

# Análise de Regressão para dados de Glucose

Samuel Medeiros

## Introdução

Este relatório apresenta uma análise de regressão realizada com base em um conjunto de dados coletados. O objetivo principal desta análise é investigar as relações entre uma variável de resposta e várias variáveis independentes, a fim de desenvolver um modelo de regressão que possa prever o valor da variável de resposta com base nas demais variáveis.

Ao longo deste relatório, serão apresentados os resultados da análise de regressão, incluindo as estatísticas dos coeficientes do modelo, a significância estatística das relações e a avaliação da qualidade de ajuste do modelo. Além disso, serão discutidas as principais conclusões obtidas a partir da análise.

Os dados utilizados neste estudo consistem em casos de diabetes, contendo informações de pacientes, como o número de gestações, nível de glicose, pressão arterial, espessura da pele, insulina, índice de massa corporal (IMC), função de pedigree de diabetes, idade e resultado (1 para positivo e 0 para negativo).

## Análise descritiva

Esta seção apresenta uma análise descritiva dos dados coletados para o modelo de regressão. O conjunto de dados consiste em informações sobre uma variável de resposta e várias variáveis independentes, como apresentado na Tabela 1.

Table 1: Variáveis disponíveis para o modelo

Variável	Nome	Descrição
$Y$	Glucose	Concentração de glicose em teste oral de tolerância à glicose
$X_1$	Pregnancies	Número de vezes que a índia engravidou
$X_2$	BloodPressure	Pressão arterial diastólica (mm Hg)
$X_3$	SkinThickness	Espessura cutânea tricipital (mm)
$X_4$	Insulin	2 horas de insulina no soro ( $\mu$ U/ml)
$X_5$	BMI	Índice de massa corporal (IMC)
$X_6$	DiabetesPedigree	Diabetes função da genealogia
$X_7$	Age	Idade (anos)
$X_8$	Outcome	Teste de diabetes (0=saudável, 1=diabético)

Inicialmente, é importante realizar uma análise de consistência dos dados, devido à possibilidade de existirem valores descritos de forma errada que podem prejudicar o desempenho do modelo. Por exemplo, valores que deveriam ser NA podem estar preenchidos com o valor 0, bem como a exclusão das observações com valores faltantes para a variável resposta, nos restando uma amostra de 763 unidades amostrais. Essa análise de consistência será realizada antes da modelagem de regressão para garantir a qualidade dos resultados. Os códigos utilizados podem ser vistos no anexo do arquivo, aqui apenas os resultados serão apresentados.

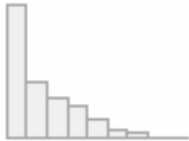
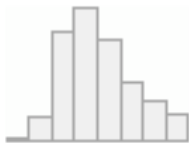
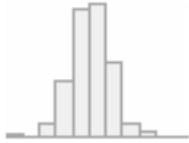
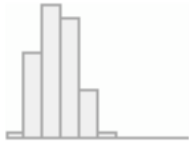
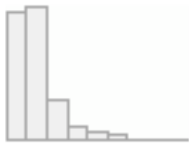
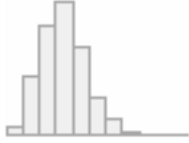
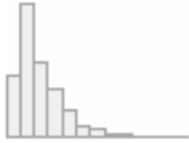
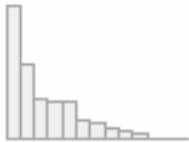
É possível observar abaixo a distribuição dos dados disponíveis bem como as devidas estatísticas para as variáveis numéricas. É perceptível um valor próximo de 50% das observações, veremos na área de modelagem se a variável agrega ou não informações para o modelo.

## Informações a Respeito dos Dados

df

Dimensões: 763 x 9

Duplicadas: 0

No	Variáveis	Estatísticas/Valores	Freq. De Válidos	Gráficos	Faltantes
1	Pregnancies [numeric]	Mean (sd) : 3.9 (3.4) min < med < max: 0 < 3 < 17 IQR (CV) : 5 (0.9)	17 Valores Distintos		0 (0.0%)
2	Glucose [numeric]	Mean (sd) : 121.7 (30.5) min < med < max: 44 < 117 < 199 IQR (CV) : 42 (0.3)	135 Valores Distintos		0 (0.0%)
3	BloodPressure [numeric]	Mean (sd) : 72.4 (12.4) min < med < max: 24 < 72 < 122 IQR (CV) : 16 (0.2)	46 Valores Distintos		35 (4.6%)
4	SkinThickness [numeric]	Mean (sd) : 29.1 (10.5) min < med < max: 7 < 29 < 99 IQR (CV) : 14 (0.4)	50 Valores Distintos		227 (29.8%)
5	Insulin [numeric]	Mean (sd) : 155.9 (118.7) min < med < max: 14 < 125 < 846 IQR (CV) : 113 (0.8)	185 Valores Distintos		370 (48.5%)
6	BMI [numeric]	Mean (sd) : 32.5 (6.9) min < med < max: 18.2 < 32.3 < 67.1 IQR (CV) : 9.1 (0.2)	246 Valores Distintos		11 (1.4%)
7	DiabetesPedigree [numeric]	Mean (sd) : 0.5 (0.3) min < med < max: 0.1 < 0.4 < 2.4 IQR (CV) : 0.4 (0.7)	516 Valores Distintos		0 (0.0%)
8	Age [numeric]	Mean (sd) : 33.3 (11.8) min < med < max: 21 < 29 < 81 IQR (CV) : 17 (0.4)	52 Valores Distintos		0 (0.0%)

