Regression and Clinical prediction models

Session 6

Introducing statistical modeling – Part 2 (Correlation and Linear regression)

Pedro E A A do Brasil
pedro.brasil@ini.fiocruz.br

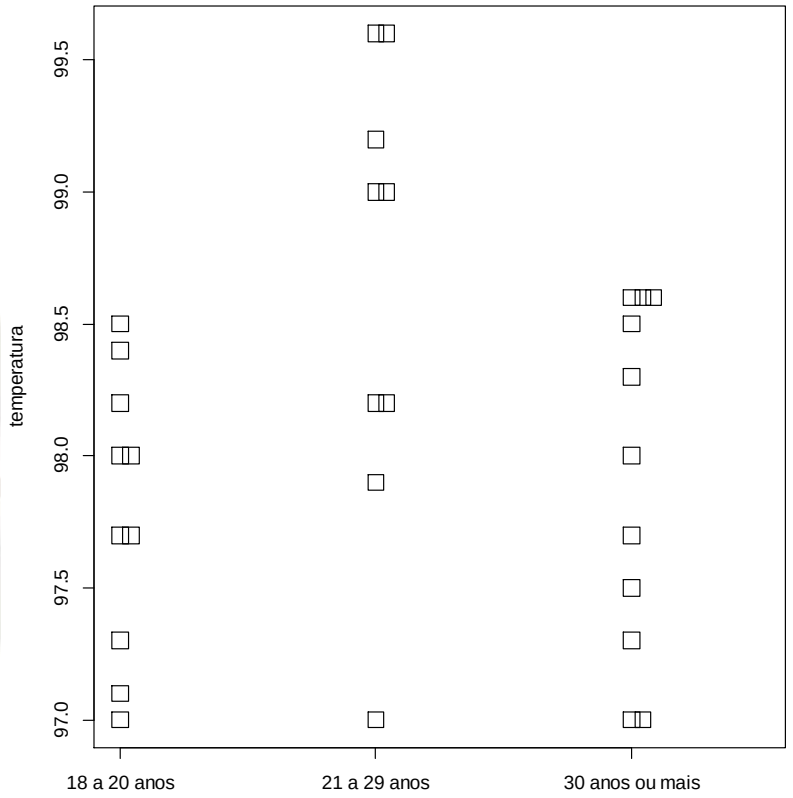
2018

Objetivos

- Continuar a introduzir conceitos e práticas em modelos estatísticos de regressão.
- Apresentar e discutir propriedades e limitações do modelo linear simples como exemplo de regressão.

Dispersão

- Quando se dispõem dados de um desfecho em escala contínua por preditor categórico (boxplot pode ser uma alternativa), uma abordagem analítica possível é a ANOVA.
- E quando o preditor também está em uma escala contínua?



Dispersão

Tabela 1: Dados de imunização e taxa de mortalidade de 11 dos 20 países amostrados

País	Porcentagem Imunização	Taxa de Mortalidade por 1000 nascidos vivos
Bolívia	77	118
Brasil	69	65
Camboja	32	184
Canadá	85	8
China	94	43
República Tcheca	99	12
Egito	89	55
Etiópia	13	208
Finlândia	95	7
França	95	9

