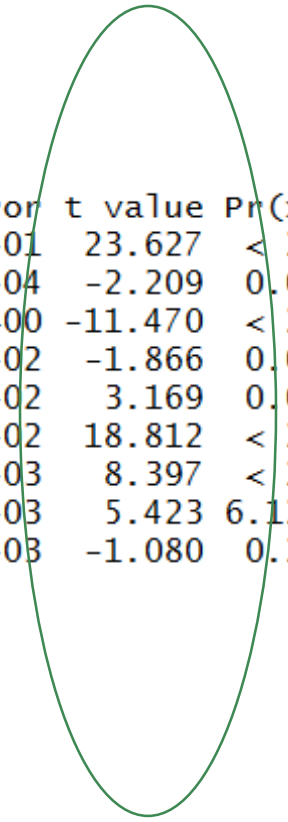
Testes de Hipóteses

```
Call:
lm(formula = dados3$mort_infantil ~ dados3$renda_per_capita +
    dados3$indice_gini + dados3$salario_medio_mensal + dados3$perc_criancas_extrem_pobres +
    dados3$perc_criancas_pobres + dados3$perc_pessoas_dom_agua_estogo_inadequados +
    dados3$perc_pessoas_dom_paredes_inadequadas + dados3$perc_pop_dom_com_coleta_lixo)

Residuals:
     Min       1Q   Median       3Q      Max
-16.5530  -2.4952  -0.3666   1.9344  20.8067

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  1.936e+01  8.196e-01  23.627  < 2e-16 ***
dados3$renda_per_capita                     -1.278e-03  5.784e-04  -2.209  0.02721 *
dados3$indice_gini                          -1.430e+01  1.247e+00 -11.470  < 2e-16 ***
dados3$salario_medio_mensal                 -1.775e-01  9.515e-02  -1.866  0.06212 .
dados3$perc_criancas_extrem_pobres           3.854e-02  1.216e-02   3.169  0.00154 **
dados3$perc_criancas_pobres                  2.159e-01  1.148e-02  18.812  < 2e-16 ***
dados3$perc_pessoas_dom_agua_estogo_inadequados  5.055e-02  6.021e-03   8.397  < 2e-16 ***
dados3$perc_pessoas_dom_paredes_inadequadas  4.297e-02  7.924e-03   5.423 6.12e-08 ***
dados3$perc_pop_dom_com_coleta_lixo         -7.045e-03  6.520e-03  -1.080  0.27999
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.038 on 5555 degrees of freedom
Multiple R-squared:  0.6804,    Adjusted R-squared:  0.6799
F-statistic:  1478 on 8 and 5555 DF,  p-value: < 2.2e-16
```



Testes de Hipóteses

Valores críticos para o teste **unicaudal à direita**:

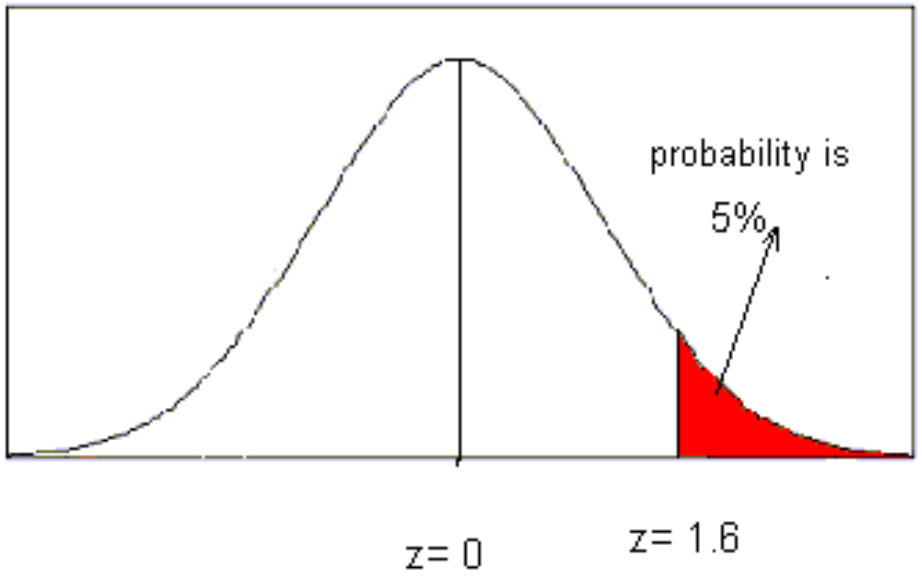
```
alpha <- 0.05;
q1 <- -qnorm(alpha);
q2 <- -qt(alpha, df=19);
q3 <- -qt(alpha, df=29);
q4 <- -qt(alpha, df=10000);
```

