Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
IMD0601 - Bioestatística

Introdução a Regressão Linear

Prof. Dr. Tetsu Sakamoto
Instituto Metrópole Digital - UFRN
Sala A224, ramal 182
Email: tetsu@imd.ufrn.br

Baixe a aula (e os arquivos)

- Para aqueles que não clonaram o repositório:

```
> git clone https://github.com/tetsufmbio/IMD0601.git
```

- Para aqueles que já tem o repositório local:

```
> cd /path/to/IMD0601
```

```
> git pull
```

Objetivos da aula

- Modelos de pesquisa em saúde pública
- Revisão: Correlação
- Regressão Linear Simples e Múltipla
- Regressão Linear em R – CPOD data

Modelos de pesquisa em Saúde Pública – *Análise de Risco*

- **Estudos observacionais** em populações
- **Estudos clínicos** com intervenções

→ **Obter um estimador para o efeito do tratamento/exposição, tendo ajustado todas as outras variáveis**

- **Estudos de relação causa-efeito**

→ **Analisar os estimadores de regressão e suas relações com o resultado (*outcome*)**

- **Modelos de Predição** (prognóstico e diagnóstico)

→ **Intervir para evitar resultados não-desejáveis**

Modelos de pesquisa em Saúde Pública

- Considerações específicas sobre dados em saúde. Informações a nível do indivíduo e da população.
- O objetivo do modelo irá determinar como ele será desenvolvido, as variáveis selecionadas e o nível de precisão das relações.
- É necessário definir claramente a questão da pesquisa!

"All models are wrong but some models are usefull"

George Box (1919-2013)

VARIAVEIS	
Características	AGE GENDER PackHistory Smoking
Doença	COPD SEVERITY CAT (COPD Assessment Test)
Habilidade de caminhada	MWT1 (Six-minute walk test) MWT2 MWT1Best
Função Pulmonar	FEV1 (Forced Expiratory Volume) FEV1PRED FVC (Forced Vital Capacity) FVCPRED
Ansiedade & Depressão	HAD
Qualidade de vida	SGRQ (St. George's respiratory questionnaire)
Comorbidades	Diabetes Muscular Hypertension AtrialFib IHD (Ischemich Heart Disease)

Doença Pulmonar Crônica Obstrutiva - COPD

Questão:

A função pulmonar FEV1 e a idade AGE são bons preditores da distância percorrida MWT1Best?

Espirometria

- Exame que mede o volume (litros) e fluxo (litros/seg) pulmonar.
- Auxilia na prevenção e permite o diagnóstico e quantificação dos distúrbios ventilatórios, como transtornos obstrutivos, restritivos ou mistos.
- Fornece informações para estudos epidemiológicos, farmacológicos, da fisiopatologia respiratória, entre outros.
- Exame que exige cooperação do paciente, empregos de técnicas padronizadas aplicadas por pessoal especialmente treinado.

Espirometria - *valores mais utilizados*

Capacidade Vita Forçada - CVF ou FVC

- Volume total máximo de ar exalado com esforço máximo, após uma inspiração máxima.
- Muito dependente do esforço.
- Expressa em litros.
- Indicativo de restrição ou obstrução.
- Normal quando superior a 70-80% do predito.

Volume Expiratório Forçado 1º seg – VEF1 ou FEV1

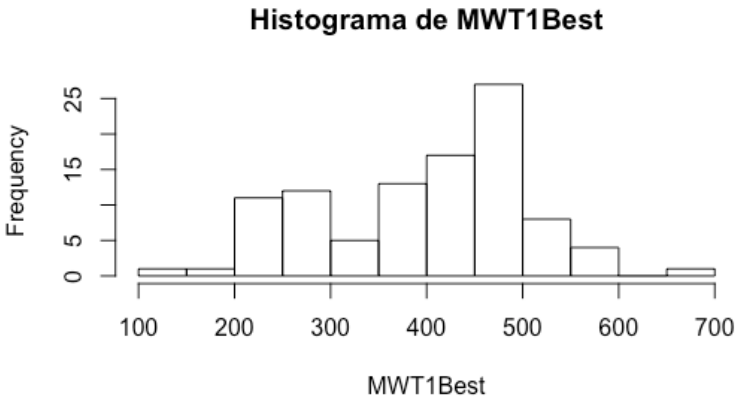
- Volume de ar exalado no 1º segundo da manobra de CVF.
- Relativamente independente do esforço.
- Expresso em litros.
- Avalia basicamente distúrbios obstrutivos.
- Obstrução importante com valores abaixo de 80% do predito.

