Modelos estatísticos: Fundamentos e Aplicações em Genética Quantitativa

Fabricio de Almeida Silva
Universidade Estadual do Norte Fluminense Darcy Ribeiro
Programa de Pós-Graduacão em Biotecnologia Vegetal

Contents

1 Distribuições estatísticas

As distribuições estatísticas são funções matemáticas que podem ser usadas para predizer as probabilidades associadas a cada possível destino de uma variável aleatória. As distribuições podem ser **discretas** ou **contínuas**, e conhecê-las é fundamental para entender os modelos estatísticos. Isso porque os modelos estatísticos possuem diversas premissas, dentre elas a distribuição teórica esperada para a variável estudada. Nas seções abaixo, exploraremos brevemente as principais distribuições de probabilidade e exemplos de aplicações nas Ciências da Vida.

1.1 Distribuição Binomial

A distribuição binomial descreve a probabilidade de observarmos $S = k$ sucessos em N tentativas (Love 2016). Ou seja:

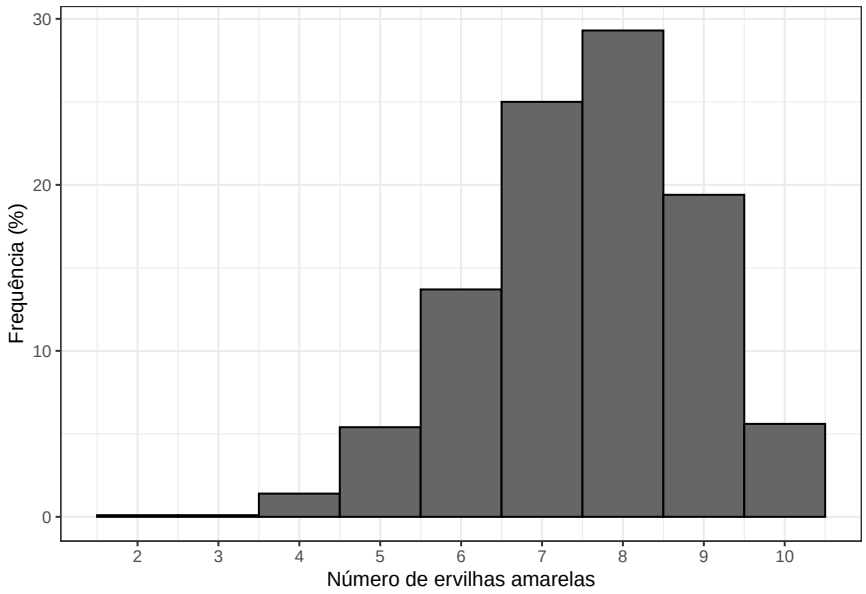
$$\Pr(S = k) = \binom{N}{k} p^k (1-p)^{N-k}$$

sendo p a probabilidade de sucesso. Por exemplo, imagine que temos uma população de 1000 ervilhas da geração F2 do experimento de Mendel. O locus V possui dominância completa e determina a cor da ervilha, sendo a cor amarela uma característica dominante. Nesse cenário, ao amostrarmos 10 ervilhas ao acaso, com que frequência poderíamos observar k ervilhas amarelas?

```r
p <- 0.75
N <- 10
k <- 0:N
tabela_dbinom <- data.frame(
    Numero_amarelas = k,
    Probabilidade = dbinom(k, size = N, prob = p)
)
tabela_dbinom$Porcentagem <- round(tabela_dbinom$Probabilidade * 100, 3)
gt::gt(tabela_dbinom)
```

Numero_amarelas	Probabilidade	Porcentagem
0	0.0000009536743	0.000
1	0.0000286102295	0.003
2	0.0003862380981	0.039
3	0.0030899047852	0.309
4	0.0162220001221	1.622
5	0.0583992004395	5.840
6	0.1459980010986	14.600
7	0.2502822875977	25.028
8	0.2815675735474	28.157
9	0.1877117156982	18.771
10	0.0563135147095	5.631

Visualizando a distribuição binomial para esse exemplo:



O histograma acima apresenta com que frequências (em porcentagem) encontraríamos cada número de ervilhas amarelas amostrando 10 ervilhas de uma população de 1000. Note que a probabilidade de obtermos 0 ou apenas 1 ervilha amarela é tão baixa que o número sequer aparece no histograma.

