Brasileiro_2021_Regression

Eduardo Cecconi

1/7/2022

Regressão Múltipla

Dados do Campeonato Brasileiro 2021 extraídos do InStat

Este artigo encerra uma sequência de três publicações com as quais demonstrei algumas habilidades adquiridas em programação na linguagem R e em modelagem estatística. Com base em dados de 379 jogos do Campeonato Brasileiro 2021 extraídos do provedor InStat, após aplicar técnicas de Análise Fatorial Confirmatória e Análise Fatorial Exploratória, agora apresento uma Regressão Múltipla.

Os processos iniciais são exatamente os mesmos dos artigos anteriores: data wrangling, criação de uma variável (Local da partida) que serve de parâmetro para outras quatro novas variáveis ausentes do banco de dados original, e posterior agrupamento das informações em métricas.

Ultrapassados os três blocos de código com a manipulação dos dados, o processo de regressão múltipla será detalhado. Nestes blocos constam a **criação de 5 variáveis** (Local, Gols Concedidos, Chutes Concedidos, Finalizações Concedidas em Bolas Paradas e Pontos Conquistados), a **criação de 11 métricas**, a seleção de um total de **24 variáveis** em um novo banco de dados, o **ajuste das escalas de mensuração**, o **arredondamento das casas decimais** e as **matrizes de variâncias, covariâncias e correlações**.

Os fatores testados são os mesmos da análise confirmatória: **ataque, defesa, posse e bolas paradas**.

Regressão de ATAQUE

Em primeiro lugar, é válido ponderar que o método de regressão não me parece o mais apropriado para lidar com dados de futebol. Nestes estudos iniciais a técnica Exploratória se mostrou mais ajustada às demandas da modalidade. Como a regressão utiliza uma variável dependente, relacionando as demais em busca de um modelo preditivo, os resultados dos testes tendem a ser baixos, assim como já aconteceu na técnica Confirmatória. Por quê? Porque o futebol é influenciado por acontecimentos imprevistos que tornam difícil prever resultados a partir de comportamentos técnico-táticos.

Como disse antes, e todos já sabem, não é incomum a equipe dominante em todos os aspectos ser derrotada por um adversário inferior que marcou o gol da vitória em um único chute e conseguiu manter a invencibilidade na defesa mesmo concedendo mais de 20 finalizações, por exemplo.

No entanto, os métodos de regressão são promissores se utilizados para elaborar probabilidades em sites de apostas esportivas, por exemplo, o que não é o foco desta análise.

Voltando ao tema, o primeiro teste de regressão foi utilizado no indicador de **Ataque**. Tanto este indicador como os demais terão a variável Pontos como parâmetro (**variável dependente**), ou seja, a regressão vai estimar o potencial preditivo das variáveis independentes para a conquista de pontos.

No caso do Ataque, as variáveis independentes selecionadas são as mesmas utilizadas nas técnicas Exploratória e Confirmatória: **Gols, Conversão, Chutes e Passes para Finalização**. Os coeficientes e os p-valores, como esperado, demonstram que o potencial preditivo da fórmula é baixo. Na prática, apenas os gols marcados tiveram relação direta com os pontos conquistados. O Multiple R-Squared foi de 0.4165 (ou seja, a fórmula explica apenas 41.65% dos pontos).

```
##
## Call:
## lm(formula = Points ~ Goals + Conversion + Attacks_Shot + Shot_Assist,
##     data = Metrics_Regression)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8927 -0.5262 -0.2534  0.8550  1.8272
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.529755   0.088531   5.984 3.37e-09 ***
## Goals         0.775675   0.052418  14.798  < 2e-16 ***
## Conversion    0.002069   0.052248   0.040    0.968
## Attacks_Shot  0.057647   0.045029   1.280    0.201
## Shot_Assist  -0.006382   0.011598  -0.550    0.582
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9806 on 753 degrees of freedom
## Multiple R-squared:  0.4165, Adjusted R-squared:  0.4134
## F-statistic: 134.4 on 4 and 753 DF,  p-value: < 2.2e-16
```

```
## Start:  AIC=-24.66
## Points ~ Goals + Conversion + Attacks_Shot + Shot_Assist
##
##                Df Sum of Sq    RSS     AIC
## - Conversion    1     0.002 724.13 -26.654
## - Shot_Assist   1     0.291 724.42 -26.351
## - Attacks_Shot  1     1.576 725.70 -25.008
## <none>                      724.12 -24.656
## - Goals         1   210.580 934.70 166.836
##
## Step:  AIC=-26.65
## Points ~ Goals + Attacks_Shot + Shot_Assist
##
##                Df Sum of Sq     RSS    AIC
## - Shot_Assist   1      0.30  724.43 -28.34
## - Attacks_Shot  1      1.66  725.78 -26.92
## <none>                      724.13 -26.65
## - Goals         1    475.76 1199.89 354.15
##
```

```
## Step:  AIC=-28.34
## Points ~ Goals + Attacks_Shot
##
##                Df Sum of Sq     RSS    AIC
## - Attacks_Shot  1      1.38  725.81 -28.89
## <none>                       724.43 -28.34
## - Goals         1    475.47 1199.90 352.16
##
## Step:  AIC=-28.89
## Points ~ Goals
##
##         Df Sum of Sq     RSS    AIC
## <none>                 725.81 -28.89
## - Goals  1    515.14 1240.95 375.66

##
## Call:
## lm(formula = Points ~ Goals, data = Metrics_Regression)
##
## Coefficients:
## (Intercept)        Goals
##      0.4808       0.7863
```