Inicialmente... Análise Exploratória dos Dados

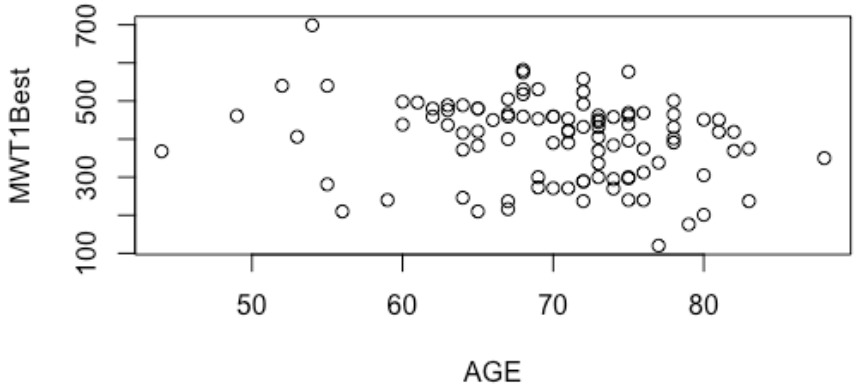
- Identificar distribuições, tendências e padrões.
- Identificar valores incomuns, impossíveis ou ausentes.
- Identificar correlações entre variáveis.
- Sugerir hipóteses a serem testadas.

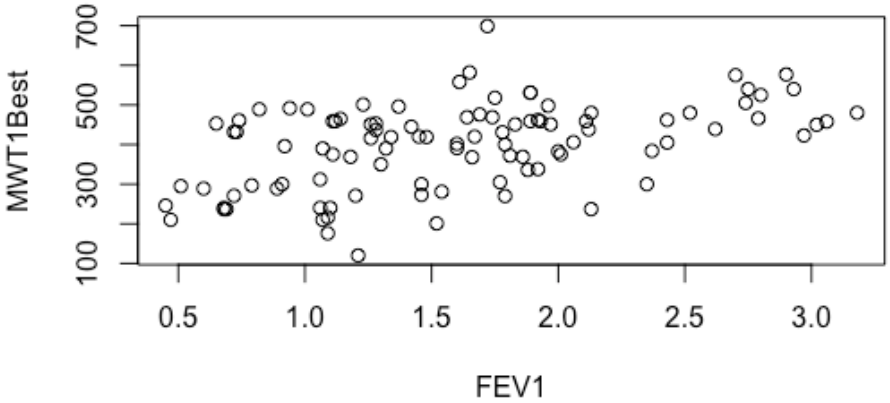
- Neste caso, todas as 3 variáveis são contínuas. Como você gostaria de examinar os dados, e quais verificações gostaria de fazer?

Iniciando a análise... EDA



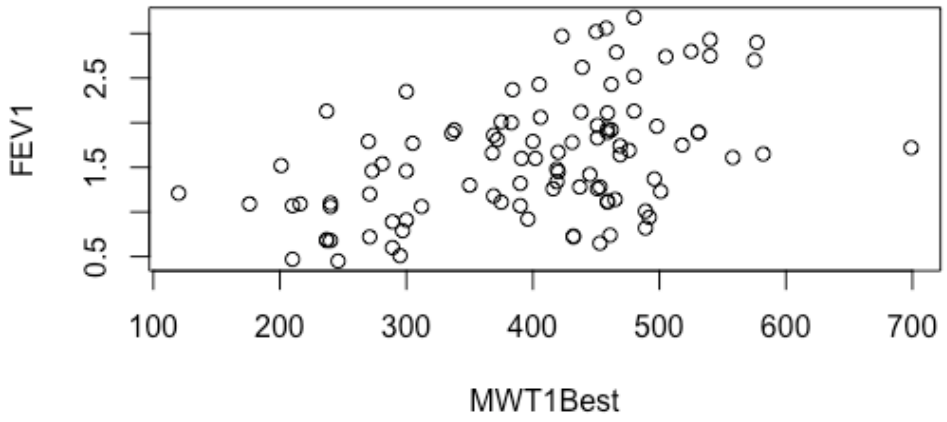

```
> hist(COPD$MWT1Best, main="Histograma de MWT1Best", xlab="MWT1Best", breaks=12)
```






```
> plot(COPD$FEV1, COPD$MWT1Best, xlab = "FEV1", ylab = "MWT1Best")
```

Correlação de Pearson



Qual das variáveis deve ser a dependente e qual a independente?

