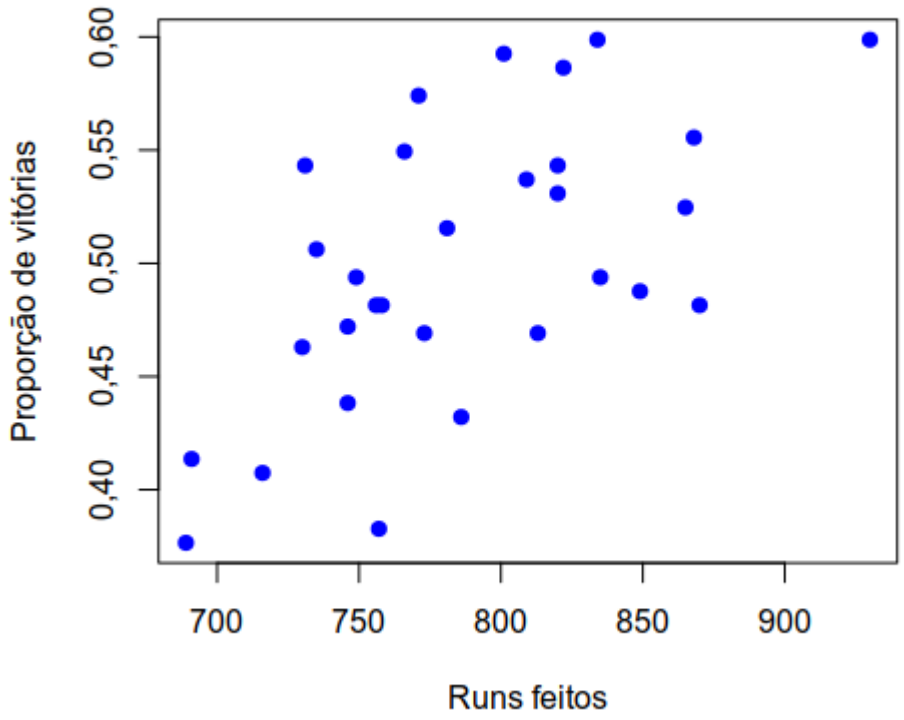
**Tarefa 10**

*Bernardo Chrispim Baron*

2)

a) Gráfico de dispersão (Runs feitos X Proporção de vitórias)



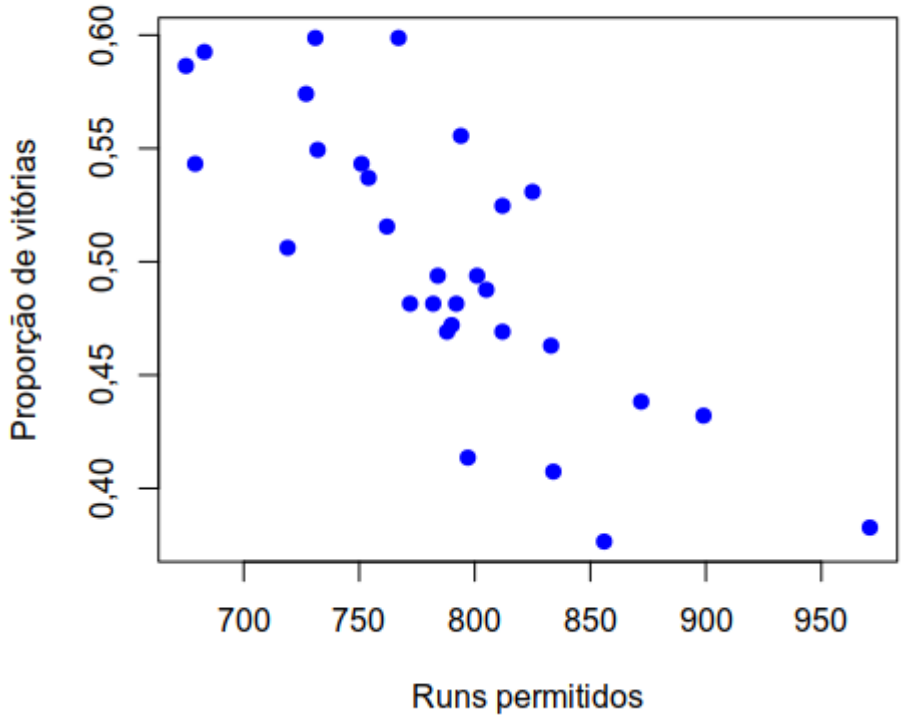

b) O coeficiente de correlação linear de Pearson foi de 0,606 entre as variáveis “runs feitos” e “proporção de vitórias”. O teste de correlação para esse mesmo par de variáveis apresentou p-valor 0,0038, favorecendo a hipótese de que há correlação entre ambas.