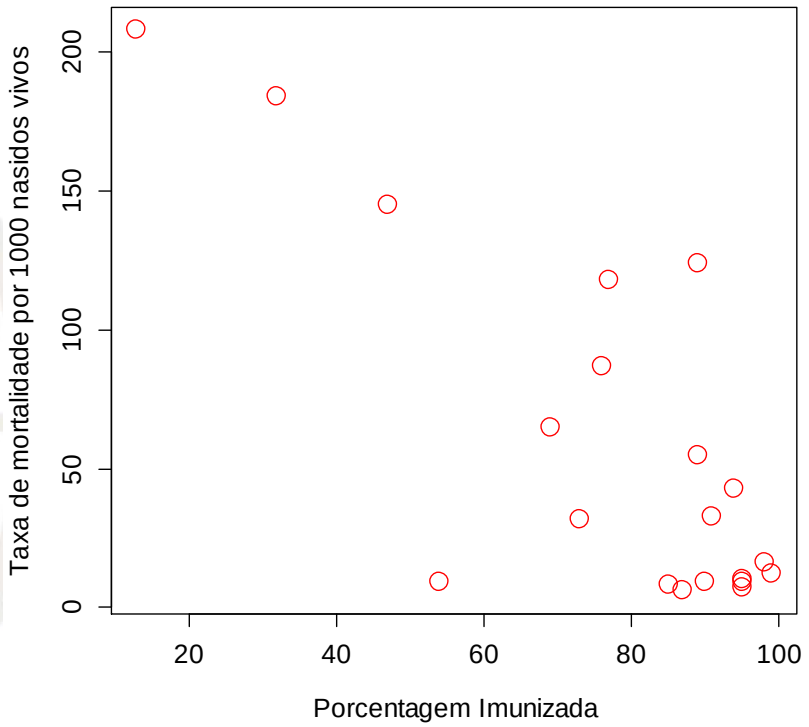
5

Dispersão

- Como relacionar o percentual de crianças imunizadas contra DPT com a taxa de mortalidade até 5 anos?
- Primeiramente, pode-se traçar um gráfico para verificar a relação entre a imunização e a taxa de mortalidade.
- Qual seria o gráfico mais apropriado a esse tipo de análise?

Dispersão

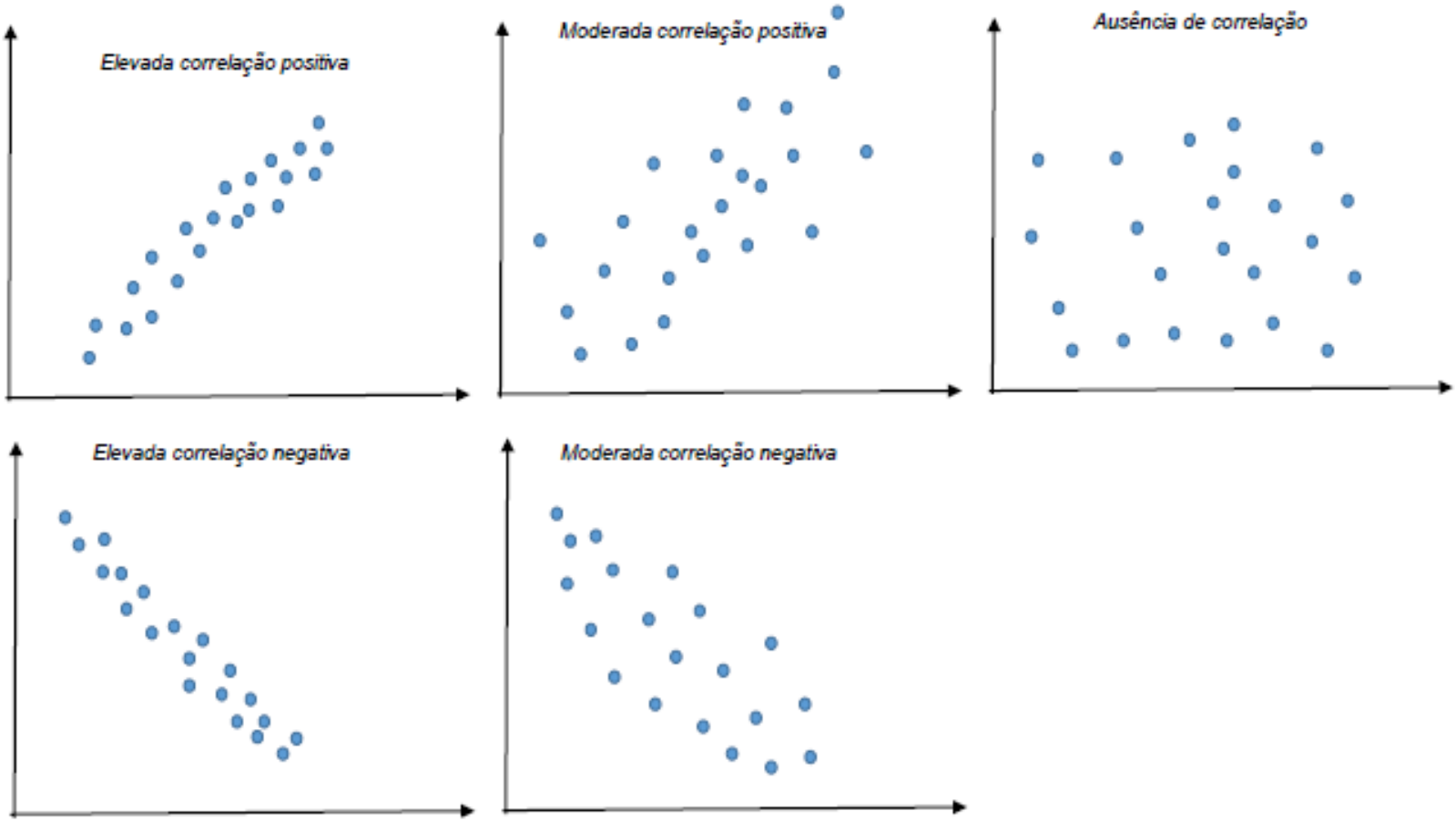
- O gráfico passa a ideia de que quanto menor a porcentagem de imunização de DPT no país, maior a taxa de mortalidade por 1000 nascidos vivos.