	MWT1Best	FEV1
1	120	1.21
2	176	1.09
3	201	1.52
4	210	0.47
5	210	1.07
6	216	1.09
7	237	0.69
8	237	0.68
9	237	2.13
10	240	1.06
11	240	1.10
12	240	0.68
13	246	0.45
14	270	1.79
15	271	1.20

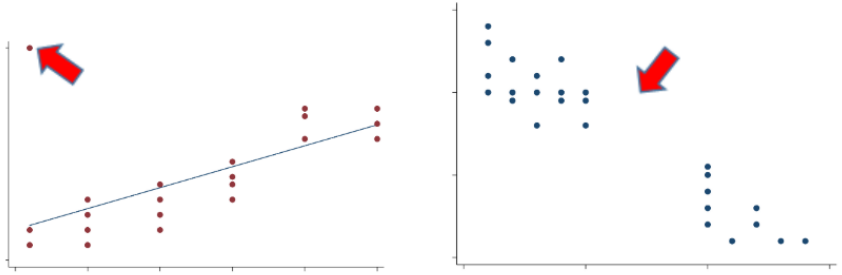
Coeficiente de Pearson

- R: medida da força da relação entre 2 variáveis contínuas.
- R: covariância em escala de -1 a +1
- Não há uma regra para interpretar o coeficiente, mas ~ 0.7 é considerado uma correlação forte.

$$r = \frac{Sxy}{Sx * Sy}$$

Premissas:

- Variáveis continuas.
- Amostras obtidas aleatoriamente.
- Distribuição próxima do normal.



Teste de Correlação de Pearson

- H0: não há correlação entre AGE e MWT1Best na população. Cor = 0
- H1: a correlação entre AGE e MWT1Best é diferente de 0.
- Cor amostra: -0.230
- IC 95%: -0.408, -0.035
- p-valor = 0.021

Correlação não implica Causa!

```
> cor.test(COPD$AGE, COPD$MWT1Best, use='complete.obs', method='pearson')

        Pearson's product-moment correlation

data:  COPD$AGE and COPD$MWT1Best
t = -2.3406, df = 98, p-value = 0.02128
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4080687 -0.0352688
sample estimates:
      cor
-0.230093
```

```
> cor.test(COPD$AGE, COPD$MWT1Best, use='complete.obs', method='spearman')

        Spearman's rank correlation rho

data:  COPD$AGE and COPD$MWT1Best
S = 211497, p-value = 0.006781
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
-0.2691106
```

Mais sobre correlação

Medem a força de associação entre 2 variáveis com valores entre -1 e +1
Exigem que as observações sejam uma amostra aleatória da população.

Pearson

- Usa valores reais para calcular a correlação.
- Assume associação linear entre as duas variáveis.
- Requer que ambas variáveis sejam contínuas.
- Exige que ambas variáveis sejam aproximadamente distribuídas normalmente.

Spearman

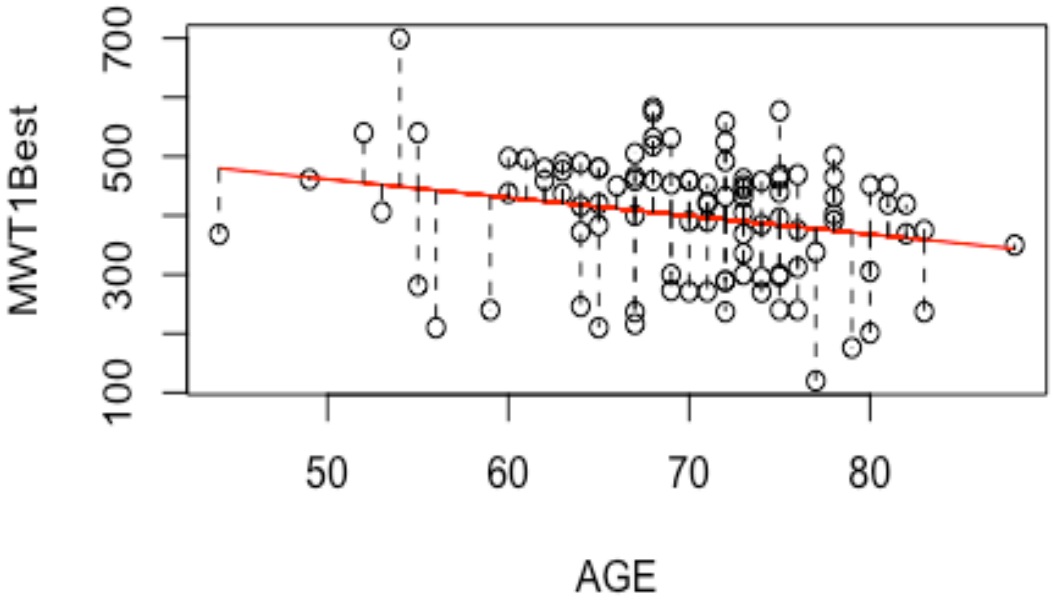
- Classifica os valores e calcula o coeficiente usando as classificações.
- Assume apenas uma relação monotônica.
- Pode ser usado para variáveis contínuas e ordinais.
- Não exige distribuição normal.

