Aula 3

Renato Rodrigues Silva

Universidade Federal de Goiás.

(2020-12-15)

Introdução

- Na aula de hoje serão vistos algumas linhas de comando referentes a medidas resumo.
- Além disso, vamos ter uma introdução a inferência estatística no R.
- Mas antes, vamos apgar todos os objetos da memória, carregar a biblioteca tidyverse e ler os dados de diabetes.

```
rm(list=ls())

library(tidyverse)

dat = read.csv("diabetes.csv", header = TRUE)
```

Introdução

- Agora vamos utilizar o que aprendemos na última aula e transformar a variável Outcome em uma variável qualitativa

```
dat = mutate(dat, Outcome = factor(Outcome))
```

- Pronto, agora podemos começar a primeira parte da aula 3.

Medidas Resumo

Como acessar vetores dentro de um data.frame

- Utlizando o símbolo $ podemos acessar um vetor dentro do data.frame.
- Exemplo queremos acessar a variável Glicose.

```
dat$Glucose
```

```
##   [1] 148  85 183  89 137 116  78 115 197 125 110 168 139 189 166 100 118
##  [19] 103 115 126  99 196 119 143 125 147  97 145 117 109 158  88  92 122
##  [37] 138 102  90 111 180 133 106 171 159 180 146  71 103 105 103 101  88
##  [55] 150  73 187 100 146 105  84 133  44 141 114  99 109 109  95 146 100
##  [73] 126 129  79   0  62  95 131 112 113  74  83 101 137 110 106 100 136
##  [91]  80 123  81 134 142 144  92  71  93 122 163 151 125  81  85 126  96
## [109]  83  95 171 155  89  76 160 146 124  78  97  99 162 111 107 132 113
## [127] 120 118 117 105 173 122 170  84  96 125 100  93 129 105 128 106 108
## [145] 154 102  57 106 147  90 136 114 156 153 188 152  99 109  88 163 151
## [163] 114 100 131 104 148 120 110 111 102 134  87  79  75 179  85 129 143
## [181]  87 119   0  73 141 194 181 128 109 139 111 123 159 135  85 158 105
## [199] 109 148 113 138 108  99 103 111 196 162  96 184  81 147 179 140 112
## [217] 109 125  85 112 177 158 119 142 100  87 101 162 197 117 142 134  79
## [235]  74 171 181 179 164 104  91  91 139 119 146 184 122 165 124 111 106
```

Medidas Resumo

- Medidas resumo reduz o conjunto de observações a respeito de uma variável em um único número (BUSSAB & MORETTIN, 1988).

Tipos de medidas de resumo

- Medidas de Tendência Central: média, moda, mediana
- Medidas de Dispersão: variância, desvio padrão.
- Quantis: 1 quartil, mediana, 3 quartil e percentis em geral.

Medidas de Tendência Central

Média aritmética

- Para calcular a média aritmética no R, basta usar a função mean.
  Exemplo: Vamos calcular a média da quantidade de glicose nos indivíduos

```
mean(dat$Glucose)
```

```
## [1] 120.8945
```

Mediana

- Mediana é o valor que ocupa a posição central do conjunto de dados ordendos. Ou seja, separa a metade dos dados absixo dela e metade acima. Exemplo: Vamos calcular a mediana dos valores de

```
median(dat$Glucose)
```

```
## [1] 117
```

Medidas de Tendência Central

Mediana Revisitada

- Para entendermos o conceito um pouco melhor, vamos usar um conjunto de dados bem pequeno

```
x = c(1,5,6,8,9)
```

- Observe que o valor 6 separa os dados em 50% inferiores $(1, 5)$ e 50% superiores $(8, 9)$. Portanto, ele é a mediana. Caso, tivessem um número par de observações a mediana seria o ponto médio entre os valores que separa a metade inferior da superior.

```
median(x)
```

```
## [1] 6
```

Medidas de Dispersão

Variância

- É uma medida que mensura a proximidade das observações com relação a média.
- Para entendermos melhor o conceito vamos usar 4 conjunto de dados pequenos

```
x = c(3,4,5,6,7)

y = c(1,3,5,7,9)

z = c(5,5,5,5,5)
```

```
mean(x)
var(x)
mean(y)
var(y)
mean(z)
var(z)
```

dados	media	variancia
x	5	2.5
y	5	10.0
z	5	0.0

Medidas de Dispersão

Variância

- Vamos calcular a variância da quantidade de glicose

```
var(dat$Glucose)
```

```
## [1] 1022.248
```

