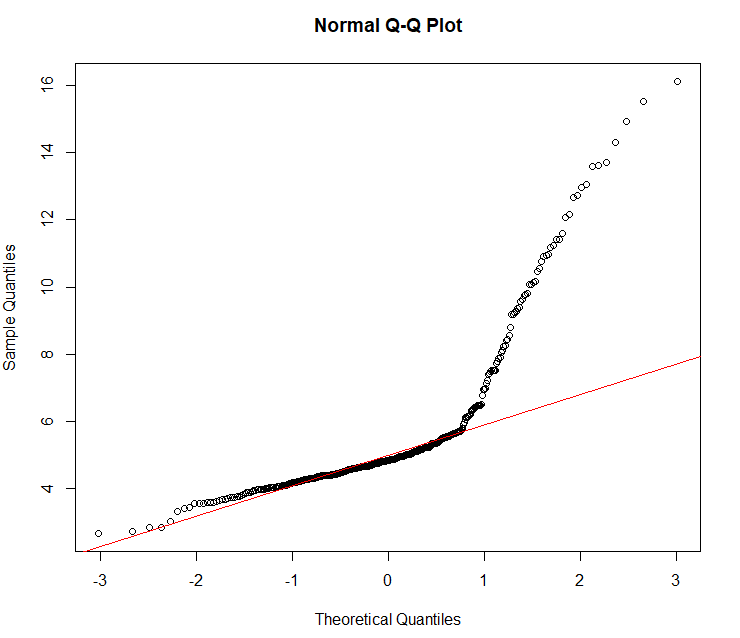
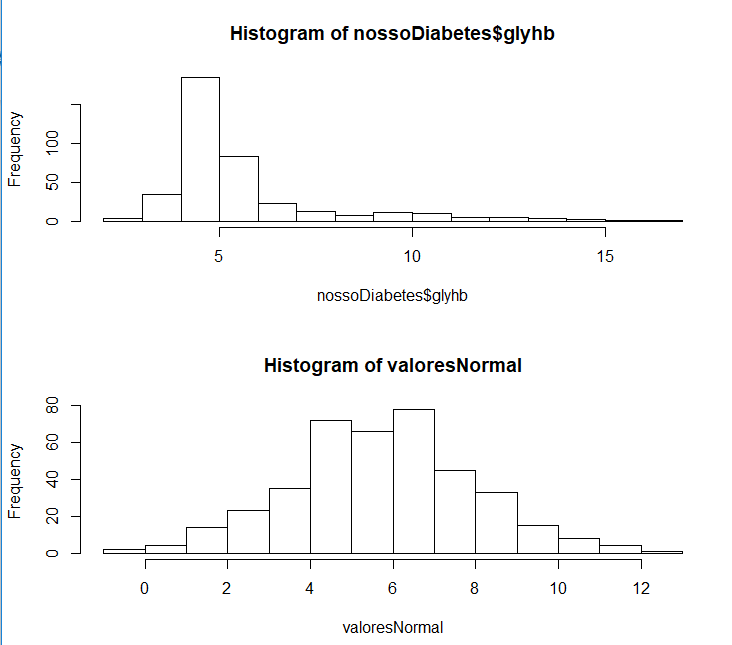
Análise Exploratória