c) O ajuste do modelo de regressão linear pelo método dos mínimos quadrados, *descartando o intercepto*, estimou um coeficiente de regressão 0,0006353 para a variável independente “runs feitos”, com p-valor < $2{,}2.10^{-16}$, e obteve um coeficiente de determinação ajustado de 0,99.

d) A análise de resíduos não descartou nenhuma das suposições necessárias a um modelo linear, no que se refere erro aleatório. O Teste de Kolmogorov-Simirnov para os resíduos normalizados apresentou p-valor 0,88, não rejeitando a hipótese de normalidade dos resíduos. O Teste de Goldfeld-Quandt retornou p-valor 0,96, não rejeitando a hipótese de homocedasticidade. Finalmente, o Teste de Durbin-Watson teve p-valor 0,28, tampouco rejeitando a hipótese de independência dos resíduos.

3)

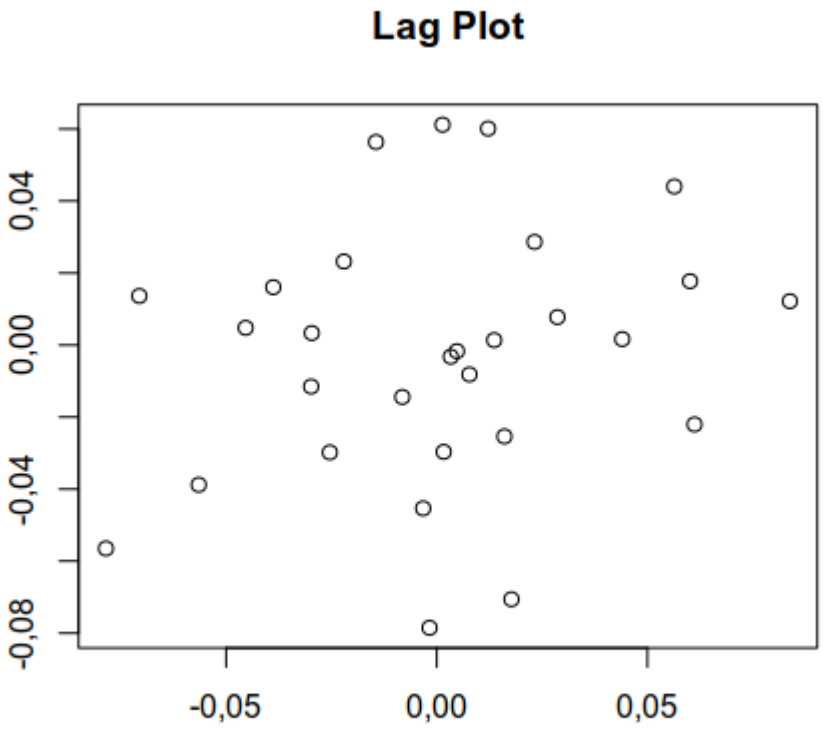
a) Gráfico de dispersão (“Runs permitidos” X “Proporção de vitórias)




b) O coeficiente de correlação linear de Pearson foi de -0,785 entre as variáveis “runs permitidos” e “proporção de vitórias”. O teste de correlação para esse mesmo par de variáveis apresentou p-valor $2{,}8.10^{-7}$, favorecendo a hipótese de que há correlação entre ambas.