Desvio Padrão

- O desvio padrão é raiz quadrada da variância. A sua utilidade é estar na mesma unidade da variável.
- Vamos calcular o desvio padrão da quantidade de glicose

```
sd(dat$Glucose)
```

```
## [1] 31.97262
```

Medidas resumo

- A função summary calcula automaticamente algumas medidas resumo de interesse, você pode aplicar ela no vetor ou no data.frame

```
summary(dat)
```

```
##   Pregnancies        Glucose      BloodPressure    SkinThickness
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##     Insulin           BMI        DiabetesPedigreeFunction      Age
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780           Min.   :21.00
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437           1st Qu.:24.00
##  Median : 30.5   Median :32.00   Median :0.3725           Median :29.00
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719           Mean   :33.24
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200           Max.   :81.00
##  Outcome
##  0:500
##  1:268
```

Inferência Estatística

Inferir

- Inferir significa: deduzir, concluir por inferência ou por dedução, geralmente partindo de indícios, de fatos ou de raciocínios (Dicionário Online de Português, 2017?).
- A inferência utiliza raciocínio dedutivo, ou seja, o raciocínio é feito do particular para o geral.

Definições:

- A inferência estatística é um processo de inferir características de uma **população** por meio da observação de uma **amostra**.
- A população é o conjunto de todos os elementos ou resultados sob investigação.
- Amostra é qualquer conjunto da população.

Alguns comentários

- Os estatísticos fazem algumas pressuposições sobre a população:

a. A população é infinita.

b. A frequência (densidade) da população é modelada por uma distribuição de probabilidade com parâmetros desconhecidos.

- Parâmetros: Quantidades que representam característica da **população** e que são **desconhecidos.**
- Estimadores: Quantidades calculadas a partir da **amostra**.

Exemplo: Associações entre variáveis relacionadas ao diabetes em mulheres indígenas do povo Pima

- Nesse exemplo, podemos considerar que o conjunto de todos os índices de massa corporal de todas as mulheres índigenas do Povo Pima é a população.
- Por outro lado, os dados que temos em nosso data.frame é apenas um subconjunto da população, portanto uma amostra.
- O objetivo da inferência é a partir da amostra tirar conclusões a respeito da população.
- **Importante !!!**: Na vida real, por muitas vezes, a população é finita, mas é inviável a mensuração de todos os elementos da amostra.
- Exemplo: pesquisa eleitoral.

Intervalo de confiança

- Acredito que a partir desse ponto, todos nós concordamos que uma amostra é parte de uma população.
- Ainda, concordamos que um dos objetivos de estimar a média amostral é ter um valor que represente a média da população que é desconhecida. Agora pense o seguinte:
- Se pesquisadores distintos tomarem amostras da população, é natural considerar que elas serão distintas, correto?
- Sendo assim, uma pergunta surge: Se nós calcularmos a média a partir de uma dessas amostra. Qual seria a precisão dessa estimativa ?

Intervalo de confiança

- Pensando sob essa perspectiva, talvez seja melhor termos um intervalo de valores que possa conter ou não o verdadeiro da média populacional, correto?
- Essa é o conceito de um intervalo de confiança.
- Um intervalo de confiança para média populacional é um intervalo que pode conter ou não o valor da média populacional dada uma certa probabilidade.
- O intervalo de confiança para média populacional nos fornecerá uma ideia de precisão da estimativa.
- Pois, intervalos com pequena amplitude, significa que há uma grande precisão na estimativa. Enquanto que intervalos com amplitudes maiores significa que há baixa precisão.

Intervalo de confiança no R

- Vamos calcular o intervalo de confiança de 95% de confiança para a média do índice de massa corporal.
- Para isso utiliza-se a função t.test

```
t.test(dat$BMI, conf.level=0.95)$conf.int
```

```
## [1] 31.43410 32.55106
## attr(,"conf.level")
## [1] 0.95
```

- Interpretação: O intervalo construído tem 95% de probabilidade de conter a média populacional.
- Veja que essa afirmação é muito diferente de dizer que a probabilidade da média populacional estar entre 31,4 e 32,55 seja 95%. Essa última afirmação está errada.