Ajuste da curva: Mínimos Quadrados (MMQ, MQO)

- Técnica de otimização matemática que procura encontrar o melhor ajuste para um conjunto de dados

- ***Minimiza a soma dos quadrados das diferenças entre o valor estimado e os dados observados***.

- Esta diferença é chamada de ***Resíduos***.



Regressão Linear

$$Y = \alpha + \beta * X + \varepsilon$$

Y = variável resultado (dependente)

X = variável preditora (independente)

α = intercepto quando X = 0

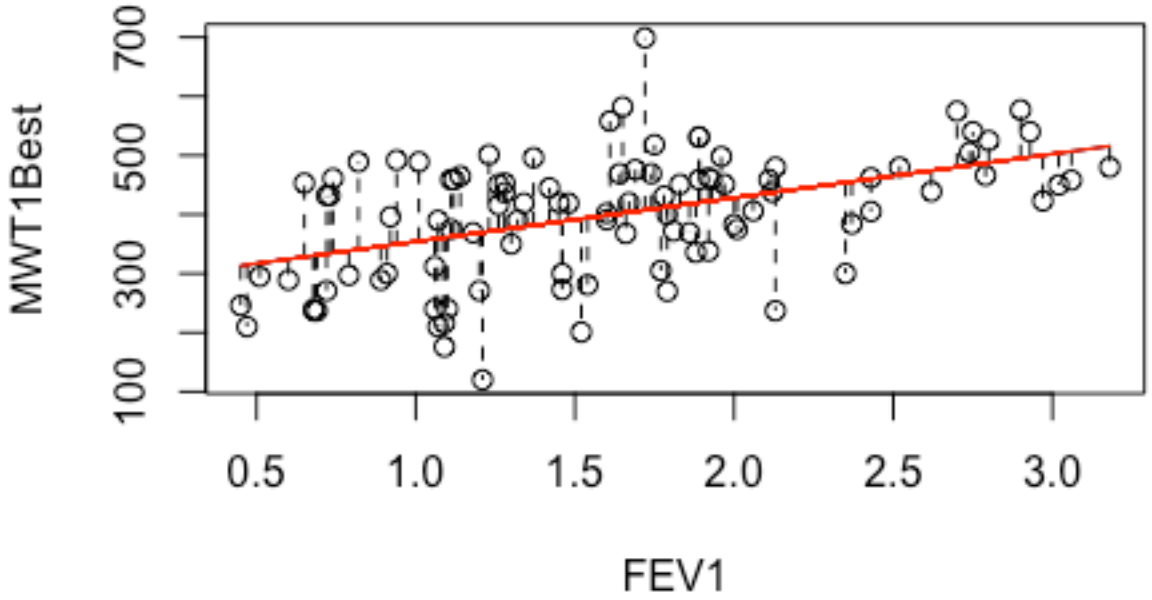
β = coeficiente linear ou inclinação da reta
(variação de Y para variação de 1 unid. de X)

ε = variação aleatória em Y, os resíduos!

Regressão Linear

- Modelo estatístico que procura estabelecer uma **relação linear** entre uma variável preditora X e uma variável resposta Y.
- Quantifica a relação ajustando uma reta aos dados que minimize a distância de todas as observações a ela.
- Método: mínimos quadrados

```
> Best_FEV1 <- lm(MWT1Best ~ FEV_1, data=COPD)
> MWT1Best = 279.92 + 74.11*FEV_1
```



Interpretando o resultado

MWT1Best = 279.92 + 74.11* FEV_1

- Escreva a equação de regressão.
- Interprete os coeficientes α e β.
- O que o IC e p-valor informam?
- Grau de liberdade: n - # var. explicativas.
- T-stats: $\beta - \beta_0$ / std.error(β)
- E quanto à estatística R2?