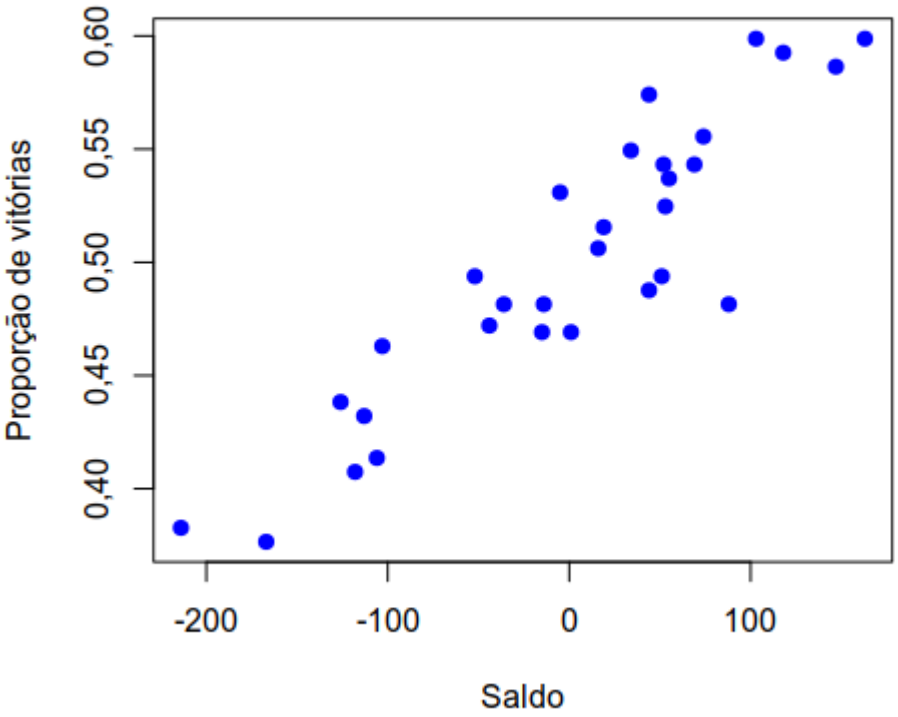
c) O ajuste do modelo de regressão linear pelo método dos mínimos quadrados, estimou um intercepto de 1,099 (p-valor: $8{,}95.10^{-13}$) e um coeficiente de regressão -0,000762 (p-valor < $2{,}8.10^{-7}$) para a variável independente “runs permitidos”, e obteve um coeficiente de determinação ajustado de 0,60.

d) A análise de resíduos não descartou nenhuma das suposições necessárias a um modelo linear, no que se refere erro aleatório. O Teste de Kolmogorov-Simirnov para os resíduos normalizados apresentou p-valor 0,98, não rejeitando a hipótese de normalidade dos resíduos. O Teste de Goldfeld-Quandt retornou p-valor 0,81, não rejeitando a hipótese de homocedasticidade. Finalmente, o Teste de Durbin-Watson teve p-valor 0,064, não sendo suficiente para rejeitar a hipótese de independência dos resíduos a um nível de significância de 5%. Neste último caso, dado o p-valor um pouco mais próximo do nível de significância, a observação do *lag plot* (a seguir) ajuda a verificar visualmente a inexistência de qualquer relação clara entre os resíduos subsequentes, validando a suposição de independência dos resíduos.




4)

a)




b) O coeficiente de correlação linear de Pearson foi de 0,91 entre as variáveis “saldo” e “proporção de vitórias”. O teste de correlação para esse mesmo par de variáveis apresentou p-valor $2{,}7.10^{-12}$, favorecendo a hipótese de que há correlação entre ambas.