Hipótese Estatística

- Uma hipótese estatística é uma afirmação acerca dos parâmetros da população. Exemplo: A média populacional do IMC é igual a 20.

Teste de Hipóteses

- Um teste de hipótese é uma **regra de decisão** para rejeitar ou não uma hipótese nula.

Hipótese Nula e Alternativa

- Hipótese nula é a hipótese a ser testada. A hipótese complementar a nula é a hipótese alternativa.

Teste de Hipóteses

- Uma vez que um teste de hipótese é uma regra de decisão, esta regra é passível de erros.
- **Erro tipo I**: Rejeitar a hipótese nula, quando a hipótese nula é verdadeira.
- **Erro tipo II**: Não rejeitar a hipótese nula, quando a hipótese nula é falsa.
- Importante, não é possível minimizar os dois erros simultaneamente. A solução encontrada pelos estatísticos é fixar o erro tipo I e minimizar o erro tipo II.
- Por convenção, utiliza-se erro tipo I igual a 5%. O erro tipo I é também chamado de nível de significância.

Teste de Hipóteses

- Agora precisamos de um critério para rejeitar ou não a hipótese nula sujeita ao erro tipo I. Usaremos o valor-p.

Valor-p

- Um valor p é uma medida da probabilidade de que uma diferença observada possa ter ocorrido apenas por acaso.
- Um valor p menor significa que há evidências mais fortes a favor da hipótese alternativa.
- Na prática, usamos a seguinte regra: se o valor p for menor que 0.05 (nível de significância), vamos rejeitar a hipótese nula.
- Explicações mais detalhadas sobre valor p pode ser visto nesta aula.

Teste de Hipóteses para média populacional no R

- Usamos a função `t.test` para aplicarmos teste de hipóteses.
- Suponha que desejamos testar que a média populacional do IMC seja igual a 25.
- Podemos fazer da seguinte forma:

```
t.test(BMI ~ 1, mu=25,data=dat)
```

```
##
##     One Sample t-test
##
## data:  BMI
## t = 24.579, df = 767, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 25
## 95 percent confidence interval:
##  31.43410 32.55106
## sample estimates:
## mean of x
##  31.99258
```

- Voce pode usar a seguinte sintaxe também: `t.test(dat$BMI, mu=25,data=dat)`.

Teste de Hipóteses - Inferência para duas médias (amostras independentes)

- Vamos testar se a média do IMC das mulheres que testaram positivo para diabetes é igual
  a média do IMC para aquelas que testaram negativo.
- Aqui, percebemos que o IMC das mulheres com diabetes foram extraídos independentemente da mulheres saudáveis, por isso vamos considerar que as amostras são independentes.

Pressuposições dos testes

- A população de cada grupo pode ser representada por distribuição normal.
- As variâncias dos duas populações são iguais.
- Alternativamente, podemos considerar que essas variâncias são diferentes.

Teste de Hipóteses - Igualdade de Variâncias

- Podemos notar que o primeiro passo para fazer o teste de média entre dois grupos com amostras independentes.
- Logo, precisamos fazer o teste de igualdade de variâncias.

```
var.test(BMI ~ Outcome, data = dat,
         alternative = "two.sided")
```

```
##
##     F test to compare two variances
##
## data:  BMI by Outcome
## F = 1.121, num df = 499, denom df = 267, p-value = 0.295
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9050986 1.3790088
## sample estimates:
## ratio of variances
##           1.121007
```

Teste de Hipóteses - Inferência para duas médias (amostras independentes)