1.2 Distribuição de Poisson

Quando a probabilidade de sucesso p é muito pequena e o número de tentativas N é muito grande, a distribuição binomial pode ser aproximada pela distribuição de Poisson com taxa $\lambda = np$ (Holmes and

Huber, n.d.). As probabilidades de uma distribuição de Poisson podem ser obtidas com:

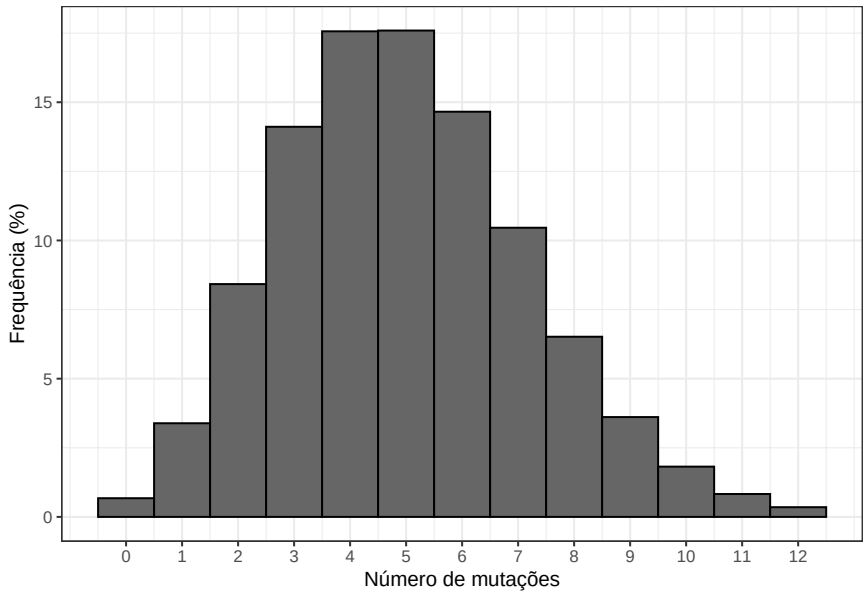
$$\Pr(S = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Por exemplo, assuma que mutações no genoma do vírus do mosaico da soja (10 kbp) ocorram aleatoriamente com taxas de 5×10^{-4} por nucleotídeo por ciclo de replicação. Isso significa que o número de mutações no genoma do vírus segue uma distribuição de Poisson com taxa 5, ou seja, após um ciclo de replicação, o número de mutações é próximo de 5. Qual é a probabilidade de obtermos de 0 a 12 mutações por ciclo de replicação nesse genoma?

```
lambda <- 5
k <- 0:12
tabela_dpois <- data.frame(
    Numero_amarelas = k,
    Probabilidade = dpois(k, lambda)
)
tabela_dpois$Porcentagem <- round(tabela_dpois$Probabilidade * 100, 3)
gt::gt(tabela_dpois)
```

Numero_amarelas	Probabilidade	Porcentagem
0	0.006737947	0.674
1	0.033689735	3.369
2	0.084224337	8.422
3	0.140373896	14.037
4	0.175467370	17.547
5	0.175467370	17.547
6	0.146222808	14.622
7	0.104444863	10.444
8	0.065278039	6.528
9	0.036265577	3.627
10	0.018132789	1.813
11	0.008242177	0.824
12	0.003434240	0.343

Visualizando a distribuição de Poisson para esse exemplo:

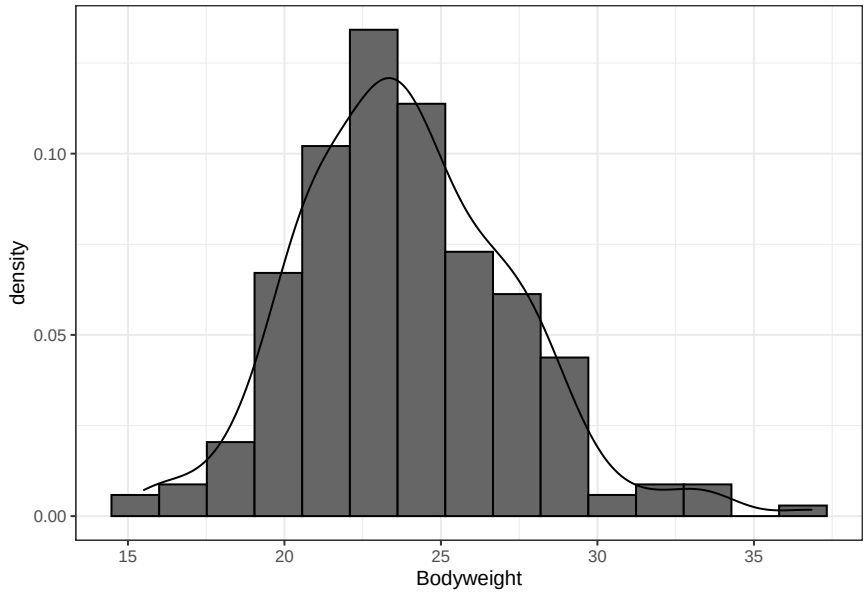


1.3 Distribuição Normal

A distribuição normal (ou Gaussiana) é uma distribuição tipicamente associada a variáveis contínuas, como altura, peso, etc. Quando a distribuição de uma variável se aproxima de uma distribuição normal, podemos calcular a probabilidade de obtermos um valor dentro de qualquer intervalo com:

$$\Pr(a < x < b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx$$

Por exemplo, vamos observar a distribuição de peso para fêmeas de camundongos. Os dados foram obtidos de Love (2016).