Com o método **Stepwise**, a fórmula foi reduzida à relação apenas entre Pontos e Gols, reiterando que estatisticamente o ditado popular se confirma: "futebol é bola na rede". Entretanto, como também afirmei nos artigos anteriores, a elaboração de métricas e indicadores no futebol acima de tudo pretende **identificar padrões de comportamento e traduzir o modelo/estilo de jogo**. Ao invés de prever resultados, **queremos prever posicionamentos, movimentos e padrões de tomadas de decisão**, o que aí sim nos permitirá **criar soluções para antecipar ameaças e oportunidades que os adversários têm a oferecer**, e como podemos nos **prevenir/aproveitar** delas - assim como para **aperfeiçoar** o nosso modelo de jogo (sabendo o que queremos executar, também saberemos como modelar os dados para **corrigir** o que não está funcionando, **aprimorar** o que está funcionando, e **ajustar estrategicamente** o modelo ao que sabemos dos **adversários a cada jogo**).

Utilizando **Gols como a variável** dependente, em substituição aos pontos, o Multiple R-Squared sobe para 58% (ainda baixo), e o método stepwise não sugere alteração à fórmula, que relaciona Conversão, Chutes certos e Passes para finalização.

```
##
## Call:
## lm(formula = Goals ~ Conversion + Attacks_Shot + Shot_Assist,
##     data = Metrics_Regression)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7524 -0.3512 -0.0783  0.2819  3.1932
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.221605   0.060976   3.634 0.000298 ***
## Conversion   0.742982   0.024198  30.704  < 2e-16 ***
## Attacks_Shot 0.260100   0.029816   8.723  < 2e-16 ***
## Shot_Assist  0.028518   0.007991   3.569 0.000381 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6813 on 754 degrees of freedom
## Multiple R-squared:  0.5799, Adjusted R-squared:  0.5782
## F-statistic: 346.9 on 3 and 754 DF,  p-value: < 2.2e-16
```

```
## Start:  AIC=-577.76
## Goals ~ Conversion + Attacks_Shot + Shot_Assist
##
##                Df Sum of Sq    RSS     AIC
## <none>                      349.99 -577.76
## - Shot_Assist   1      5.91 355.90 -567.07
## - Attacks_Shot  1     35.32 385.31 -506.88
## - Conversion    1    437.60 787.59   35.03
```

```
##
## Call:
## lm(formula = Goals ~ Conversion + Attacks_Shot + Shot_Assist,
##     data = Metrics_Regression)
##
## Coefficients:
##  (Intercept)    Conversion  Attacks_Shot   Shot_Assist
##      0.22161       0.74298       0.26010       0.02852
```

Regressão de DEFESA

O potencial preditivo da fórmula de defesa é ainda mais baixo que o da fórmula de ataque. Na prática, apenas os gols sofridos tiveram relação inversa com os pontos conquistados, seguindo o que aconteceu no indicador anterior (somente os gols feitos se relacionaram com os pontos).