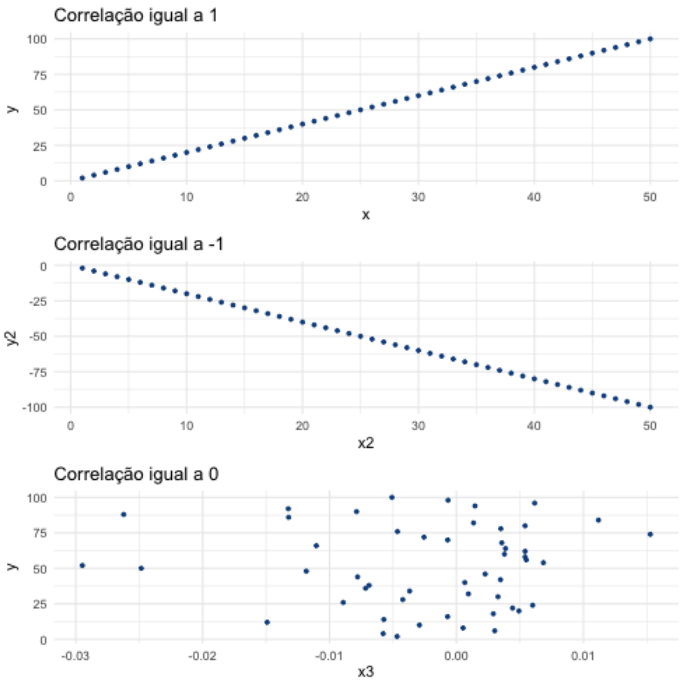
- Uma vez que não se rejeita a hipótese nula de igualdade de variâncias, procede-se o teste de comparação de médias de dois grupos da seguinte forma:

```
t.test(BMI ~ Outcome, data = dat,
       paired = FALSE, var.equal = FALSE,
       alternative = "two.sided")
```

```
##
##     Welch Two Sample t-test
##
## data:  BMI by Outcome
## t = -8.6193, df = 573.47, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.940864 -3.735811
## sample estimates:
## mean in group 0 mean in group 1
##        30.30420        35.14254
```

Análise de Correlação

- Correlação de Pearson (r), mede a dependência linear entre duas variáveis quantitativa contínua (x, y).
- A correlação assume valores entre -1 á 1.
- Podemos fazer um teste de correlação paramétrica para verificar a hipótese se a correlação é



Análise de Correlação no R

- Para fazer análise de correlação no R, vamos usar a função `cor`
- Aqui vamos calcular a correlação entre todas variáveis quantitativas contínuas no conjunto de dados de diabetes: Glucose, BloodPressure , SkinThickness , Insulin , Age ,

```
vars = c( "Pregnancies", "Outcome" )

datq = select(dat, -one_of(vars))

cor(datq)
```

Análise de Correlação no R

```
##                             Glucose BloodPressure SkinThickness     Insuli
## Glucose                  1.00000000    0.15258959    0.05732789  0.3313571
## BloodPressure            0.15258959    1.00000000    0.20737054  0.0889333
## SkinThickness            0.05732789    0.20737054    1.00000000  0.4367825
## Insulin                  0.33135711    0.08893338    0.43678257  1.0000000
## BMI                      0.22107107    0.28180529    0.39257320  0.1978590
## DiabetesPedigreeFunction 0.13733730    0.04126495    0.18392757  0.1850709
## Age                      0.26351432    0.23952795   -0.11397026 -0.0421629
##                                 BMI DiabetesPedigreeFunction         Age
## Glucose                  0.22107107              0.13733730  0.26351432
## BloodPressure            0.28180529              0.04126495  0.23952795
## SkinThickness            0.39257320              0.18392757 -0.11397026
## Insulin                  0.19785906              0.18507093 -0.04216295
## BMI                      1.00000000              0.14064695  0.03624187
## DiabetesPedigreeFunction 0.14064695              1.00000000  0.03356131
## Age                      0.03624187              0.03356131  1.00000000
```