OBS: *Nunca use um modelo para predizer valores de variáveis fora dos limites reais!*

```
> Best_FEV1 <- lm(MWT1Best ~ FEV1, data=COPD)
> summary(Best_FEV1)

Call:
lm(formula = MWT1Best ~ FEV1, data = COPD)

Residuals:
     Min       1Q   Median       3Q      Max
-249.592  -58.227    7.881   63.551  291.612

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   279.92      24.55   11.40  < 2e-16 ***
FEV1           74.11      14.09    5.26 8.47e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 94.57 on 98 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.2202,    Adjusted R-squared:  0.2122
F-statistic: 27.67 on 1 and 98 DF,  p-value: 8.469e-07

> confint(Best_FEV1)
                2.5 %   97.5 %
(Intercept) 231.19004 328.6456
FEV1         46.15031 102.0710
```

Interpretando o resultado

- R^2: medida da variabilidade explicada pelo modelo, como proporção da variabilidade total dos dados.
- R^2 = variância explicada pelo modelo/ variância total. Valores entre 0 e 1.
- Os resíduos indicam quanta variabilidade é deixada sem explicação.

```
> Best_FEV1 <- lm(MWT1Best ~ FEV1, data=COPD)
> summary(Best_FEV1)

Call:
lm(formula = MWT1Best ~ FEV1, data = COPD)

Residuals:
     Min       1Q   Median       3Q      Max
-249.592  -58.227    7.881   63.551  291.612

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   279.92      24.55   11.40  < 2e-16 ***
FEV1           74.11      14.09    5.26 8.47e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 94.57 on 98 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.2202,    Adjusted R-squared:  0.2122
F-statistic: 27.67 on 1 and 98 DF,  p-value: 8.469e-07

> confint(Best_FEV1)
                2.5 %   97.5 %
(Intercept) 231.19004 328.6456
FEV1         46.15031 102.0710
```

Interpretando o resultado

`MWT1Best = 379.58 + 30.51*gender1`

- Escreva a equação de regressão.
- Verifique como **gender** foi codificado.
- Interprete os coeficientes α e β.
- O que o IC, p-valor e R2 informam?
- O que muda quando '**male**' é a categoria de referência?

OBS: variáveis com mais de duas categorias.
Uma categoria será a referência e todas as outras (n-1) serão comparadas a ela.

```
> Best_gender <- lm(MWT1Best ~ gender, data=COPD)
> summary(Best_gender)

Call:
lm(formula = MWT1Best ~ gender, data = COPD)

Residuals:
    Min      1Q  Median      3Q     Max
-290.09  -87.96   18.66   74.92  288.91

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   379.58      17.68  21.473   <2e-16 ***
gender1        30.51      22.10   1.381     0.17
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 106.1 on 98 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.01908,   Adjusted R-squared:  0.009073
F-statistic: 1.906 on 1 and 98 DF,  p-value: 0.1705

> confint(Best_gender)
                2.5 %    97.5 %
(Intercept) 344.50270 414.66397
gender1     -13.34038  74.36121
```

Regressão Linear - Suposicões

A relação entre o resultado y e o preditor x é aproximadamente linear.

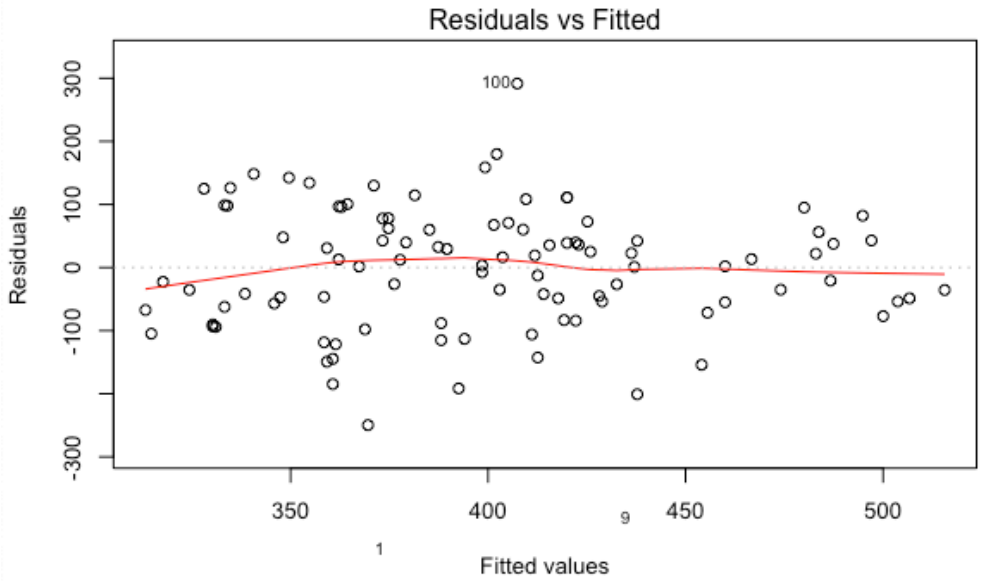
A variável dependente é normalmente distribuída entre os valores do preditor, ou seja, para cada valor do preditor, o valor da resposta é normalmente distribuída.