Calculando média e desvio padrão do peso:

```r
media <- mean(female_control$Bodyweight)
media
```

```
## [1] 23.89338
```

```r
desvio_padrao <- sd(female_control$Bodyweight)
desvio_padrao
```

```
## [1] 3.424056
```

Observe que a distribuição se assemelha a uma distribuição normal, com média = 23.9 e desvio padrão = 3.42. Sabendo disso, qual é a probabilidade de, tomando ao acaso, obtermos um camundongo cujo peso é inferior a 16 g? E superior a 30 g? Para isso, basta calcular as probabilidades assumindo uma distribuição normal teórica com esses valores de média e desvio padrão.

```r
# P(inferior a 16 g)
pnorm(16, mean = media, sd = desvio_padrao)
```

```
## [1] 0.01057569
```

```r
# P(superior a 30 g)
1 - pnorm(30, mean = media, sd = desvio_padrao)
```

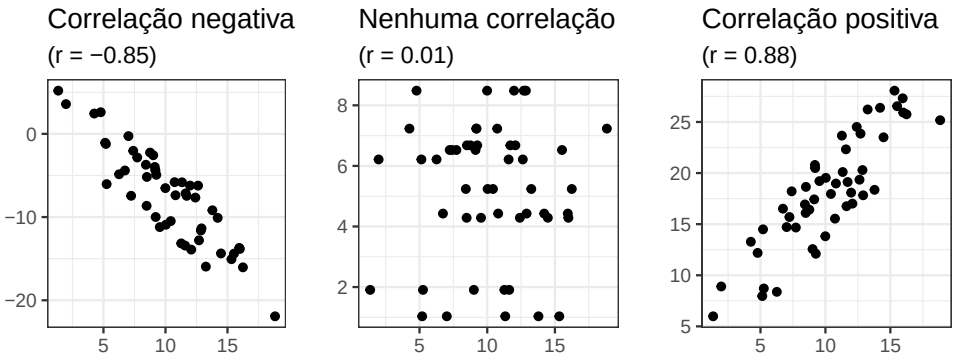
```
## [1] 0.03725679
```

2 Relações entre variáveis: correlação e regressão

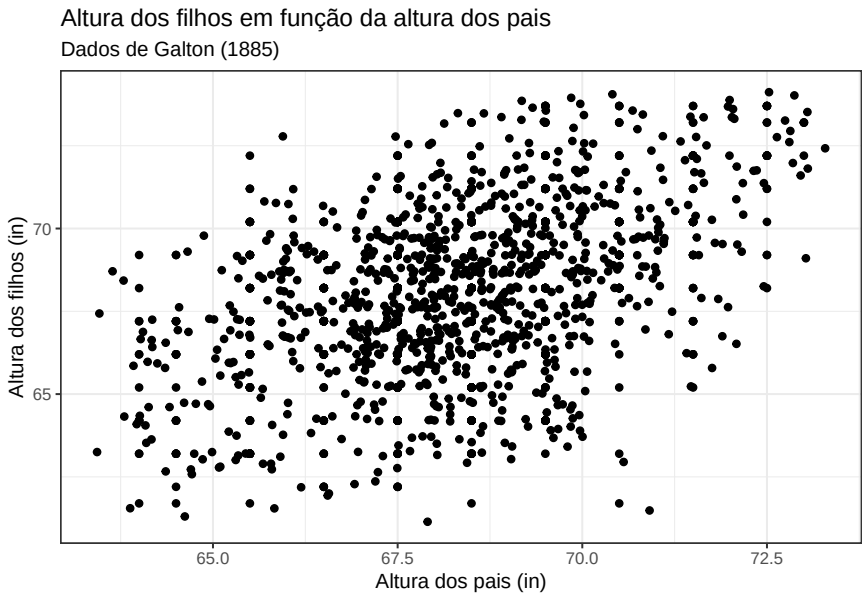
Correlação e regressão são duas formas comuns de explorar a relação entre duas variáveis. Embora comumente confundidas, seus objetivos são diferentes.

2.1 Correlação

A correlação é uma medida de associação entre duas variáveis. Usamos testes de correlação para medir a força e direção da asssociação entre duas variáveis, além de atribuir um grau de significância para essa associação (através de um valor de P). A força de associação é medida com um **coeficiente de correlação**, que varia de -1 a 1.



Por exemplo, vamos calcular a correlação entre a altura de homens e seus filhos. Esses dados foram obtidos por Galton (1886), e foram estudados à época para estudar a relação existente entre a altura de pais e filhos.




```
cor.test(altura$child, altura$parent)
```

```
##
##  Pearson's product-moment correlation
##
## data:  altura$child and altura$parent
## t = 15.711, df = 926, p-value < 0.00000000000000022
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4064067 0.5081153
## sample estimates:
##       cor
## 0.4587624
```

As alturas dos pais e de seus filhos apresentam correlação significativa ($P < 0.05$) e moderada (r = 0.46). Nesse exemplo, estamos calculando o **coeficiente de correlação de Pearson**, que assume que as distribuições de ambas as variáveis se aproximam de uma distribuição normal com média μ e variância σ^2. Portanto, essa é uma estatística **paramétrica**. Se esse não for o caso dos dados em questão, pode-se calcular o **coeficiente de correlação de Spearman**, que é uma estatística não-paramétrica.

```
cor.test(altura$child, altura$parent, method = "spearman")
```

```
## Warning in cor.test.default(altura$child, altura$parent, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  altura$child and altura$parent
## S = 76569964, p-value < 0.00000000000000022
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.4251345
```

Observe que como a distribuição de altura dos pais e dos filhos se aproxima de uma distribuição normal, usar o método de Pearson ou de Spearman leva praticamente ao mesmo coeficiente de correlação.