No	Variáveis	Estatísticas/Valores	Freq. De Válidos	Gráficos	Faltantes
9	Outcome [factor]	1. 0 2. 1	497 (65.1%) 266 (34.9%)		0 (0.0%)

Podemos ver abaixo a interação, através da análise de correlação, de cada variável explicativa com a variável reposta, note porém que aqui não será considerada a variável `Outcome` por se tratar de uma variável categórica.

```
library(corrplot)
```

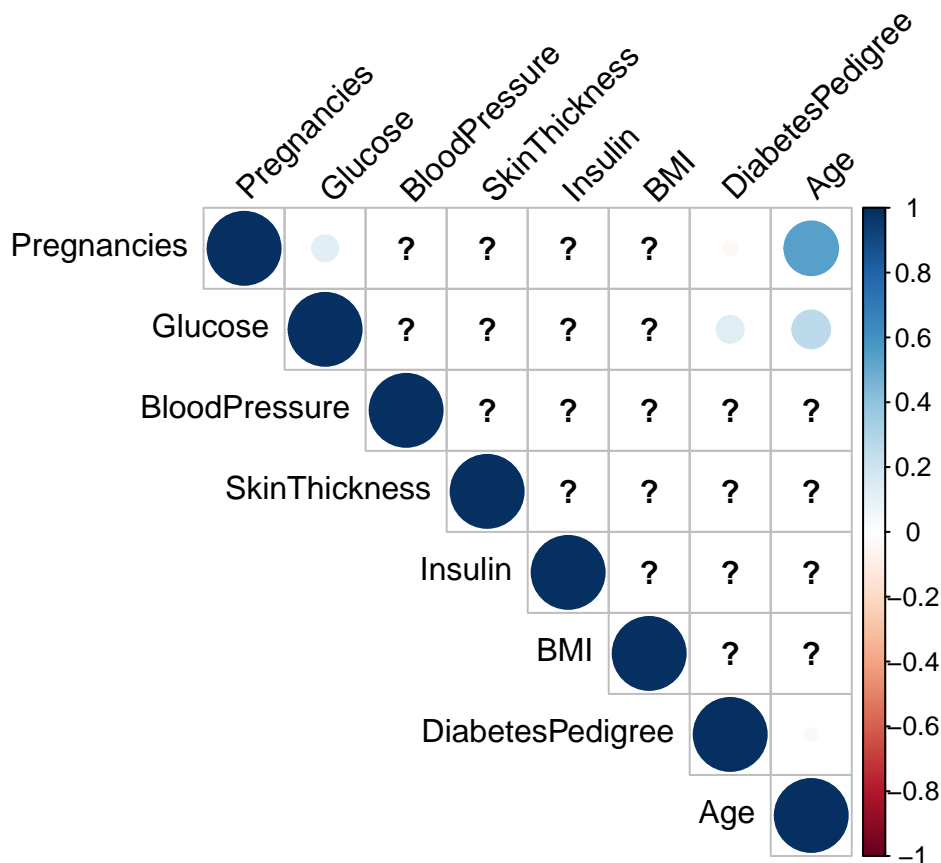
```
## corrplot 0.92 loaded
```

```
# Calcular a matriz de correlação
```

```
matriz_cor <- cor(df %>% select(-Outcome))
```

```
# Plotar um gráfico de correlação
```

```
corrplot(matriz_cor, method = "circle", type = "upper", tl.col = "black", tl.srt = 45)
```



## Modelagem

Nesta seção, serão apresentados os resultados da modelagem de regressão realizada com base nos dados coletados. O objetivo principal é desenvolver um modelo de regressão que possa prever o valor da variável

de resposta com base em diversas variáveis independentes. Serão exploradas as possíveis interações entre as variáveis e, caso necessário, será realizada uma seleção de variáveis com base na análise de diagnóstico do modelo.

## Modelo 1

Inicialmente, foi desenvolvido um modelo de regressão incluindo todas as variáveis disponíveis, bem como as interações relevantes entre elas. Neste modelo inicial, buscamos investigar as relações entre as variáveis independentes e a variável de resposta. Serão apresentados os coeficientes estimados, seus respectivos intervalos de confiança e a significância estatística das relações. O modelo inicial segue como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \epsilon_i$$

Onde cada  $\beta_i$  corresponde a influência da covariável  $X_i$  na variável resposta