A variância do resultado deve ser a mesma, com relação aos valores do preditor.

Regressão Linear - Suposicões

1. **Relação linear e homegeneidade da variância.**

➢ *Distribuição dos resíduos segue uma normal com média zero e variância constante através dos valores do preditor.*

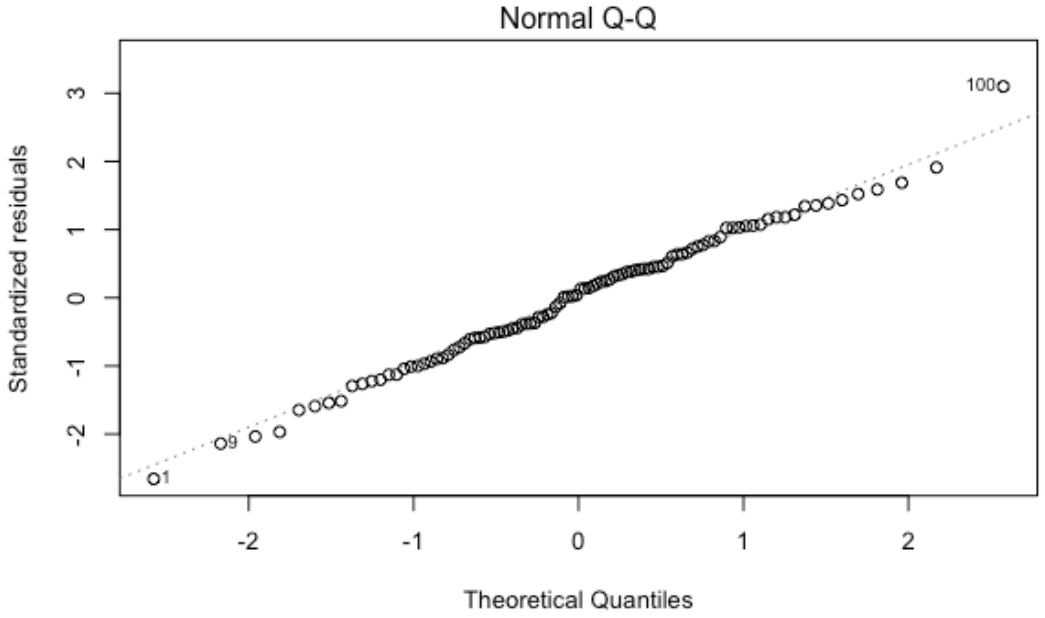



```
> plot(Best_FEV1)
```

Regressão Linear - Suposicões

2. **Resíduos seguem distribuição normal**

- ***Q-Q plot:*** *quartis dos resíduos contra os quartis da distribuição normal teórica. As observações seguem linha reta se a suposição é aceita.*