Obs. Note o sinal '-' nas fórmulas acima

Regra de rejeição da hipótese nula:

$t_{stat} >$ valor crítico

Obs. Não se aplica o valor absoluto nesse caso



Testes de Hipóteses

Valores críticos para o teste **unicaudal à esquerda**:

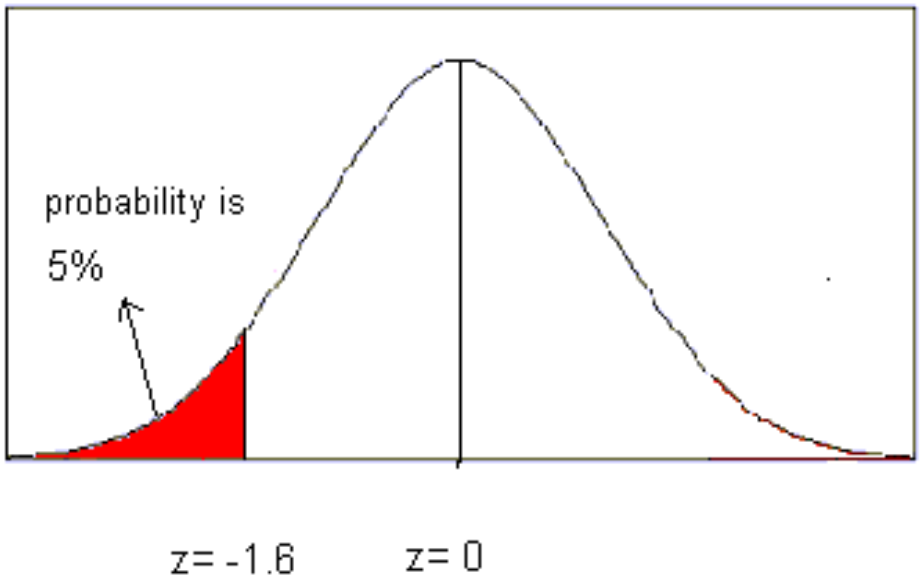
```
alpha <- 0.05;
q1 <- qnorm(alpha);
q2 <- qt(alpha, df=19);
q3 <- qt(alpha, df=29);
q4 <- qt(alpha, df=10000);
```

Obs. Note que não há mais o sinal '-' nas fórmulas acima

Regra de rejeição da hipótese nula:

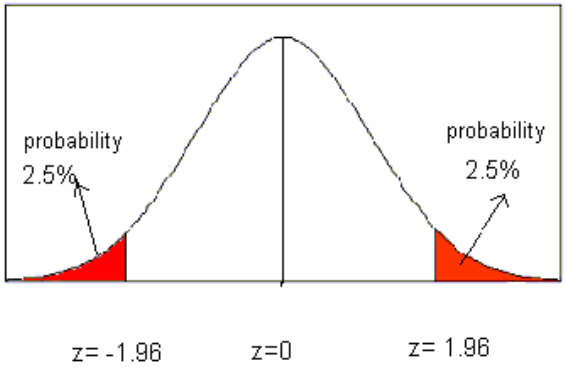
$t_{stat} <$ valor crítico

Obs. Não se aplica o valor absoluto nesse caso



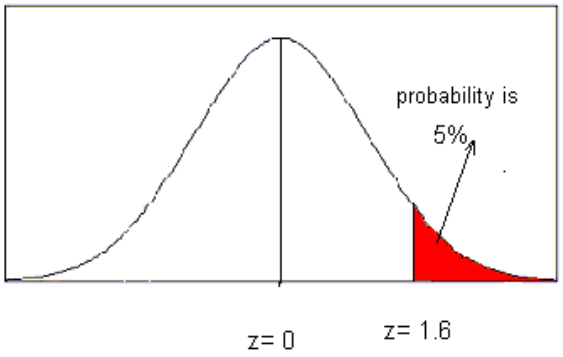
- Teste bicaudal :