O Multiple R-Squared foi de 0.3427 (ou seja, **a fórmula explica apenas 34.3% dos pontos**). Foram relacionados Gols concedidos, Conversão concedida, Pressão, PPDA, Contenção, e Chutes concedidos.

```
##
## Call:
## lm(formula = Points ~ Goals_Conceded + Pressing + Conversion_Conceded +
##     Contention + PPDA + Shots_Conceded, data = Metrics_Regression)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9466 -0.8952 -0.2019  0.8256  3.0680
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.119503   0.060929  34.786   <2e-16 ***
## Goals_Conceded      -0.676240   0.060354 -11.205   <2e-16 ***
## Pressing             0.008311   0.041869   0.199   0.8427
## Conversion_Conceded -0.018643   0.054027  -0.345   0.7301
## Contention           0.105448   0.042179   2.500   0.0126 *
## PPDA                 0.095681   0.040581   2.358   0.0186 *
```

```
## Shots_Conceded      -0.016232   0.047033  -0.345   0.7301
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.042 on 751 degrees of freedom
## Multiple R-squared:  0.3427, Adjusted R-squared:  0.3374
## F-statistic: 65.25 on 6 and 751 DF,  p-value: < 2.2e-16

## Start:  AIC=69.64
## Points ~ Goals_Conceded + Pressing + Conversion_Conceded + Contention +
##     PPDA + Shots_Conceded
##
##                       Df Sum of Sq    RSS     AIC
## - Pressing             1     0.043 815.77  67.676
## - Conversion_Conceded  1     0.129 815.86  67.757
## - Shots_Conceded       1     0.129 815.86  67.757
## <none>                             815.73  69.636
## - PPDA                 1     6.038 821.77  73.226
## - Contention           1     6.789 822.52  73.918
## - Goals_Conceded       1   136.363 952.09 184.808
##
## Step:  AIC=67.68
## Points ~ Goals_Conceded + Conversion_Conceded + Contention +
##     PPDA + Shots_Conceded
##
##                       Df Sum of Sq    RSS     AIC
## - Conversion_Conceded  1     0.112 815.88  65.780
## - Shots_Conceded       1     0.173 815.95  65.837
## <none>                             815.77  67.676
## - PPDA                 1     6.600 822.37  71.784
## - Contention           1     7.792 823.56  72.882
## - Goals_Conceded       1   137.081 952.85 183.413
##
## Step:  AIC=65.78
## Points ~ Goals_Conceded + Contention + PPDA + Shots_Conceded
##
##                   Df Sum of Sq     RSS     AIC
## - Shots_Conceded  1     0.111  816.00  63.884
## <none>                         815.88  65.780
## - PPDA            1     6.526  822.41  69.819
## - Contention      1     7.795  823.68  70.988
## - Goals_Conceded  1   282.328 1098.21 289.033
##
## Step:  AIC=63.88
## Points ~ Goals_Conceded + Contention + PPDA
##
##                   Df Sum of Sq     RSS    AIC
## <none>                         816.00  63.88
## - PPDA            1      6.62  822.62  68.01
## - Contention      1      7.69  823.68  68.99
## - Goals_Conceded  1    385.28 1201.28 355.03

##
## Call:
```

```
## lm(formula = Points ~ Goals_Conceded + Contention + PPDA, data = Metrics_Regression)
##
## Coefficients:
##    (Intercept)  Goals_Conceded      Contention            PPDA
##         2.1251         -0.6974          0.1061          0.0974
```

O método stepwise sugere a **retirada dos itens Pressing e Shots_Conceded**, mantendo Gols_Conceded, Contention e PPDA, o que praticamente não intereferiu no Multiple R-Squared (34.24%) - ou seja, os itens exluídos não se correlacionaram com os pontos.

```
##
## Call:
## lm(formula = Points ~ Goals_Conceded + Contention + PPDA, data = Metrics_Regression)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9275 -0.8966 -0.2041  0.8314  3.0500
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.12512    0.05573  38.134  < 2e-16 ***
## Goals_Conceded -0.69743    0.03696 -18.868  < 2e-16 ***
## Contention      0.10606    0.03979   2.666  0.00785 **
## PPDA            0.09740    0.03937   2.474  0.01359 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.04 on 754 degrees of freedom
## Multiple R-squared:  0.3424, Adjusted R-squared:  0.3398
## F-statistic: 130.9 on 3 and 754 DF,  p-value: < 2.2e-16
```