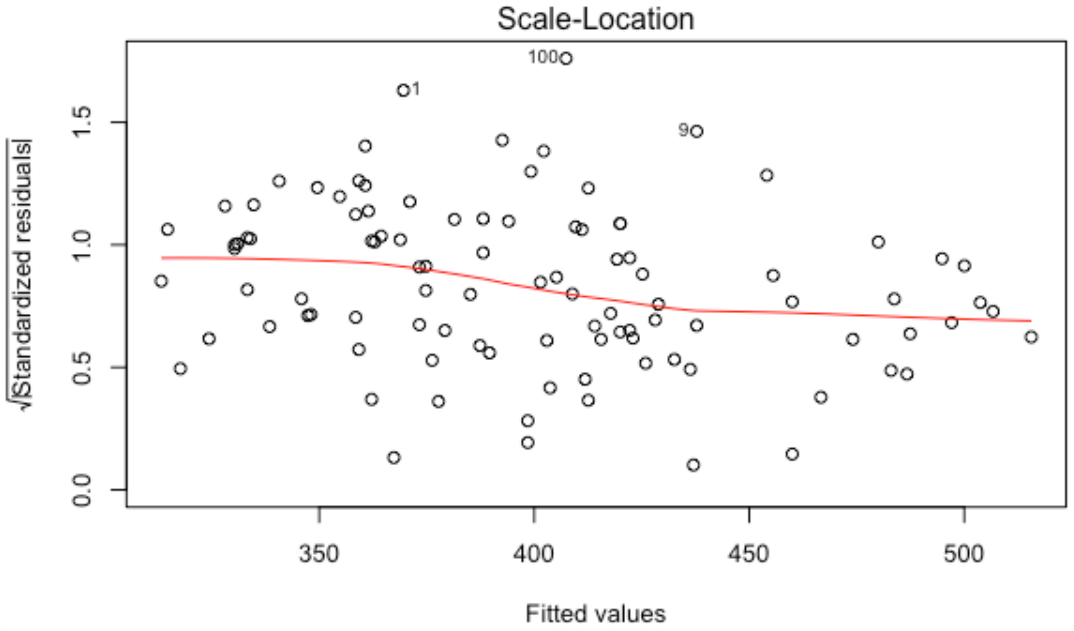



```
> plot(Best_FEV1)
```

Regressão Linear - Suposicões

3. **Homogeneidade da variância**

- *A distribuição da variância será constante através dos valores do preditor.*

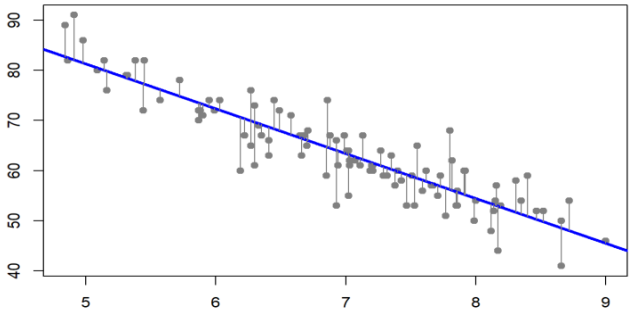



```
> plot(Best_FEV1)
```

Regressão Linear Simples *vs.* Regressão Linear Múltipla

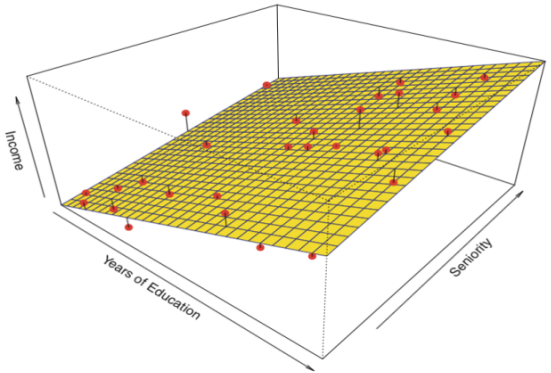
- Uma variável preditora.
- Linha que melhor se ajusta ao dados.

$$Y = \alpha + \beta1*X1 + \varepsilon$$



- Mais de uma variável preditora.
  *Supõe que o **efeito** dos preditores é **aditivo** e o modelo, uma combinação linear.*
- Plano que melhor se ajusta aos dados.

$$Y = \alpha + \beta1*X1 + \beta2*X2 + \varepsilon$$



Regressão Linear Multipla

MWT1Best = 428.1 - 7.7*Diabetic
- 72.0*AtrialFib - 130.1*DAF1

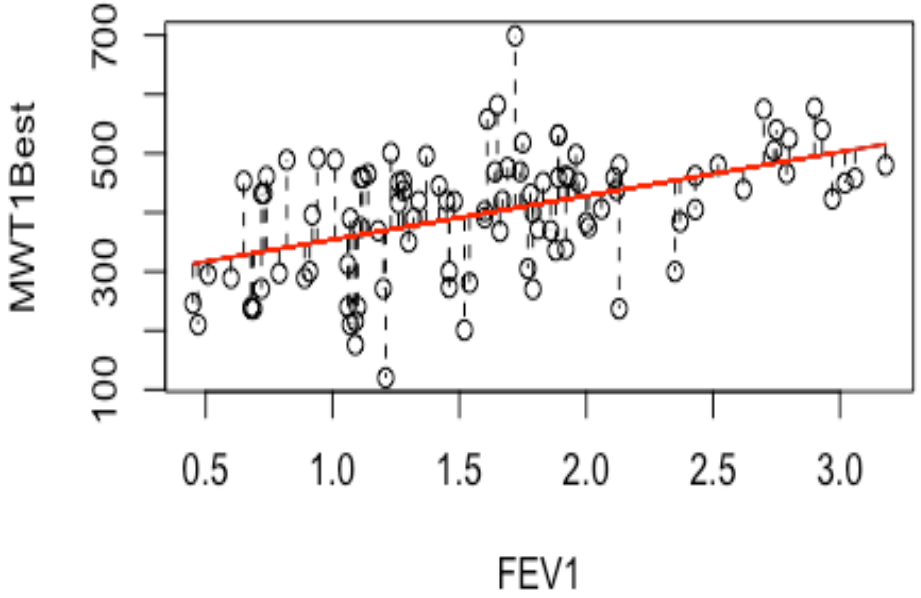
- Qual é a distância em média que uma pessoa sem diabetes e com fibrilação arterial caminha?
- E uma com ambas as comorbidades?
- O efeito de cada uma é simplesmente aditivo? Porquê?