2.2 Regressão

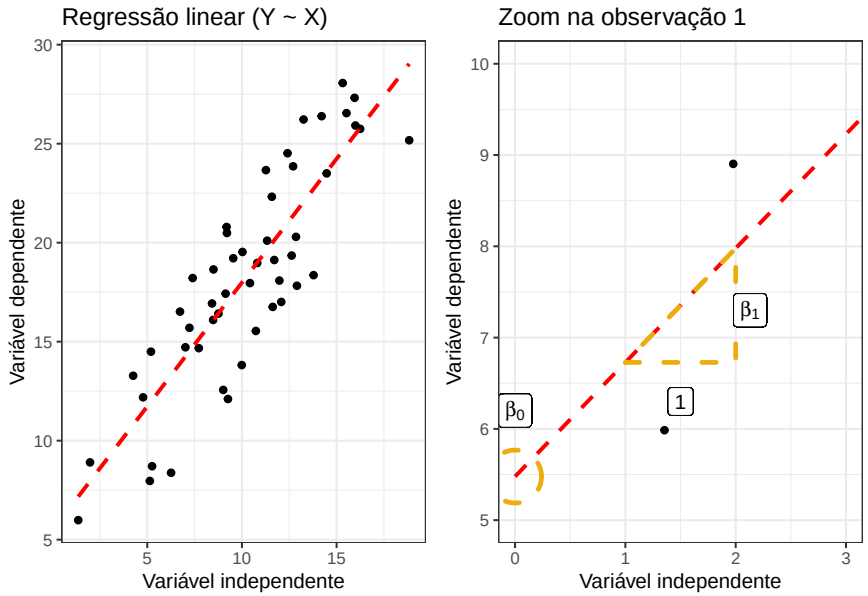
Enquanto a correlação mede a força e direção da associação entre duas variáveis, a regressão é usada para medir como a mudança em uma variável influencia a mudança em outra. Dessa forma, podemos usar uma variável preditora (X, ou variável independente) para predizer a mudança em uma variável de interesse (Y, ou variável dependente). Existem diversos modelos de regressão (*e.g.,* linear, quadrática, polinomial, etc.), mas vamos focar nos modelos lineares e suas aplicações.

2.2.1 Modelos lineares simples

Um modelo de regressão simples pode ser definido matematicamente da seguinte forma:

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

onde β_0 é o intercepto (ponto da linha que cruza o eixo Y), β_1 é o coeficiente angular (peso aplicado a X_i), X_i é o valor da variável preditora no ponto i, Y_i é o valor da variável dependente no ponto i, e ϵ é a variável de erro (resíduo), que segue distribuição normal.



O melhor modelo de regressão linear é aquele que minimiza a **soma dos quadrados dos resíduos**, sendo os resíduos a diferença entre o valor observado e o valor predito no modelo. Vamos usar os dados de Galton (1886) novamente para ajustar um modelo de regressão linear para tentar predizer a altura dos filhos usando a altura dos pais como variável preditora.

```
altura <- UsingR::galton
head(altura)
```

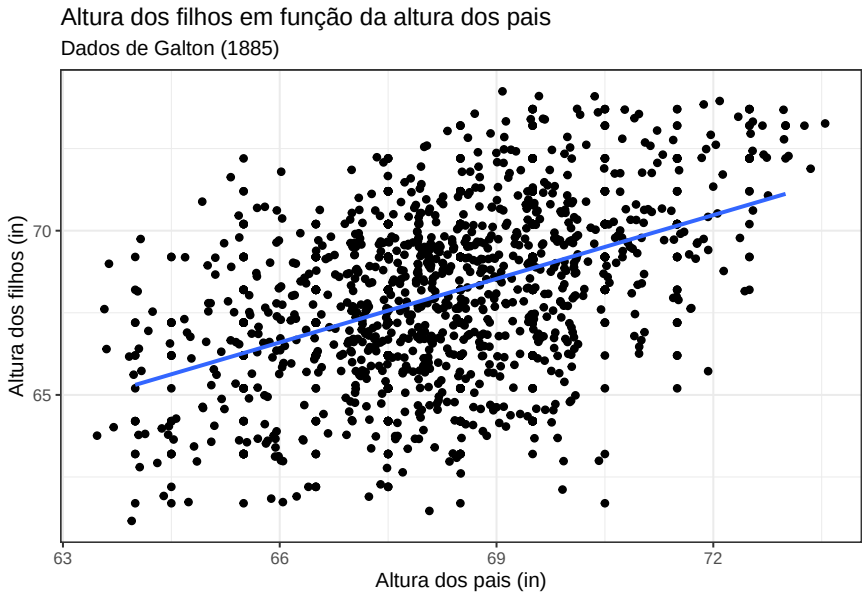
```
##   child parent
## 1  61.7   70.5
## 2  61.7   68.5
## 3  61.7   65.5
## 4  61.7   64.5
## 5  61.7   64.0
## 6  62.2   67.5
```

```
# Ajustando modelo de regressão linear
modelo_lm <- lm(child ~ parent, data = altura)
summary(modelo_lm)
```

```
##
## Call:
## lm(formula = child ~ parent, data = altura)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8050 -1.3661  0.0487  1.6339  5.9264
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 23.94153    2.81088   8.517 <0.0000000000000002 ***
## parent       0.64629    0.04114  15.711 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.239 on 926 degrees of freedom
## Multiple R-squared:  0.2105, Adjusted R-squared:  0.2096
## F-statistic: 246.8 on 1 and 926 DF,  p-value: < 0.00000000000000022
```