Ao contrário do indicador de ataque, quando a **variável dependente** passa a ser os **Gols concedidos**, a fórmula se fortalece, alcançando **64.21% de relação** com as variáveis preditoras. O método stepwise são sugeridas as **exclusões de PPDA e Pressing**. Ou seja, a fórmula **não encontrou relação preditora entre dois itens de pressão) e a eficiência defensiva**.

```
##
## Call:
## lm(formula = Goals_Conceded ~ Pressing + Conversion_Conceded +
##     Contention + PPDA + Shots_Conceded, data = Metrics_Regression)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.27059 -0.37055 -0.04097  0.32459  2.56319
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.50467    0.03188  15.828  < 2e-16 ***
## Pressing            -0.03943    0.02526  -1.561 0.118900
## Conversion_Conceded  0.62981    0.02320  27.150  < 2e-16 ***
## Contention          -0.09173    0.02526  -3.631 0.000302 ***
## PPDA                 0.03318    0.02449   1.355 0.175875
```

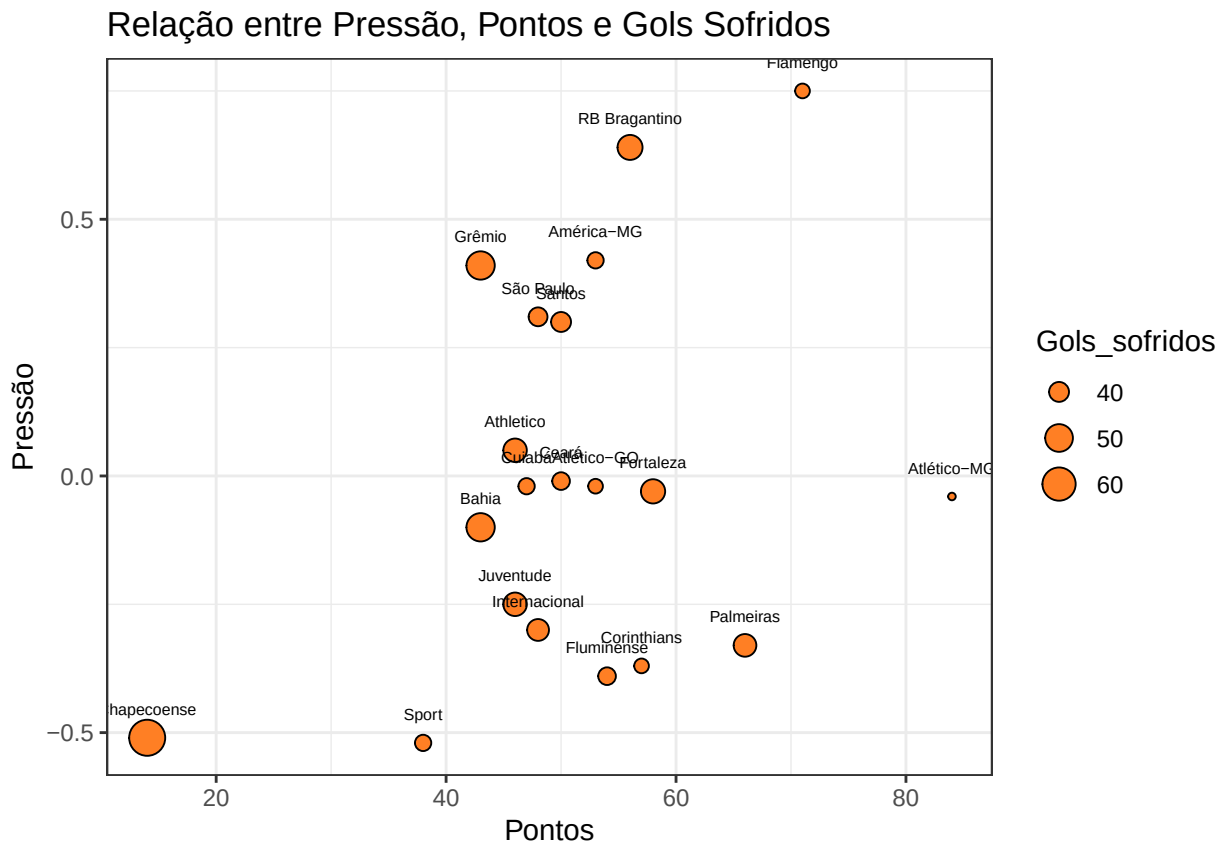
```
## Shots_Conceded       0.40510    0.02428  16.687  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6297 on 752 degrees of freedom
## Multiple R-squared:  0.6421, Adjusted R-squared:  0.6397
## F-statistic: 269.8 on 5 and 752 DF,  p-value: < 2.2e-16

## Start:  AIC=-695.17
## Goals_Conceded ~ Pressing + Conversion_Conceded + Contention +
##     PPDA + Shots_Conceded
##
##                       Df Sum of Sq    RSS     AIC
## - PPDA                 1     0.728 298.92 -695.32
## <none>                             298.19 -695.17
## - Pressing             1     0.966 299.16 -694.72
## - Contention           1     5.227 303.42 -684.00
## - Shots_Conceded       1   110.412 408.60 -458.40
## - Conversion_Conceded  1   292.286 590.48 -179.31
##
## Step:  AIC=-695.32
## Goals_Conceded ~ Pressing + Conversion_Conceded + Contention +
##     Shots_Conceded
##
##                       Df Sum of Sq    RSS     AIC
## - Pressing             1     0.663 299.58 -695.65
## <none>                             298.92 -695.32
## - Contention           1     7.020 305.94 -679.73
## - Shots_Conceded       1   111.757 410.68 -456.56
## - Conversion_Conceded  1   296.590 595.51 -174.88
##
## Step:  AIC=-695.65
## Goals_Conceded ~ Conversion_Conceded + Contention + Shots_Conceded
##
##                       Df Sum of Sq    RSS     AIC
## <none>                             299.58 -695.65
## - Contention           1      8.77 308.35 -675.77
## - Shots_Conceded       1    127.72 427.30 -428.48
## - Conversion_Conceded  1    299.30 598.88 -172.60

##
## Call:
## lm(formula = Goals_Conceded ~ Conversion_Conceded + Contention +
##     Shots_Conceded, data = Metrics_Regression)
##
## Coefficients:
##         (Intercept)  Conversion_Conceded           Contention
##              0.5066               0.6278              -0.1098
##      Shots_Conceded
##              0.4161
```