OBS: *é possível analisar também interações entre variáveis numéricas e categóricas e entre numéricas!*

```
> Best_DAF <- lm(MWT1Best ~ factor(Diabetes)+factor(AtrialFib)+
> summary(Best_DAF)

Call:
lm(formula = MWT1Best ~ factor(Diabetes) + factor(AtrialFib) +
    factor(DAF), data = COPD)

Residuals:
    Min      1Q  Median      3Q     Max
-218.15  -51.88   18.70   51.85  270.86

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         428.14      10.39  41.200  < 2e-16 ***
factor(Diabetes)1    -7.69      28.02  -0.274  0.78436
factor(AtrialFib)1  -72.05      29.21  -2.467  0.01541 *
factor(DAF)1       -130.11      47.70  -2.727  0.00759 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 86.32 on 96 degrees of freedom
Multiple R-squared:  0.3635,    Adjusted R-squared:  0.3437
F-statistic: 18.28 on 3 and 96 DF,  p-value: 1.841e-09
```

Regressão Linear – Overfiting!

- Ocorre quando há muitas variáveis incluídas e modelo ajusta muito bem os dados para estas variáveis específicas.
- Começa a descrever o erro aleatório nos dados, ao invés das relações.
- Para comparar modelos, usar R^2 ajustado, que penaliza o número de variáveis preditoras utilizadas.
- ***Como selecionar as variáveis de maior influência?***



Seleção Automática de Variáveis

Forward Selection

- Inicia com um modelo Nulo (Null).
- Acrescenta variáveis se elas melhoram significativamente o modelo.

Backward Selection

- Inicia com um modelo Completo (Full).
- Remove as variáveis menos significas a cada iteração.

Stepwise Selection

- Inicia com um modelo Nulo.
- Acrescenta variáveis se elas melhoram significativamente o modelo.
- Remove as variáveis menos significativas a cada iteração.

Medidas estatísticas	Critérios
R-Squared	Quanto maior melhor *(> 0.70)*
R-Squared Adj	Quanto maior melhor
F-Statistic	Quanto maior melhor
AIC	Quanto menor melhor
BIC	Quanto menor melhor
RMSE root mean squared error	Quanto menor melhor
MAE Mean absolute error	Quanto menor melhor

Criando um modelo de Regressão Linear...

1. Examine as distribuições das variáveis usando estatísticas resumidas, distribuições e gráficos.
2. Examine as relações entre os preditores candidatos, usando tabulações cruzadas e correlações.
3. Examine cada uma das relações entre os preditores candidatos e variável resposta, ajustando o modelo para cada variável por vez.
4. Peque uma xícara de café e reserve um tempo para examinar e interpretar os resultados.
5. Descreva apropriadamente seu modelo de regressão.

Atenção aos erros comuns de regressão!

1. Usar a regressão para analisar uma relação não-linear.
2. Avaliar correlação como sinônimo de causalidade (corr. espúrias).
3. Não avaliar a causalidade reversa.
4. Não considerar possível viés da variável omitida.
5. Incluir variáveis explicativas colineares.
6. Explorar para além dos dados.
7. Incluir variáveis demais (mineração de dados).
8. Incluir preditores se e somente se eles forem estatisticamente significativos no nível de 5%. Se forem preditores *válidos*, devem ser incluídos independentemente de seu p-valor!

Regressão Linear em R

Referências

- Coursera ”*Introduction to Statistics & Data Analysis in Public Health*”
- Dancey, CP; Reidy JG; Rowe, R. *Estatística Sem Matemática para as Ciências da Saúde*. Porto Alegre: Penso, 2017.
- Wheelan, Charles. *Estatística: o que é, para que serve, como funciona*. Rio de Janeiro: Zahar, 2016.
- Vieira, S. *Introdução à Bioestatística*. 4ª ed. Rio de Janeiro: Elsevier, 2008.