Observe que a altura dos pais não é uma boa variável preditora para a altura dos filhos ($R^2 = 0.21$). O coeficiente de determinação R^2 representa a proporção da variação em Y explicada por X.

O coeficiente angular β_1 para o modelo é 0.646. Isso significa que para cada aumento de uma unidade em X (altura dos pais), Y (altura dos filhos) aumenta em 0.64 unidade. Visualizando o modelo:



A partir do modelo predito, podemos predizer o valor de Y_i para qualquer valor de X_i. Por exemplo, qual é a altura predita de uma criança cujo pai tem 65 polegadas de altura? E de uma criança cujo pai tem 67 polegadas?

```
pais <- data.frame(parent = c(65,67))
predict(modelo_lm, newdata = pais)
```

```
##        1        2
## 65.95042 67.24300
```

Em certos casos, podemos predizer o valor de uma variável dependente usando 2 ou mais variáveis independentes. Nesse caso, trata-se de uma **regressão linear multivarada ou regressão linear múltipla**, que

pode ser expressa da seguinte maneira:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_p X_{i,p} + \epsilon_i$$

Ao levar em consideração outras variáveis, podemos aumentar a acurácia do nosso modelo ajustado. Como exemplo, vamos usar uma regressão multivariada para predizer o efeito de diversas variáveis independentes na variável fertilidade. Os dados já vêm disponíveis numa instalação da linguagem R (R Core Team 2021), e apresentam a medidade de fertilidade e indicadores socioeconômicos de 47 províncias francófonas da Suiça em 1888.

```
head(swiss)
```

```
##              Fertility Agriculture Examination Education Catholic
## Courtelary        80.2        17.0          15        12     9.96
## Delemont          83.1        45.1           6         9    84.84
## Franches-Mnt      92.5        39.7           5         5    93.40
## Moutier           85.8        36.5          12         7    33.77
## Neuveville        76.9        43.5          17        15     5.16
## Porrentruy        76.1        35.3           9         7    90.57
##              Infant.Mortality
## Courtelary               22.2
## Delemont                 22.2
## Franches-Mnt             20.2
## Moutier                  20.3
## Neuveville               20.6
## Porrentruy               26.6
```

```
modelo2 <- lm(Fertility ~ ., data = swiss)
summary(modelo2)
```

```
##
## Call:
## lm(formula = Fertility ~ ., data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2743  -5.2617   0.5032   4.1198  15.3213
##
## Coefficients:
##                  Estimate Std. Error t value    Pr(>|t|)
## (Intercept)      66.91518   10.70604   6.250 0.000000191 ***
## Agriculture      -0.17211    0.07030  -2.448     0.01873 *
## Examination      -0.25801    0.25388  -1.016     0.31546
## Education        -0.87094    0.18303  -4.758 0.000024306 ***
## Catholic          0.10412    0.03526   2.953     0.00519 **
## Infant.Mortality  1.07705    0.38172   2.822     0.00734 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.165 on 41 degrees of freedom
## Multiple R-squared:  0.7067, Adjusted R-squared:  0.671
## F-statistic: 19.76 on 5 and 41 DF,  p-value: 0.0000000005594
```

É importante ressaltar que a interpretação dos coeficientes para esse tipo de regressão é diferente. Os coeficientes indicam o efeito (em unidades) da variável preditora na variável resposta **mantendo todas as outras variáveis preditoras constantes**. Por exemplo, o modelo acima indica um decréscimo de 0.17

unidade na fertilidade como consequência da agricultura, e um decréscimo de 0.87 unidade como consequência da educação.