Sem Pressing e PPDA, o Multiple R-Squared praticamente não se alterou (0.6404)

##

```
## Call:
## lm(formula = Goals_Conceded ~ Conversion_Conceded + Contention +
##     Shots_Conceded, data = Metrics_Regression)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.25266 -0.35973 -0.04701  0.29851  2.54570
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.50658    0.03169  15.988  < 2e-16 ***
## Conversion_Conceded  0.62779    0.02287  27.446  < 2e-16 ***
## Contention          -0.10979    0.02337  -4.698 3.12e-06 ***
## Shots_Conceded       0.41607    0.02321  17.929  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6303 on 754 degrees of freedom
## Multiple R-squared:  0.6404, Adjusted R-squared:  0.639
## F-statistic: 447.6 on 3 and 754 DF,  p-value: < 2.2e-16
```

Como o indicador de **Pressão** foi excluído em todos os cenários, produzi um gráfico de dispersão relacionando essa métrica (relembrando, Pressing é composta por três itens de scout coletados pelo InStat: **Team Pressing + High Pressing Accurate + Recoveries High**) com **Pontos e** Gols sofridos**:



Primeira constatação é a concentração de 15 equipes no intervalo entre 40 e 60 pontos,

separando o extremo inferior com Chapecoense e Sport (ambas equipes com os mais baixos índices de pressão) e o extremo superior de desempenho com Atlético-MG, Flamengo e Palmeiras (as três com distintas abordagens de pressão).

Enquanto o Atlético-MG sofreu poucos gols mesmo alcançando indicador de pressão abaixo da média, Bragantino e Grêmio concederam muitos gols mesmo com valores altos de pressão. Fica evidente como esta métrica não se relacionou com a eficiência defensiva de maneira significante no Brasileiro 2021, diante da diversidade de combinações que vemos no gráfico ao cruzarmos as três variáveis (pressão, pontos e gols concedidos).

Regressão de POSSE

Partindo-se do que foi observado em ataque e defesa, o resultado da regressão que associa as métricas de posse com os pontos conquistados não surpreende: Multiple R-Squared de apenas 0.06468.