# Teste à normalidade

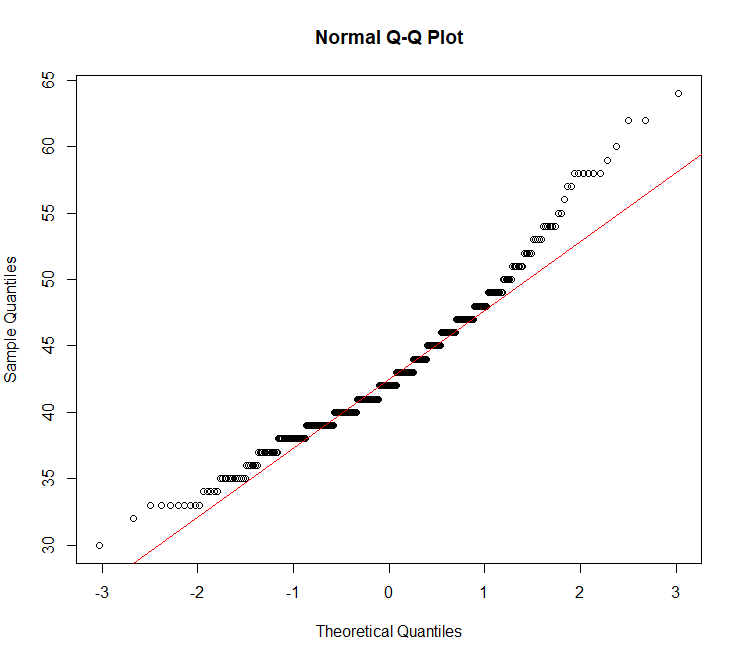
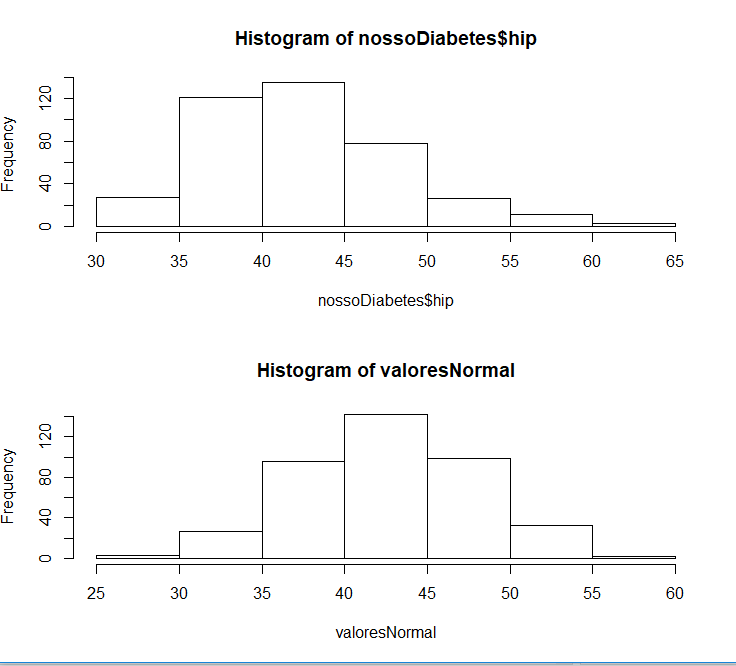
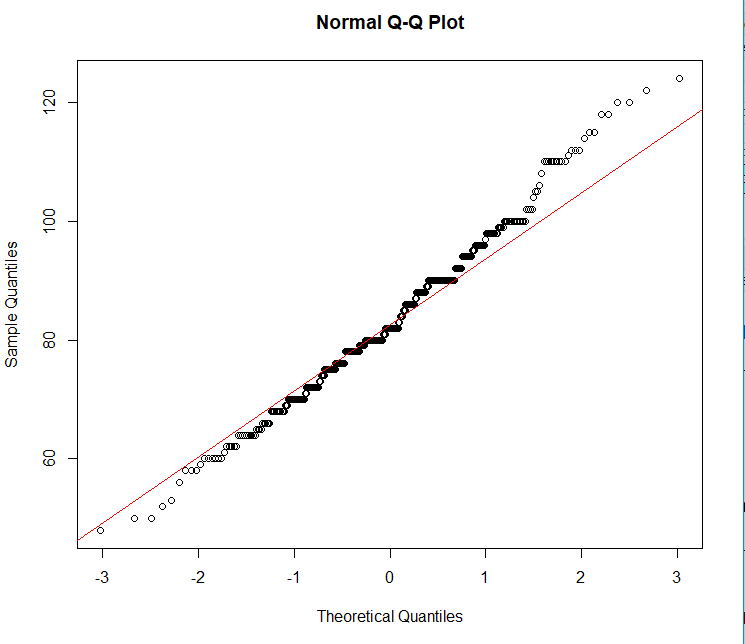
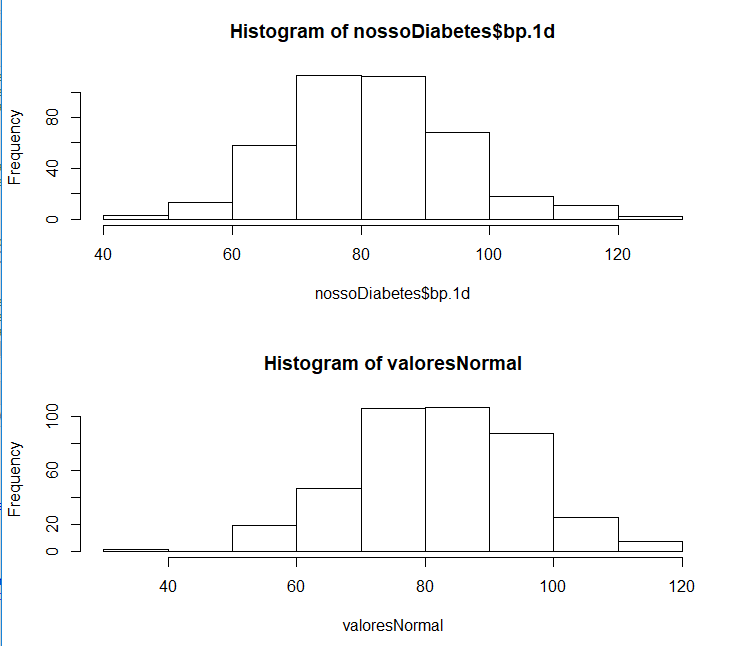
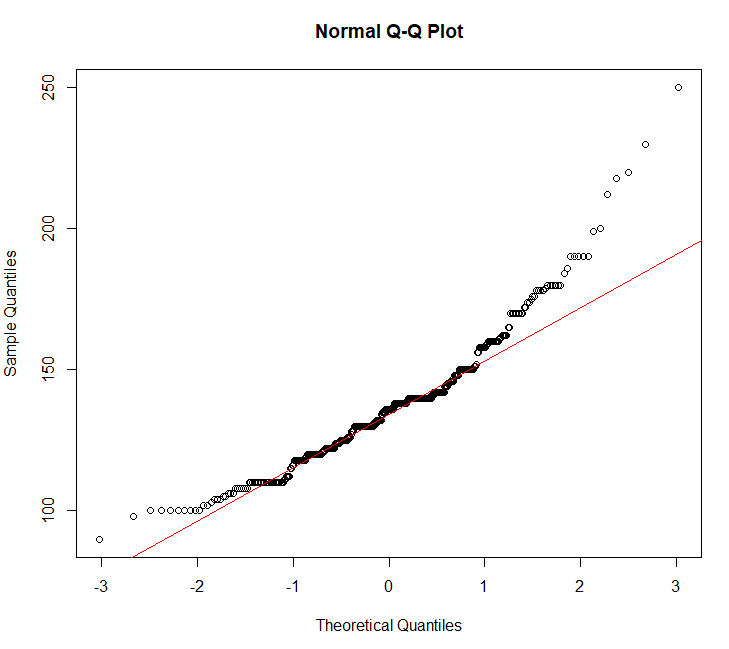