2.2.2 Modelos lineares com interação

Quando duas variáveis independentes têm efeito aditivo na variável dependente, o modelo pode ser ajustado com uma regressão multivariada, como exemplificado acima. Entretanto, se as duas variáveis têm efeito sinérgico na variável dependente, diz-se que as variáveis interagem. Nesse caso, deve-se adicionar um termo de interação no modelo, que passa a ser representado da seguinte forma:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,1} X_{i,2} + \epsilon_i$$

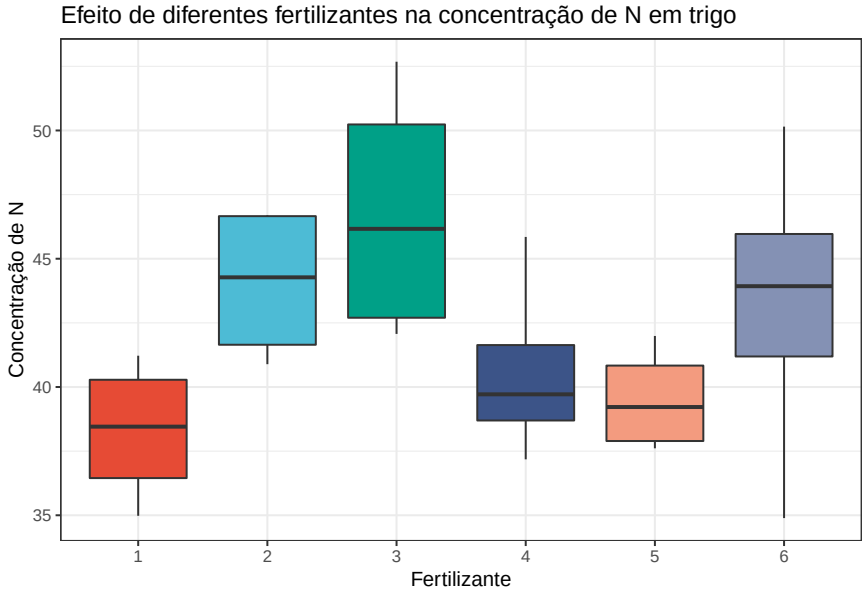
2.2.3 Modelos lineares de efeitos mistos

Um modelo linear simples tem algumas premissas. Uma delas é que as observações são independentes entre si. No entanto, dados biológicos são frequentemente obtidos com fatores de agrupamento, como local de coleta, população, blocos (em agronomia), etc. Como a dependência entre algumas observações viola uma premissa do modelo linear simples, usamos modelos lineares de efeitos mistos, que incluem tanto efeitos fixos (os mesmos do modelo linear simples) quanto efeitos aleatórios (variável que agrupa as observações). Esses modelos podem ser definidos como:

$$Y_i = \underbrace{\beta X_i}_{\text{Efeito fixo}} + \underbrace{Z\gamma + e_i}_{\text{Efeito aleatório}}$$

Para exemplificar, vamos usar dados obtidos de um experimento que buscava avaliar o efeito da aplicação de diferentes fertilizantes na concentração de nitrogênio em trigo.

```
##   concentracao bloco fertilizante
## 1        40.89     1            2
## 2        37.99     1            5
## 3        37.18     1            4
## 4        34.98     1            1
## 5        34.89     1            6
## 6        42.07     1            3
```

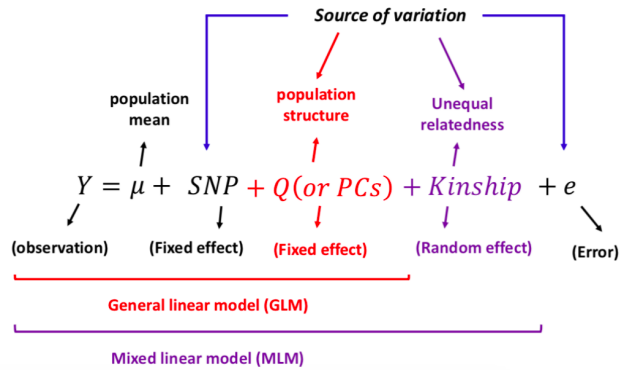


Esse experimento foi realizado num delineamento em blocos. As plantas dentro do mesmo bloco não são independentes, violando a premissa de independência das observações de um modelo linear simples. Portanto, vamos ajustar um modelo misto usando a variável bloco como efeito aleatório.

```
modelo_misto <- lmer(concentracao ~ fertilizante + (1|bloco), data = trigo)
emmeans::emmeans(modelo_misto,
                 pairwise ~ fertilizante,
                 adjust = "Tukey")
```

```
## $emmeans
##  fertilizante emmean   SE   df lower.CL upper.CL
##  1              38.3 2.06 6.78     33.4     43.2
##  2              44.0 2.06 6.78     39.1     48.9
##  3              46.8 2.06 6.78     41.9     51.7
##  4              40.6 2.06 6.78     35.7     45.5
##  5              39.5 2.06 6.78     34.6     44.4
##  6              43.2 2.06 6.78     38.3     48.1
##
## Degrees-of-freedom method: kenward-roger
## Confidence level used: 0.95
##
## $contrasts
##  contrast estimate  SE df t.ratio p.value
##  1 - 2      -5.755 1.9 15  -3.033  0.0742
##  1 - 3      -8.492 1.9 15  -4.476  0.0049
##  1 - 4      -2.337 1.9 15  -1.232  0.8150
##  1 - 5      -1.232 1.9 15  -0.650  0.9849
##  1 - 6      -4.947 1.9 15  -2.607  0.1553
##  2 - 3      -2.737 1.9 15  -1.443  0.7025
##  2 - 4       3.417 1.9 15   1.801  0.4934
##  2 - 5       4.522 1.9 15   2.383  0.2226
##  2 - 6       0.807 1.9 15   0.426  0.9978
##  3 - 4       6.155 1.9 15   3.244  0.0505
##  3 - 5       7.260 1.9 15   3.826  0.0168
##  3 - 6       3.545 1.9 15   1.868  0.4559
##  4 - 5       1.105 1.9 15   0.582  0.9907
##  4 - 6      -2.610 1.9 15  -1.376  0.7402
##  5 - 6      -3.715 1.9 15  -1.958  0.4079
##
## Degrees-of-freedom method: kenward-roger
## P value adjustment: tukey method for comparing a family of 6 estimates
```