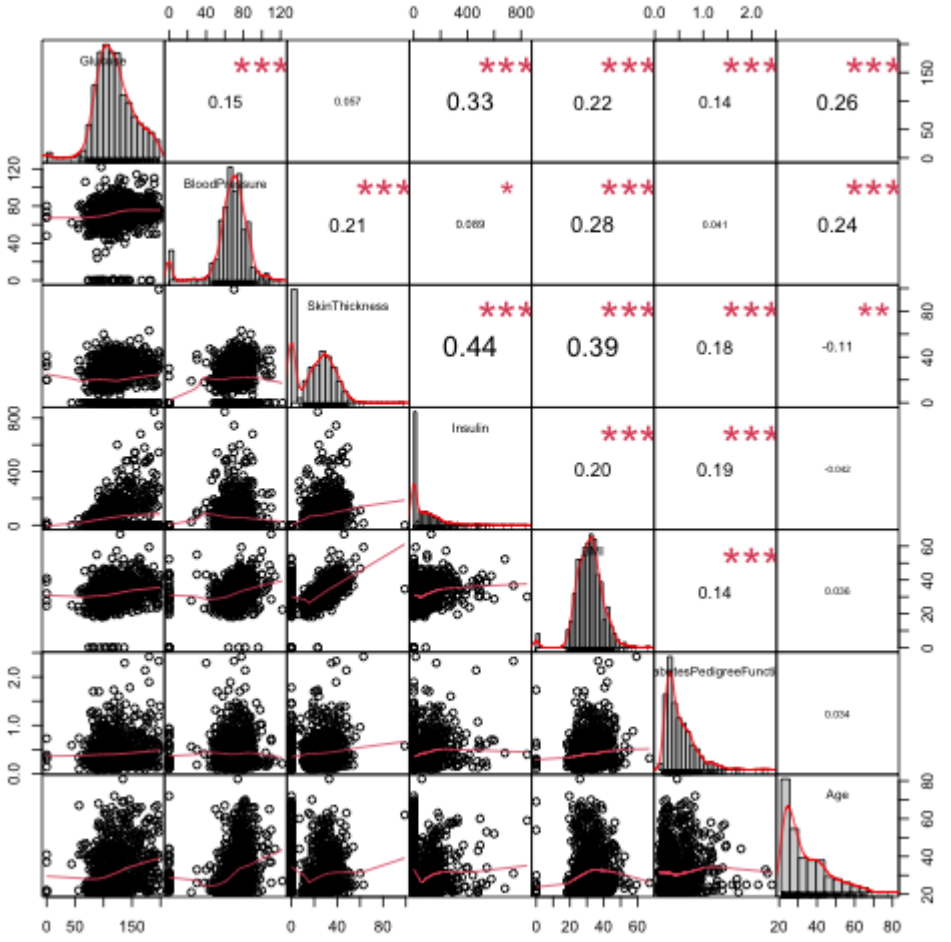
Gráfico com análise de correlação no R

- Se instalarmos a biblioteca `PerformanceAnalytics`, podemos fazer alguns gráficos interessantes com a função `chart.Correlation`

```
#install.packages("PerformanceAnalytics")
library("PerformanceAnalytics")


chart.Correlation(
  datq ,
  histogram = TRUE,
  method = c("pearson")
)
```

Gráfico com análise de correlação no R



Regressão Linear Simples

- Modelos de Regressão são indicados para verificar a relação entre uma variável resposta e uma ou mais variáveis explicativas.
- Um modelo de regressão linear modela a relação entre uma variável resposta e uma variável explicativa na forma $y = ax + b$.

Rgressão Linear Simples no R

- Podemos fazer um modelo de regressão linear simples no R com o comando `lm`.
- O comando `summary` fornece informações a respeito do R2, teste F para regressão, estimativa dos coeficientes, teste de hipótese para coeficientes, entre outras coisas.
- Exemplo : Queremos saber a relação entre quantidade de espessura da pele e IMC.

Regressão Linear Simples no R

```
mod = lm(Glucose ~ Insulin, data = dat)

summary(mod)
```

```
##
## Call:
## lm(formula = Glucose ~ Insulin, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -115.673  -21.231   -3.559   17.441   85.441
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.136e+02  1.325e+00   85.69   <2e-16 ***
## Insulin     9.193e-02  9.458e-03    9.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.19 on 766 degrees of freedom
## Multiple R-squared:  0.1098,    Adjusted R-squared:  0.1086
## F-statistic: 94.48 on 1 and 766 DF,  p-value: < 2.2e-16
```