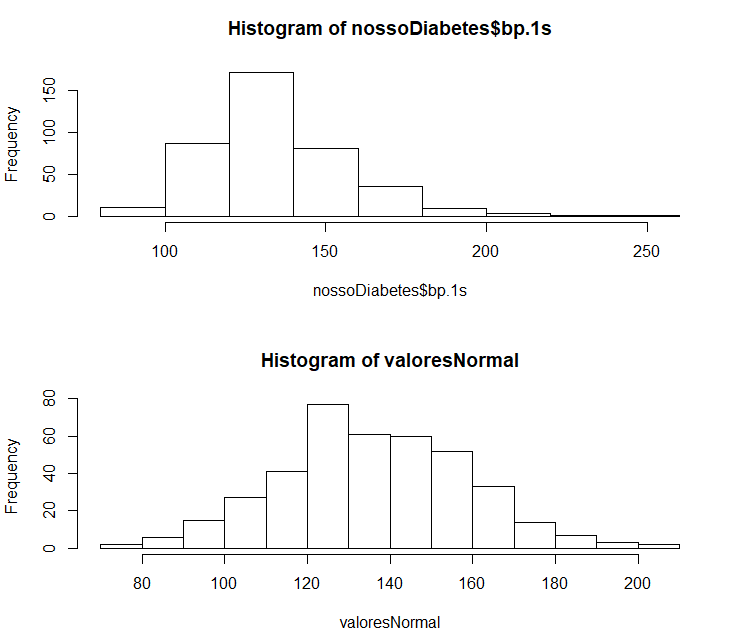
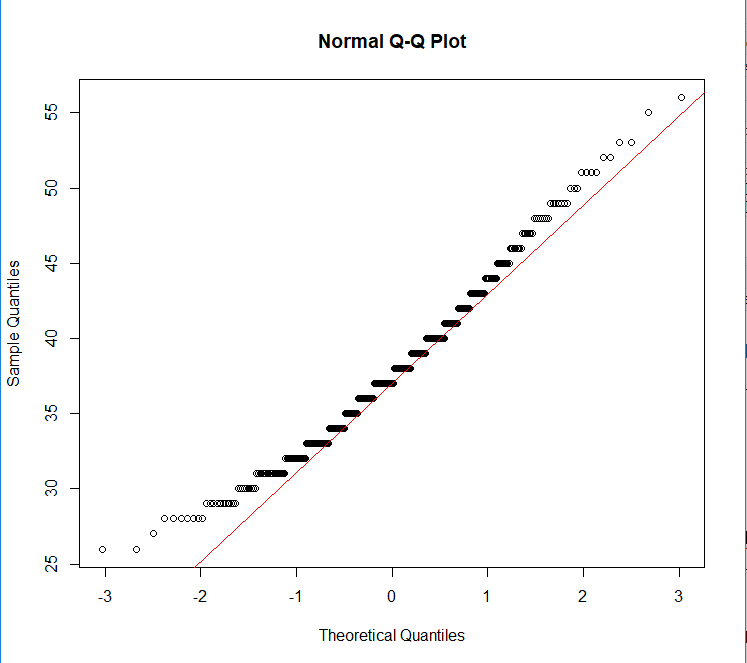
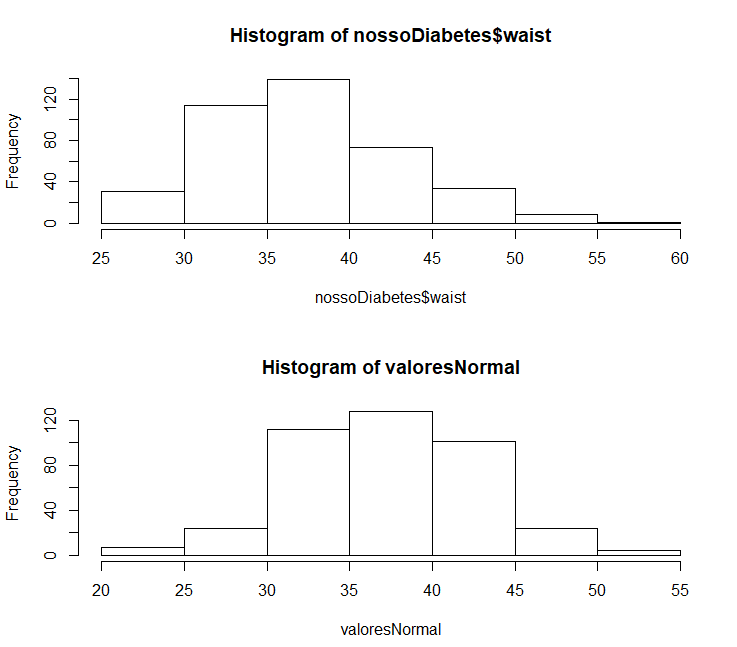
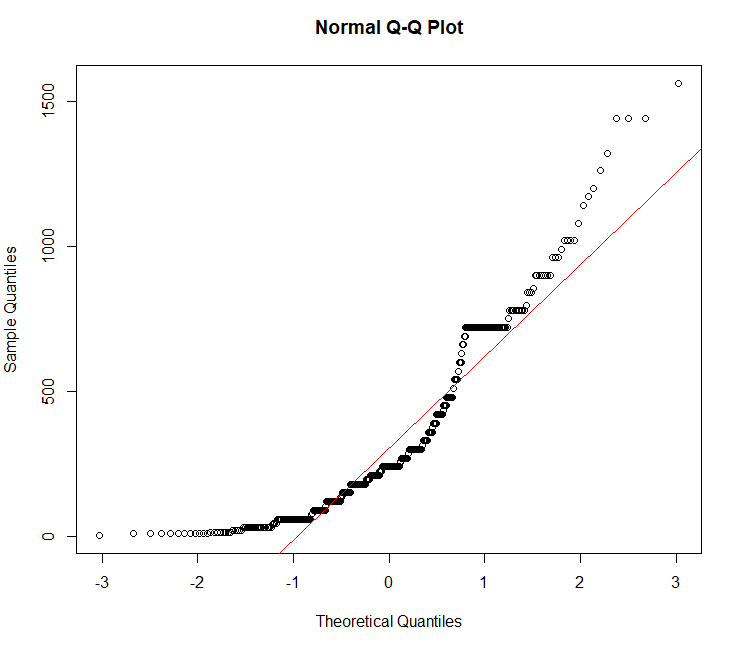
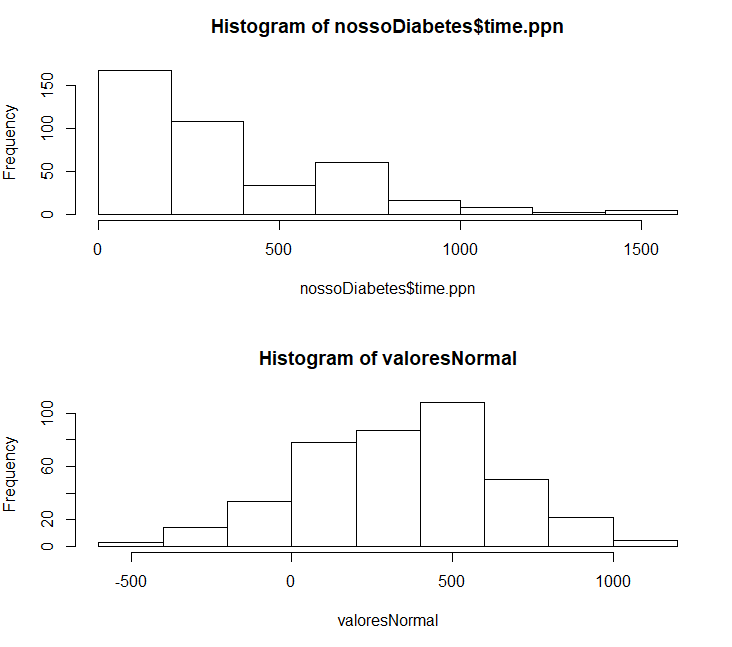
Por forma a excluir alguns métodos que pressupõe certas distribuições da variável resposta ou preditores, procedemos a um teste à normalidade de cada um dos elementos quantitativos do nosso modelo. Para isso comparamos os histogramas de cada variável com o histograma dos valores gerados segundo uma distribuição normal com a média e desvio padrão da variável. Comparamos também com os gráficos de teste da normalidade.