Os modelos lineares de efeitos mistos são muito populares nos estudos de associação genômica ampla (GWAS), pois permitem levar em consideração o parentesco entre os genomas estudados (Yu et al. 2006). Se desconsiderado, o parentesco pode ser um fator de confusão e levar a associações espúrias entre polimorfismos de nucleotídeo único (SNPs) e fenótipos de interesse. Os modelos mistos em GWAS se configuram da seguinte forma:



2.2.4 Modelos lineares generalizados (GLMs)

Como o nome indica, os modelos lineares generalizados (GLMs, do inglês *Generalized Linear Models*) são extensões de modelos lineares. Isso significa que eles são muito similares a modelos lineares simples, mas são mais permissivos em relação a algumas premissas.

Modelo linear simples:

1. Observações (Y_i) independentes
2. Resíduos seguem distribuição normal com média μ e variância σ^2
3. $\mu_i = X_i^T \beta$

Modelo linear generalizado:

1. Observações (Y_i) independentes
2. Resíduos seguem qualquer distribuição da família exponencial
3. $g(\mu_i) = X_i^T \beta$, onde g é a função de link

Como apontado na comparação acima, a principal diferença entre GLMs e modelos lineares simples é que os resíduos de um GLM podem seguir qualquer distribuição da família exponencial, não apenas uma distribuição normal. Além disso, para ajustar um GLM, é preciso usar uma função de link que conecta μ_i aos preditores. Abaixo estão as distribuições da família exponencial e suas respectivas funções de link por padrão em linguagem R.

Família	Função.de.link
Binomial	logit
Gaussiana	identidade
Gama	inversa
Gaussiana inversa	1/mu^2
Poisson	log
Quasi-binomial	logit
Quasi-Poisson	log

Para exemplificar a aplicação de um GLM, vamos ajustar uma **regressão logística** em um conjunto de dados onde a variável dependente é binária, não contínua. Os dados apresentam os níveis de glicose no sangue de indivíduos e o diagnóstico de diabetes (variável binária, sim ou não). O objetivo é predizer se o indivíduo tem diabetes

```
##     pregnant glucose pressure triceps insulin mass pedigree age diabetes
## 637        5     104       74      NA      NA 28.8    0.153  48      neg
## 254        0      86       68      32      NA 35.8    0.238  25      neg
## 18         7     107       74      NA      NA 29.6    0.254  31      pos
## 663        8     167      106      46     231 37.6    0.165  43      pos
```

```
## 280        2     108       62      10     278 25.3    0.881  22      neg
## 534        6      91       NA      NA      NA 29.8    0.501  31      neg
```

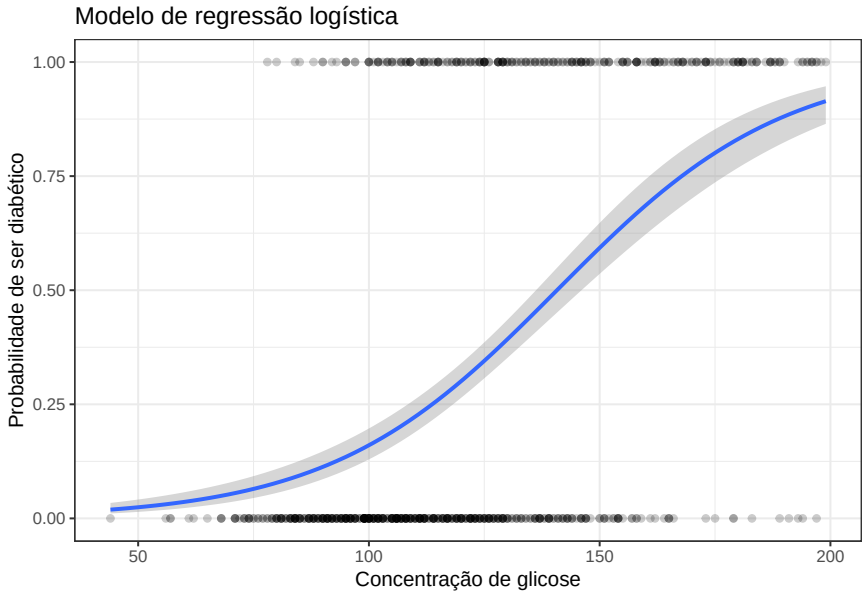
```
modelo <- glm(diabetes ~ glucose, data = diabetes_dados, family = binomial)
summary(modelo)
```

```
##
## Call:
## glm(formula = diabetes ~ glucose, family = binomial, data = diabetes_dados)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1829  -0.7758  -0.5083   0.8246   2.2914
##
## Coefficients:
##              Estimate Std. Error z value            Pr(>|z|)
## (Intercept) -5.719815   0.443532  -12.90 <0.0000000000000002 ***
## glucose      0.040637   0.003421   11.88 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.62  on 745  degrees of freedom
## Residual deviance: 769.56  on 744  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 773.56
##
## Number of Fisher Scoring iterations: 4
```