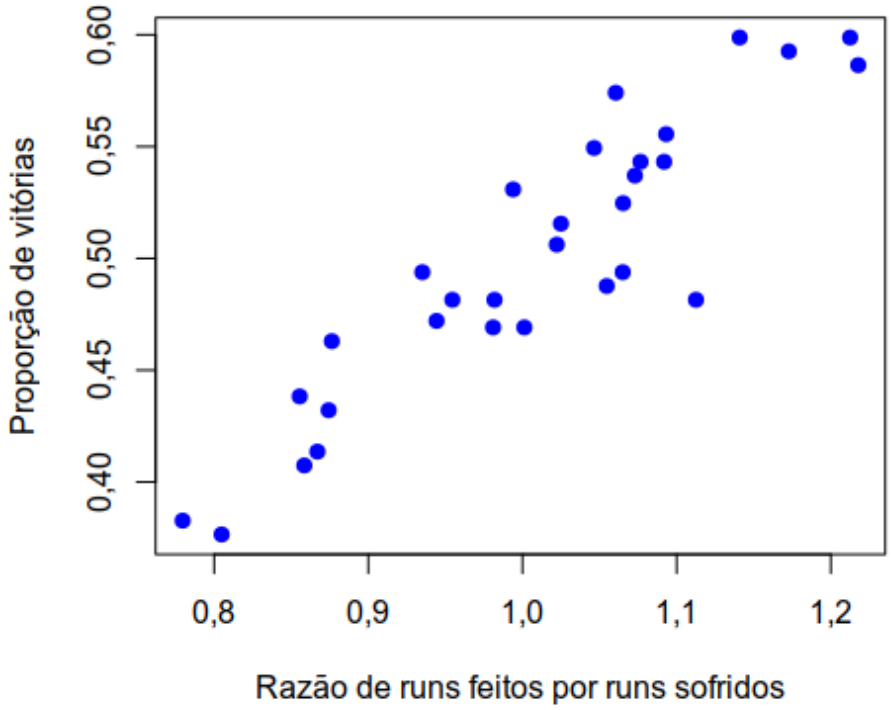
c) O ajuste do modelo de regressão linear pelo método dos mínimos quadrados, estimou um intercepto de 0,4996 (p-valor: $2.10^{-16}$) e um coeficiente de regressão -0,0006083 (p-valor < $2{,}7.10^{-7}$)

para a variável independente “runs permitidos”, e obteve um coeficiente de determinação ajustado de 0,82.

d) A análise de resíduos não descartou nenhuma das suposições necessárias a um modelo linear, no que se refere erro aleatório. O Teste de Kolmogorov-Simirnov para os resíduos normalizados apresentou p-valor 0,76, não rejeitando a hipótese de normalidade dos resíduos. O Teste de Goldfeld-Quandt retornou p-valor 0,95, não rejeitando a hipótese de homocedasticidade. Finalmente, o Teste de Durbin-Watson teve p-valor 0,42, não sendo suficiente para rejeitar a hipótese de independência dos resíduos a um nível de significância de 5%.

5)

a) Todos os modelos apresentaram baixos p-valores no Teste F, demonstrando que as variáveis selecionadas são todas relevantes para estimar o comportamento da variável resposta (“proporção de vitórias”). Assim, tomando como o principal indicador comparar a relevância do modelo o coeficiente de determinação ajustado, o melhor modelo para a proporção de vitórias seria o ajuste a partir da variável “*runs* feitos”, sem intercepto ($R^2$-ajustado: 0,99). Por outro lado, outro indicador relevante é o índice de correlação de Pearson entre as duas variáveis – que, nesse caso, ficou em apenas 0,6. Nesse aspecto, o saldo parece ser uma variável com melhor correlação (0,911), atingindo também um coeficiente de determinação ajustado bastante razoável (0,82). Por fim, também se testou um modelo linear para a variável “proporção de vitórias” em função da taxa de *runs* feitos por *runs* permitidos (gráfico de dispersão a seguir), que apresenta ligeira melhora em relação ao saldo tanto para o coeficiente de correlação de Pearson (0,915) quanto para o coeficiente de determinação ajustado (0,83).



b) Levando em consideração o módulo dos coeficientes de regressão obtidos pelos modelos da proporção de vitórias em função do número de *runs* feitos e do número de *runs* sofridos, a proporção de vitórias é influenciada com mais intensidade pela capacidade de não sofrer *runs* do que pela capacidade de fazê-los. Isso ocorre porque, segundo os modelos propostos, cada *run* sofrido reduz em 0,076% a proporção de vitórias de um time, ao passo que cada *run* feito aumenta essa mesma proporção em apenas 0,064%.

6) Dependendo do modelo utilizado, a proporção de vitórias estimada para um time que realize 844 *runs* e sofra 722 varia de 53,6% de vitórias (com 95% de chance da proporção verdadeira estar entre 43,3 e 63,9%), para o ajuste em função dos *runs* feitos; até 57,95% de vitórias (com 95% de certeza de a proporção verdadeira estar entre 52,5% e 63,4% de vitorias), no caso do ajuste proposto, em função da taxa de *runs* feitos por *runs* sofridos.