Dispersão

- Assim, o gráfico de dispersão pode ser entendido como uma demonstração visual (ou nuvem de pontos) relacionando as quantidades alcançadas para cada unidade (observação) na variável X (eixo horizontal) e Y (eixo vertical).
- No exemplo, cada ponto (unidade de observação) representa um país diferente da tabela acima.
- Assim, no gráfico os valores de taxa de mortalidade e cobertura vacinal de cada país correspondem a coordenadas de pontos, por exemplo o Brasil possui coordenadas X = 69 e Y = 65, ou seja, (X = 69, Y = 65).

Dispersão



Os padrões possivelmente observados entre duas variáveis X e Y.

Correlação vs Regressão

- Correlação – é o grau de relação linear de duas variáveis quantitativas. Neste caso, desconsidera-se uma relação de dependência ou preditor/desfecho na associação.
  - Exemplos: Coeficiente de correlação de Pearson, Coeficiente de correlaçao de Spearman.

- Regressão – uma variável é definida como resposta (quantitativa ou qualitativa) e usa uma ou mais variáveis como potenciais variáveis explicativas ou preditoras.
  - O objetivo da regressão é predizer o valor do desfecho (resposta) ou explicar o comportamento (variação) do desfecho (resposta).

Correlação

País	%Imunização	Taxa de Mortalidade		(Xi-X̄)	(Yi-Ȳ)	(Xi-X̄)(Yi-Ȳ)	$(Xi-\bar{X})^2$	$(Yi-\bar{Y})^2$
Etiópia	13	208		-64,4	149	-9595,6	4147,36	22201
Camboja	32	184		-45,4	125	-5675	2061,16	15625
Senegal	47	145		-30,4	86	-2614,4	924,16	7396
Grécia	54	9		-23,4	-50	1170	547,56	2500
Brasil	69	65		-8,4	6	-50,4	70,56	36
Federação Russa	73	32		-4,4	-27	118,8	19,36	729
Turquia	76	87		-1,4	28	-39,2	1,96	784
Bolívia	77	118		-0,4	59	-23,6	0,16	3481
Canadá	85	8		7,6	-51	-387,6	57,76	2601
Japão	87	6		9,6	-53	-508,8	92,16	2809
Índia	89	124		11,6	65	754	134,56	4225
Egito	89	55		11,6	-4	-46,4	134,56	16
Reino Unido	90	9		12,6	-50	-630	158,76	2500
México	91	33		13,6	-26	-353,6	184,96	676
China	94	43		16,6	-16	-265,6	275,56	256
França	95	9		17,6	-50	-880	309,76	2500
Finlândia	95	7		17,6	-52	-915,2	309,76	2704
Itália	95	10		17,6	-49	-862,4	309,76	2401
Polônia	98	16		20,6	-43	-885,8	424,36	1849
República Tcheca	99	12		21,6	-47	-1015,2	466,56	2209
MÉDIA	77,40	59,00	SOMA	0,00	0,00	-22706,00	10630,80	77498,00
DESVIO-PADRÃO	23,65	63,87						