$$H_0: \gamma_1 = a$$
$$H_A: \gamma_1 \neq a$$



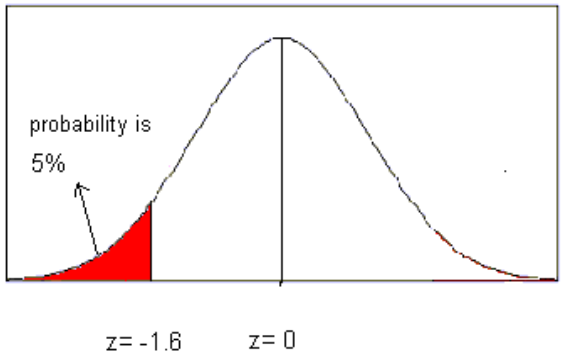
- Teste unicaudal à direita:

$$H_0: \gamma_1 \leq a$$
$$H_A: \gamma_1 > a$$



- Teste unicaudal à esquerda:

$$H_0: \gamma_1 \geq a$$
$$H_A: \gamma_1 < a$$



Testes de Hipóteses e P-Valores

- Na discussão acima, vimos que, para rejeitar a hipótese nula, precisamos comparar o valor da estatística teste com o valor crítico
- Essa comparação depende de se o teste é do tipo bicaudal, unicaudal à direita, ou unicaudal à esquerda
- Uma outra forma de se verificar se rejeitamos ou não a hipótese nula é através da utilização do que chamamos de *p-valor*
- Para entender o conceito de p-valor, considere um teste bicaudal para o coeficiente γ_1 da regressão para testar a discriminação de salários entre homens e mulheres
- As hipóteses nulas e alternativas são:

$$H_0: \gamma_1 = a$$
$$H_A: \gamma_1 \neq a$$

Testes de Hipóteses e P-Valores

- Para rejeitar a hipótese nula com nível de significância de 1%, no teste bi-caudal, precisamos que a estatística teste satisfaça (de acordo com a normal padronizada):

$$|t_{stat}| = \left|\frac{\hat{\gamma}_1}{s.e._{\hat{\gamma}_1}}\right| > 2.58$$

- Para rejeitar a hipótese nula com nível de significância de 5%, no teste bi-caudal, precisamos que a estatística teste satisfaça (de acordo com a normal padronizada):

$$|t_{stat}| = \left|\frac{\hat{\gamma}_1}{s.e._{\hat{\gamma}_1}}\right| > 1{,}96$$

- Para rejeitar a hipótese nula com nível de significância de 10%, no teste bi-caudal, precisamos que a estatística teste satisfaça (de acordo com a normal padronizada):

$$|t_{stat}| = \left|\frac{\hat{\gamma}_1}{s.e._{\hat{\gamma}_1}}\right| > 1{,}645$$

- No caso de usarmos uma distribuição t-Student, os valores são um pouco maiores, dependendo do número de graus de liberdade

Testes de Hipóteses e P-Valores

- Portanto, podemos considerar a seguinte regra de rejeição da hipótese nula, com base nos *p-valores*
    - Se p-valor < 0.05, então rejeitamos a hipótese nula com probabilidade de erro tipo I igual a 5%
    - Se p-valor < 0.10, então rejeitamos a hipótese nula com probabilidade de erro tipo I igual a 10%
    - Se p-valor < 0.01, então rejeitamos a hipótese nula com probabilidade de erro tipo I igual a 1%
- Diversos softwares estatísticos indicam o nível de significância da rejeição da hipótese nula, utilizando símbolos como, por exemplo, *, **, ***. O significado de cada um desses símbolos é indicado juntamente com a tabela de resultados

Significância dos Parâmetros em Modelos de Regressão

```
> mod <- lm(Altura ~ Peso + Genero , data = dados)
> summary(mod)

Call:
lm(formula = Altura ~ Peso + Genero, data = dados)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8004 -2.4620 -0.5442  1.8289  8.9335

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 157.96899    6.33119  24.951 7.84e-15 ***
Peso          0.03802    0.10274   0.370 0.715936
GeneroM      14.86611    3.02935   4.907 0.000133 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.082 on 17 degrees of freedom
Multiple R-squared:  0.8081,    Adjusted R-squared:  0.7856
F-statistic:  35.8 on 2 and 17 DF,  p-value: 8.041e-07
```

Modelos de Regressão

- para o modelo de regressão anterior, quais os coeficientes estatisticamente significantes a 1%, 5%, e 10%? Quais coeficientes não são significantes nem mesmo a 10%?