## Variável Resposta – glyhb



Como pode ser observado nas imagens acima, a variável resposta não aparenta seguir uma distribuição normal. Vemos isto pois esta distancia-se da linha do gráfico do teste da normal e também porque o seu histograma é bastante diferente do histograma com os valores simulados vindos da mesma distribuição.

## Preditores



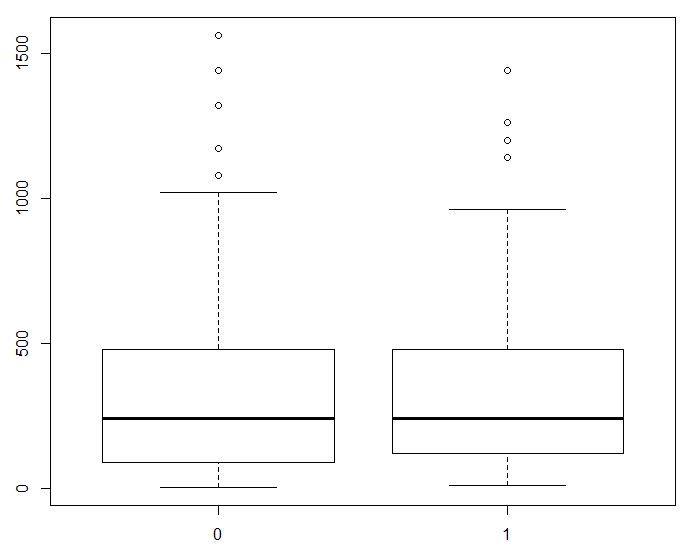
* **time.ppn:** podemos ver que tanto no histograma como no gráfico do teste da normalidade não se aproxima da distribuição normal.
* **waist:** no histograma podemos verificar que se aproxima da distribuição normal. No gráfico também se aproxima muito no meio, sendo que se afasta um pouco nos extremos.
* **hip:** pelo histograma podemos ver que se aproxima da distribuição normal. No gráfico também se aproxima, no entanto nos extremos afasta-se mais um pouco.
* **bp.1s:** os histogramas diferem ligeiramente. No gráfico vemos uma aproximação muito elevada no meio, no entanto nos extremos afasta-se muito.
* **bp.1d:** no histograma verificamos que se aproxima da distribuição normal. No gráfico também se parece aproximar.

Concluímos então que os preditores que se aproximam da distribuição normal são:

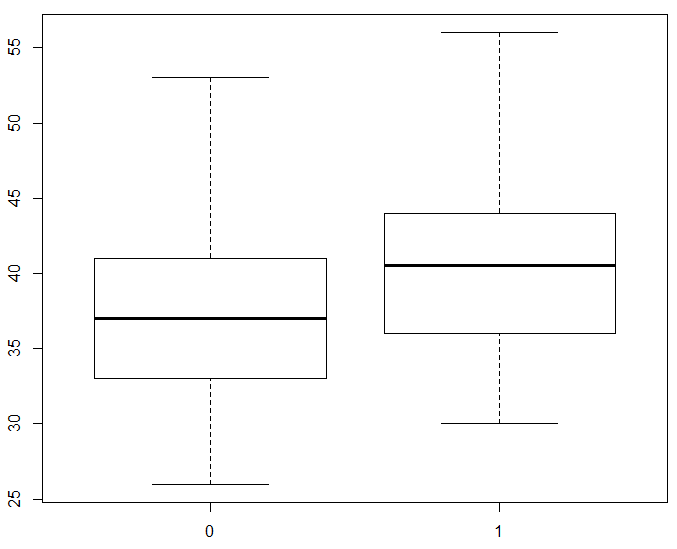
* waist
* hip
* bp.1d

# Teste à influência à variável binária

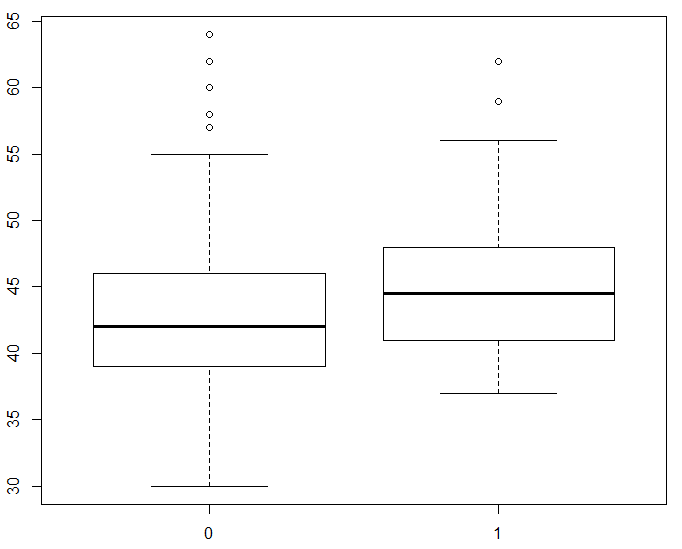
Visto que o nosso modelo tem bastantes valores em falta, decidimos realizar um teste à influência das variáveis na variável resposta binária por forma a selecionar logo à partida os preditores mais adequados, ficando assim com um maior número de casos para teste e treino.



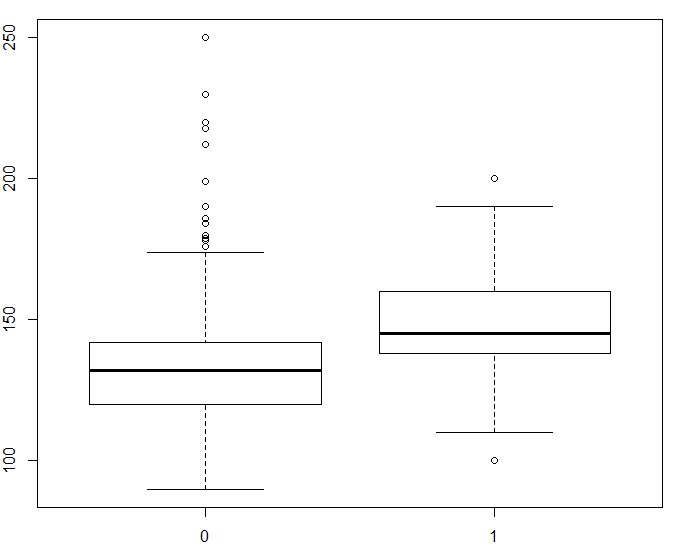
Time.ppn



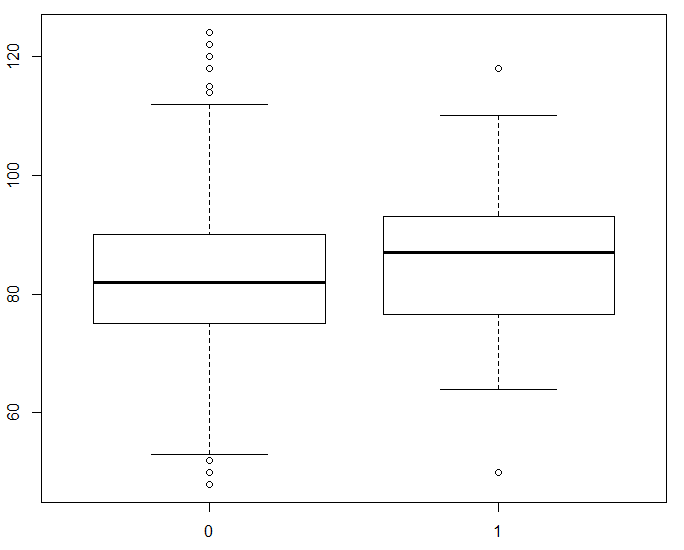
Waist



Hip



Bp.1s



Bp.1d

Na imagem acima descrita podemos verificar que apenas o time.ppn não parece ter influência na variável resposta. As restantes parecem afetar pela ordem seguinte:

1. Waist
2. Bp.1s
3. Hip
4. Bp.1d

Sendo que todas afetam positivamente (quanto maior mais probabilidade de ter diabetes).