CORRELAÇÃO= -0,79

$$r = \frac{-22706}{(20-1)\times 23{,}65\times 63{,}87}$$

Correlação

- r → +1 indica correlação linear positiva (r = 1 correlação linear positiva perfeita)
- r → -1 indica correlação linear negativa (r = -1 correlação linear negativa perfeita)
- r → 0 indica ausência de correlação linear

Correlação

- Para fazer inferência da existência de correlação para a população?
- r = -0.79 é a correlação amostral e não a populacional.
- No teste de hipótese para a correlação, as hipóteses podem ser escritas como:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

- Ou seja, essas hipóteses testam se a correlação na população é igual/diferente de zero.

Correlação

- Pulando a especificação da estatística do teste de correlação e o valor tabelado...
- Repetindo o exemplo da vacinação e mortalidade infantil:

```
> cor.test(banco$porc,banco$tm)

        Pearson's product-moment correlation

data:  banco$porc and banco$tm
t = -5.4864, df = 18, p-value = 3.281e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9137250 -0.5362744
sample estimates:
       cor
-0.7910654
```

- Existe evidência estatística suficiente ao nível de 5%, para rejeitar a hipótese nula e afirmar que a correlação na população é diferente de 0.

Correlação

- Limitações:
  - O coeficiente só quantifica o grau de relação linear (relações não monotônicas não são capturadas)
  - O coeficiente de correlação é sensível a dados extremos
  - Não pode ser extrapolado para valores além dos observados nas variáveis X e Y
  - Não permite uma interpretação de efeito de X em Y
- Alternativa não paramétrica
  - Coeficiente de relação de postos (Spearman)
    - Exatamente a mesma operação mas com postos ao invés de valores observados e suas médias.

Regressão linear

- É o mais simples modelo de regressão
- A regressão linear pretende verificar a relação linear entre uma ou mais variáveis explicativas/preditoras/independentes (que podem ser qualitativas ou quantitativas) e uma variável quantitativa (resposta/desfecho/dependente)
- A regressão linear pode ser dividida em duas estratégias de análise, de acordo com o número de variáveis envolvidas: simples ou múltipla.
- A regressão linear simples envolve apenas duas variáveis:
  - 1ª: variável resposta (ou desfecho)
  - 2ª: variável preditor (ou explicativa, ou independente)
- Assim, permite verificar para cada mudança na variável preditora, a mudança esperada na variável resposta. Com isso podemos estimar ou prever a variável resposta.

Regressão linear

- A regressão linear simples pode ser definida por uma equação com componentes determinístico e aleatório.

$$Y_i = \underbrace{B_0 + B_i X_i}_{\text{determinístico}} + \underbrace{\varepsilon_i}_{\text{aleatório}}$$

- Onde *Xi* é o valor da variável explicativa na i-ésima observação, *Yi* é o valor da variável resposta ou desfecho na i-ésima observação.
- i-ésima observação = observação sendo analisada, pode ser a primeira, segunda, ..., última (i = índice ou a posição nos dados)

Regressão linear

- Os nomes das estimativas da regressão linear diferem dos nomes dos parâmetros:

Parâmetro

$$Y_i = B_0 + B_i X_i + \varepsilon_i$$

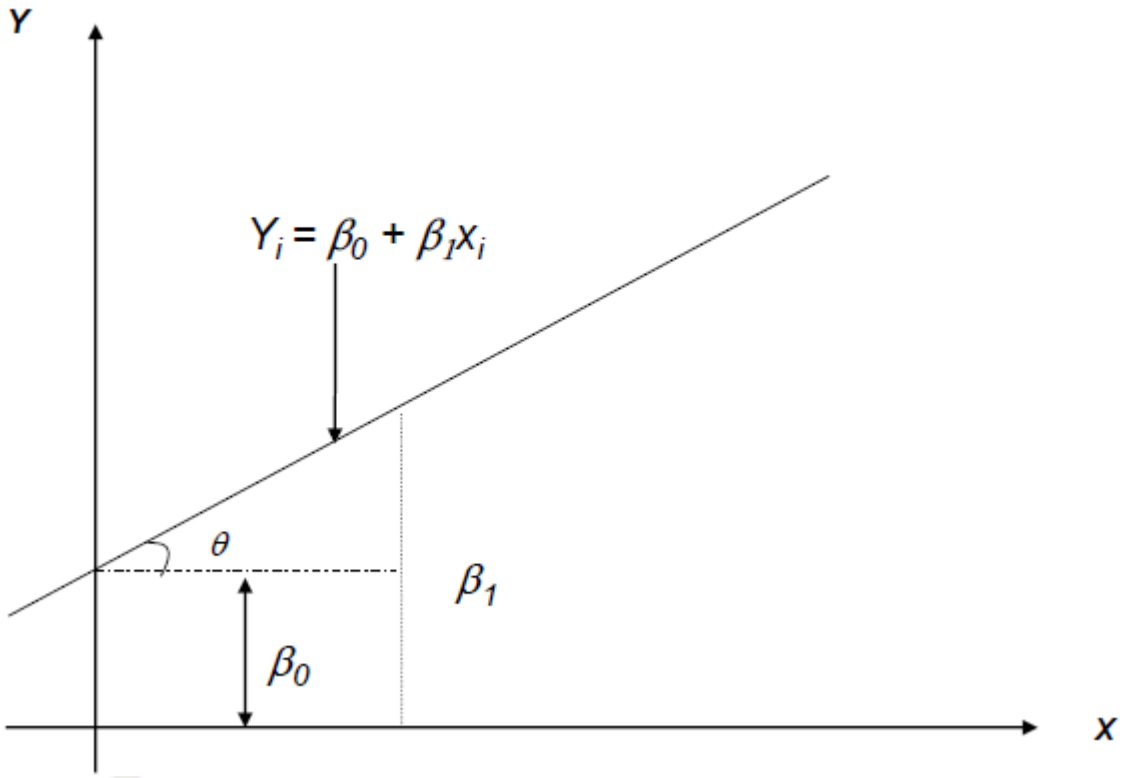
determinístico aleatório

Estimativa

$$\hat{Y}_i = b_0 + b_i x_i$$

Regressão linear

- β_0 é o intercepto, o valor onde a reta de regressão corta o eixo da resposta Y (vertical), onde x (preditor) = 0
- β_1 determina a inclinação da reta de regressão (*slope*). Mudança esperada em y para cada aumento de uma unidade em x.

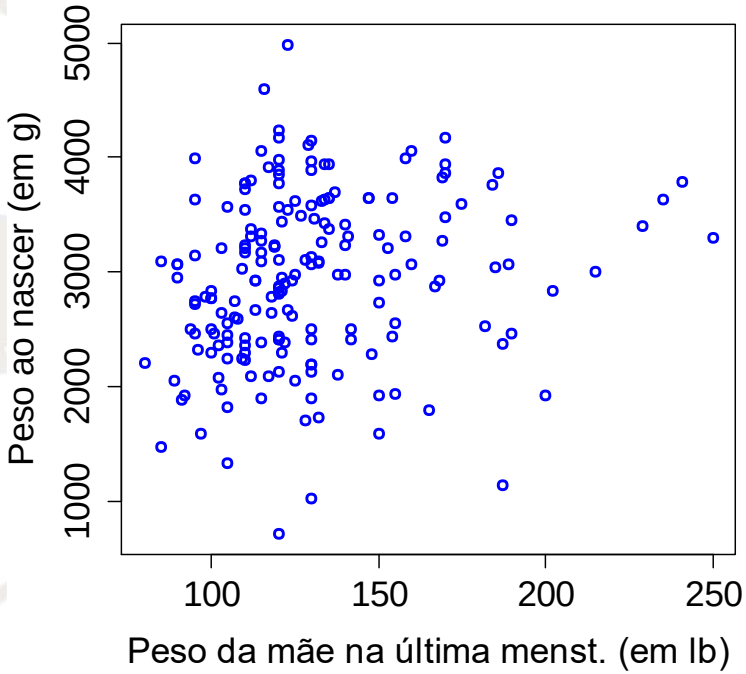


Regressão linear

- Exemplificando com um estudo conduzido para identificar preditores de baixo peso ao nascer, conduzido no *Baystate Medical Center, Springfield*, durante o ano de 1986.
- Neste estudo foram disponibilizadas as variáveis peso da mãe no último período menstrual e peso ao nascer (em gramas) para 189 nascimentos.
- Como analisar a relação entre peso da mãe e peso do bebê?
- Quais seriam as variáveis resposta e explicativa?
- Qual a melhor reta que se ajusta aos dados?