Gráfico de Regressão Linear Simples no R

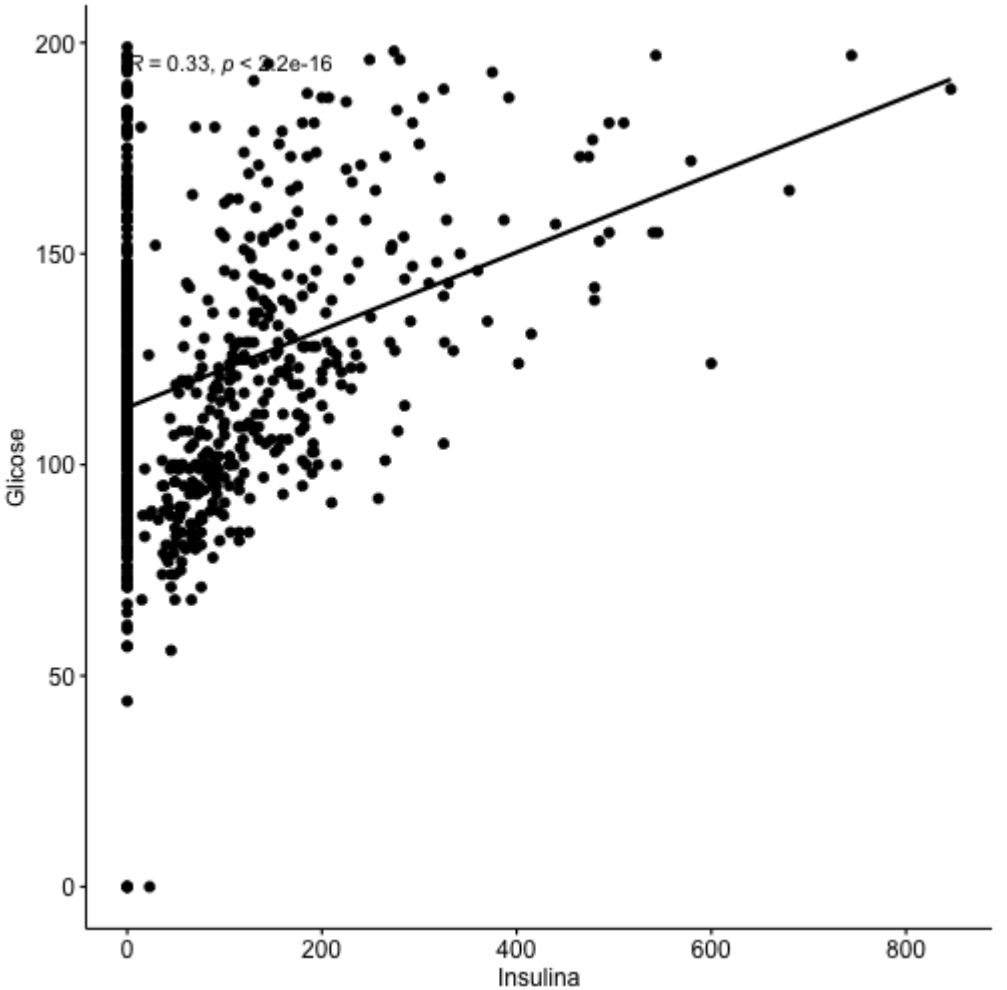
- Podemos fazer um gráfico bem completo sobre regressão linear simples no R com a biblioteca ggpubr.
- Nesse gráfico podemos ter informações sobre o ajuste da reta, a raiz quadrada do R2

```
#install.packages("ggpubr")

library("ggpubr")

ggscatter(dat, x = "Insulin", y = "Glucose",
          add = "reg.line", conf.int = FALSE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Insulina", ylab = "Glicose")
```

Gráfico de Regressão Linear Simples no R



Regressão Linear Múltipla no R

- Modelos de Regressão múltipla é quando se tem mais de uma variável explicativa.
- Podemos fazer regressão múltipla no R, usando o mesmo comando `lm`.
- Exemplo: Modelar a Glicose em função da Insulina e do Resultado do teste.

Regressão Linear Múltipla no R

```
mod = lm(Glucose ~ Outcome + Insulin, data = dat)

summary(mod)
```

```
##
## Call:
## lm(formula = Glucose ~ Outcome + Insulin, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -133.599  -17.693   -2.411   15.873   89.271
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.047e+02  1.339e+00  78.192   <2e-16 ***
## Outcome1    2.887e+01  2.057e+00  14.036   <2e-16 ***
## Insulin     7.633e-02  8.512e-03   8.967   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.94 on 765 degrees of freedom
## Multiple R-squared:  0.2921,    Adjusted R-squared:  0.2903
## F-statistic: 157.8 on 2 and 765 DF,  p-value: < 2.2e-16
```

Referências

- WICKHAM, H.; GROMULRMUND, G. R for Data Science, 2017. O'Reilly Media. Disponível em: https://r4ds.had.co.nz/explore-intro.html. Acesso em: 26 de nov. de 2020.
- VIEIRA, S. Bioestatística. Tópicos Avançados. 4 edição. 2018.
- BUSSAB, W. O.; MORETTIN, P. A. – Estatística Básica. Atual Editora, São Paulo, 1988.
- Dicionário Online de Português. [2017?]. Disponível em: https://www.dicio.com.br/inferir/. Acesso em: 03 de dez. de 2020.