Com o modelo ajustado, podemos predizer se uma pessoa tem ou não diabetes do tipo 2 usando os níveis de glicose como variável preditora. A predição do modelo nos fornece probabilidades. Aqui, vamos considerar um limiar de 50% para definir se o indivíduo tem diabetes, isto é, indivíduos com mais de 50% de probabilidade de serem diabéticos serão considerados diabéticos.

```
newdata <- data.frame(glucose = c(20,  180))
prob <- modelo %>%
  predict(newdata, type = "response")
predicao <- ifelse(prob > 0.5, "pos", "neg")
predicao
```

```
##     1     2
## "neg" "pos"
```



3 *Big data* na agricultura: aprendizagem de máquina e integração de dados

Nas últimas décadas, dados fisiológicos e moleculares de plantas acumularam em bancos de dados públicos. Esse fenômeno abriu caminhos para algoritmos destinados à integração de dados com o objetivo de extrair informações relevantes e fazer predições (Benos et al. 2021), em especial ao campo da aprendizagem de máquina (ML, do inglês *machine learning*). Esses algoritmos usam informações de centenas de observações para diversas variáveis para identificar padrões (Benos et al. 2021) que podem ser usados para predição em novos dados.

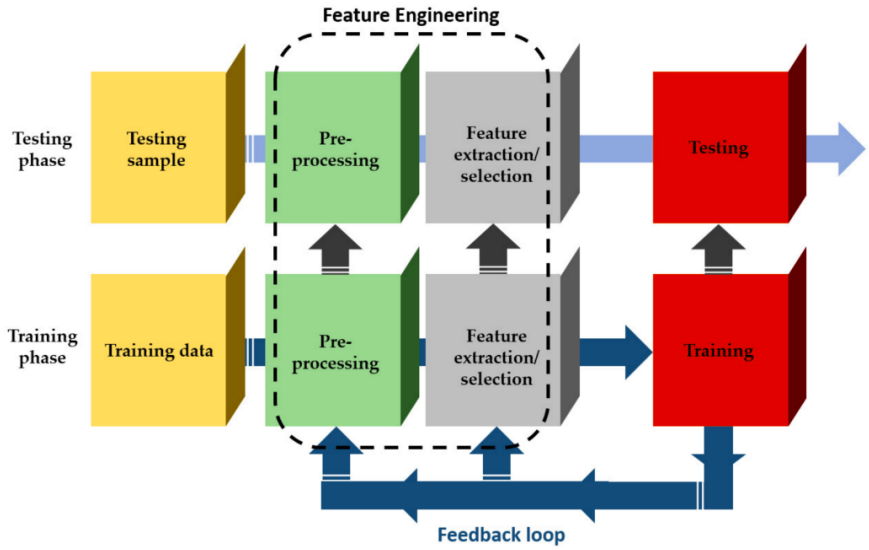


Figure 1: Figure extraída de Benos et al., 2021

A principal vantagem desses algoritmos em relação aos modelos tradicionais é a sua maior precisão, pois os modelos de classificação são complexos, envolvendo muitas variáveis preditoras. Ainda, alguns algoritmos de ML são capazes de decompor as variáveis preditoras em camadas (*deep learning*), aumentando ainda mais a

precisão da predição.

Referências

Benos, Lefteris, Aristotelis C Tagarakis, Georgios Dolias, Remigio Berruto, Dimitrios Kateris, and Dionysis Bochtis. 2021. "Machine Learning in Agriculture: A Comprehensive Updated Review." *Sensors* 21 (11): 3758.

Galton, Francis. 1886. "Regression Towards Mediocrity in Hereditary Stature." *The Journal of the Anthropological Institute of Great Britain and Ireland* 15: 246–63.

Holmes, S, and W Huber. n.d. "Modern Statistics for Modern Biology, 2019." Cambridge: Cambridge University Press.

Love, Michael I. 2016. "Statistical Models." In *Data Analysis for the Life Sciences with r*, 231–58. Chapman; Hall/CRC.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Yu, Jianming, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, et al. 2006. "A Unified Mixed-Model Method for Association Mapping That Accounts for Multiple Levels of Relatedness." *Nature Genetics* 38 (2): 203–8.