Modelos de Regressão

- **2ª Lista de exercícios para entregar em 19/11/2018.**
  - Os exercícios podem ser entregues em grupos de 2 alunos, e o grupo deve submeter o código em R utilizado para responder ao exercício, juntamente com a discussão dos resultados.
  - Utilize a base de dados do IDH brasil 2010 (IDH_Brasil_2010.csv)
  - Rode a regressão de acordo com o modelo abaixo:

```
mod2.ex <- lm(dados3$mort_infantil ~ dados3$renda_per_capita
        + dados3$indice_gini
        + dados3$salario_medio_mensal
        + dados3$perc_criancas_extrem_pobres
        + dados3$perc_criancas_pobres
        + dados3$perc_pessoas_dom_agua_estogo_inadequados
        + dados3$perc_pessoas_dom_paredes_inadequadas
        + dados3$perc_pop_dom_com_coleta_lixo
        + dados3$perc_pop_rural
        + as.factor(dados3$Regiao))
```

  - Questão 7: Com base nos resultados dessa nova equação, qual o efeito das regiões Norte, Sul, Nordeste e Sudeste, mesmo depois de “controlarmos” para as variáveis incluídas no modelo?

Modelos de Regressão

- **2ª Lista de exercícios para entregar em 19/11/2018(continuação):**
  - Rode agora a regressão com efeitos das Regiões sobre a mortalidade infantil:

```
mod3.ex <- lm(dados3$mort_infantil ~ dados3$renda_per_capita
        + dados3$indice_gini
        + dados3$salario_medio_mensal
        + dados3$perc_criancas_extrem_pobres
        + dados3$perc_criancas_pobres
        + dados3$perc_pessoas_dom_agua_estogo_inadequados
        + dados3$perc_pessoas_dom_paredes_inadequadas
        + dados3$perc_pop_dom_com_coleta_lixo
        + dados3$perc_pop_rural
        + as.factor(dados3$Regiao)
        + as.factor(dados3$Regiao)*dados3$renda_per_capita)
```

  - Questão 8: Com base nos resultados dessa nova equação, como o efeito da renda per capita, sobre mortalidade infantil, se altera de acordo com a macrorregião do município?
  - Questão 9: Houve uma melhora no R2 ajustado quando adicionamos os efeitos das macrorregiões sobre o coeficiente da renda per capita (mod3 versus mod2)?

Modelos de Regressão

- **2ª Lista de exercícios para entregar em 19/11/2018 (Continuação).**
  - Vamos incluir agora uma interação entre a macrorregião e a variável "perc_pessoas_dom_agua_estogo_inadequados":

```
mod3.ex <- lm(dados3$mort_infantil ~ dados3$renda_per_capita
        + dados3$indice_gini
        + dados3$salario_medio_mensal
        + dados3$perc_criancas_extrem_pobres
        + dados3$perc_criancas_pobres
        + dados3$perc_pessoas_dom_agua_estogo_inadequados
        + dados3$perc_pessoas_dom_paredes_inadequadas
        + dados3$perc_pop_dom_com_coleta_lixo
        + dados3$perc_pop_rural
        + as.factor(dados3$Regiao)
        + as.factor(dados3$Regiao)*dados3$renda_per_capita
        + as.factor(dados3$Regiao)*dados3$perc_pessoas_dom_agua_estogo_inadequados)
```

  - Questão 10: Vamos assumir que a variável "perc_pessoas_dom_agua_estogo_inadequados" seja uma variável direta de política pública. De acordo com os resultados da regressão acima, em qual região políticas de melhoria do acesso a água e esgoto seriam mais eficazes para reduzir a mortalidade infantil?

Obrigado.