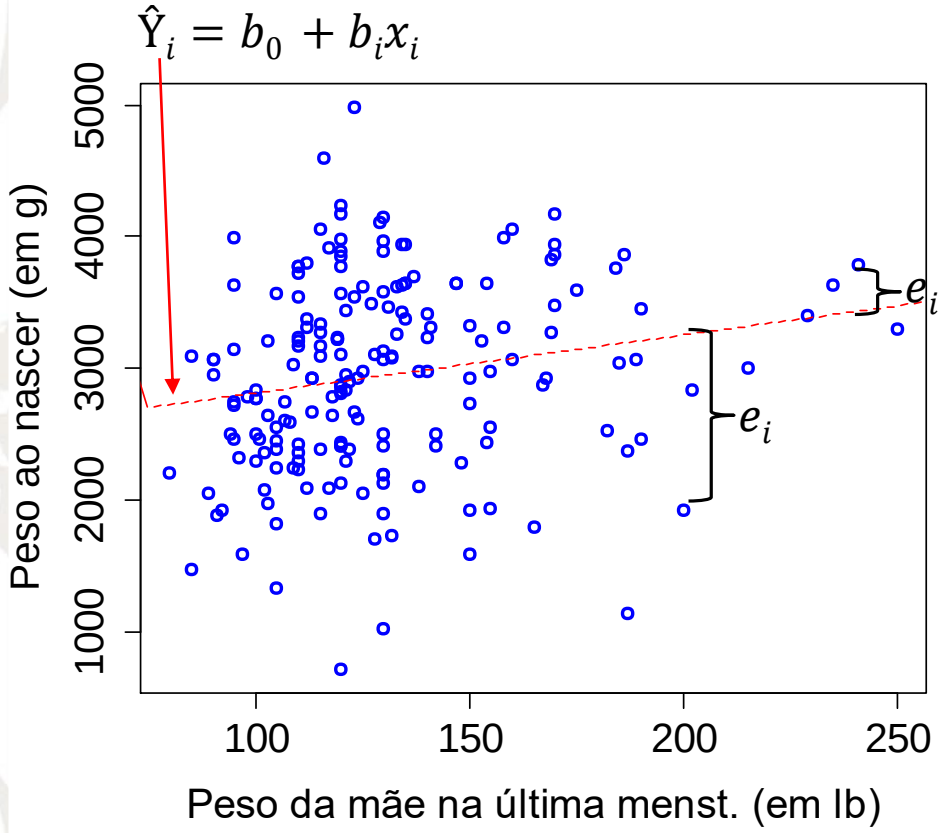
Regressão linear

- A relação pode ser investigada através de uma regressão linear simples onde o peso ao nascer é a variável resposta e peso da mãe a variável preditora.



Regressão linear

- Mesmo gráfico com a reta do melhor ajuste.



Resíduo

$$e_i = Y_i - \hat{Y}_i$$

Regressão linear

- Cálculo da tabela de análise de variância (ANOVA)
  - Através da soma dos quadrados, pode-se construir a tabela de análise de variância. Como objetivo de calcular a variabilidade do modelo e dos resíduos. (onde k é o número de coeficientes a serem estimados)

Fonte de Variação	Soma dos quadrados	gl	Quadrado médio	F
Modelo	$\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 = SQM$	k-1	QMM= SQM/k-1	QMM/QMR
Resíduo	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = SQR$	n-k	QMR= SQR/n-k	
Total	$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = SQT$	n-1	QMT= SQR/n-1	

Regressão linear

- A tabela de ANOVA testa se existe ao menos um coeficiente significativo no modelo (diferente de zero).

$$H_0: B_1 = 0$$
$$H_1: B_1 \neq 0$$

Linear simples

$$H_0: B_1 = B_2 = B_3 = B_p$$
$$H_1: B_1 \neq 0 \; ou \; B_2 \neq 0 \; ou \; B_3 \neq 0 \; ou \; B_p \neq 0$$

Linear múltiplo

- Estatística-teste (F de Snedecor): $F = \frac{QMM}{QMR}$
  - Se p-valor < α. Rejeito H_0, existe ao menos um coeficiente diferente de zero (ou significativo) no modelo.
  - Caso deseje olhar a tabela F, o valor crítico é determinado por (k-1) e (n-k) graus de liberdade.

Regressão linear

- Como achar o melhor ajuste (melhor reta)?
  - A estimação dos coeficientes do modelo, utiliza o método dos mínimos quadrados, que minimizam a soma do quadrado dos desvios da reta ajustada pelo modelo.

minimiza $$SQR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\text{Intercepto} : b_0 = \overline{Y} - b_1 \overline{X}$$

$$\text{Inclinação} : b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - n(\overline{X})^2}$$

Regressão linear

```
> summary(aov(bwt ~ lwt, data = birthwt))
             Df   Sum Sq Mean Sq F value Pr(>F)
lwt           1  3448639 3448639   6.681 0.0105 *
Residuals   187 96521017  516155
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> reg <- lm(bwt ~ lwt, data = birthwt)
> summary(reg)

Call:
lm(formula = bwt ~ lwt, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max
-2192.12  -497.97    -3.84   508.32  2075.60

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2369.624    228.493  10.371   <2e-16 ***
lwt            4.429      1.713   2.585   0.0105 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 718.4 on 187 degrees of freedom
Multiple R-squared:  0.0345,             Adjusted R-squared:  0.02933
F-statistic: 6.681 on 1 and 187 DF,  p-value: 0.0105
```

Regressão linear

- Este teste verifica se cada coeficiente é significante no modelo, ou seja, se ele é ou não igual a zero.

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

- Estatística-teste (t de student):

$$t = \frac{\beta_i - 0}{s_{\beta_i}}$$

- No caso de regressão linear simples, note que o p-valor da ANOVA será igual ao do teste do coeficiente, pois as hipóteses são iguais, o que não acontece na regressão múltipla.

Regressão linear

- Como fazer uma estimativa de um valor de peso ao nascimento do bebê dado que se sabe o peso da mãe ao início da gestação, por exemplo de 150 libras?

$$\hat{Y}_i = b_0 + b_i x_i$$

$$\hat{Y}_i = 2369.624 + 4.429\ x\ 150 = 3033.59$$

  - Inclinação:
    - Espera-se que o peso médio dos nascidos aumente em 4.429 gramas a cada aumento de 1 libra no peso da mãe na última menstruação.
  - Intercepto:
    - O peso médio em gramas dos nascidos é de 2369.624, caso X fosse igual azero.

Regressão linear

- A qualidade de ajuste pode ser medida pelo coeficiente de determinação (R2)
  - Quanto mais próximo de 1, melhor o ajuste do modelo
  - A interpretação da medida pode ser realizada em porcentagem. No geral, é difícil encontrar valores muito altos para o R2.
  - No modelo de regressão linear simples ele pode ser obtido pelo quadrado do coeficiente de correlação linear de Pearson.

$$R^2 = \frac{SQM}{SQT}$$

Regressão linear

- Pressupostos da regressão linear
  - Independência: Os valores da variável resposta são independentes.
  - Linearidade: A variável resposta é uma função linear das variáveis dependentes.
  - Homocedasticidade: Variância constante
  - Distribuição normal: Y tem distribuição normal quando condicionado aos valores de X ($\mu_{Y|X}$)
- O erro aleatório (resíduo) do modelo deve ter as seguintes propriedades:
  - Média igual a zero, variância constante e distribuição aproximadamente normal.
  - Além disso, os erros não devem estar correlacionados.
- Mais adiante, será mostrado como verificar essas propriedades através de gráficos de resíduos

fim

Session 6

Introducing statistical modeling – Part 6 (Multivariable linear regression)

Pedro E A A do Brasil
pedro.brasil@ini.fiocruz.br

2018