Ou seja, **a fórmula de posse explica menos de 7% dos pontos conquistados**. Isso significa dizer que estatiscamente importa pouco a maneira como as equipes lidam com a posse, desde que o produto final (gols feitos > gols concedidos) seja alcançado. Eficiência ofensiva e defensiva, tão somente.

Fica evidente que a qualidade da posse não influenciou resultados noBrasileiro 2021, e suas estatísticas são mais úteis para identificar estilos/modelos de jogo.

```
##
## Call:
## lm(formula = Points ~ Possession_Time + Passing_Speed + Ball_Care +
##     Build_Up + Progression + Imposition, data = Metrics_Regression)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2213 -1.0835 -0.3214  1.3607  2.4101
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.57781    1.18362   1.333    0.183
## Possession_Time  0.07474    1.22626   0.061    0.951
## Passing_Speed   -0.09089    0.27122  -0.335    0.738
## Ball_Care        0.29491    0.11573   2.548    0.011 *
## Build_Up        -0.33253    0.06919  -4.806 1.86e-06 ***
## Progression     -0.26220    0.05956  -4.403 1.22e-05 ***
## Imposition       0.06448    0.04793   1.345    0.179
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.243 on 751 degrees of freedom
## Multiple R-squared:  0.06468,    Adjusted R-squared:  0.05721
## F-statistic: 8.656 on 6 and 751 DF,  p-value: 3.999e-09
```

```
## Start:  AIC=336.97
## Points ~ Possession_Time + Passing_Speed + Ball_Care + Build_Up +
##     Progression + Imposition
##
```

```
##                   Df Sum of Sq    RSS    AIC
## - Possession_Time  1     0.006 1160.7 334.97
## - Passing_Speed    1     0.174 1160.9 335.08
## - Imposition       1     2.797 1163.5 336.79
## <none>                         1160.7 336.97
## - Ball_Care        1    10.036 1170.7 341.50
## - Progression      1    29.956 1190.6 354.29
## - Build_Up         1    35.700 1196.4 357.93
##
## Step:  AIC=334.97
## Points ~ Passing_Speed + Ball_Care + Build_Up + Progression +
##     Imposition
##
##                 Df Sum of Sq    RSS    AIC
## - Passing_Speed  1     0.245 1160.9 333.13
## - Imposition     1     2.869 1163.6 334.85
## <none>                       1160.7 334.97
## - Ball_Care      1    22.425 1183.1 347.48
## - Progression    1    35.797 1196.5 356.00
## - Build_Up       1    35.980 1196.7 356.11
##
## Step:  AIC=333.13
## Points ~ Ball_Care + Build_Up + Progression + Imposition
##
##               Df Sum of Sq    RSS    AIC
## - Imposition   1     2.858 1163.8 333.00
## <none>                     1160.9 333.13
## - Ball_Care    1    24.349 1185.3 346.87
## - Progression  1    35.563 1196.5 354.01
## - Build_Up     1    35.735 1196.7 354.11
##
## Step:  AIC=333
## Points ~ Ball_Care + Build_Up + Progression
##
##               Df Sum of Sq    RSS    AIC
## <none>                     1163.8 333.00
## - Ball_Care    1    24.595 1188.4 346.85
## - Progression  1    32.705 1196.5 352.01
## - Build_Up     1    39.087 1202.9 356.04

##
## Call:
## lm(formula = Points ~ Ball_Care + Build_Up + Progression, data = Metrics_Regression)
##
## Coefficients:
## (Intercept)    Ball_Care     Build_Up  Progression
##      1.3522       0.2892      -0.3420      -0.2382
```

O método stepwise sugere a **retirada de 3 métricas: imposição, tempo de posse e velocidade de circulação (tempo de retenção)**, reiterando a percepção anterior (controle da posse não foi parâmetro de sucesso no Brasileiro). Além disso, assim como nos métodos aplicados anteriormente (análises fatoriais exploratória e confirmatória), o indicador que estima a VEL foi descartado pela regressão, uma quebra de paradigma interessante. O Multiple R-Squared foi de 0.06218.

```
##
## Call:
## lm(formula = Points ~ Ball_Care + Build_Up + Progression, data = Metrics_Regression)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2126 -1.0771 -0.3091  1.3636  2.4339
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.35219    0.04513  29.965  < 2e-16 ***
## Ball_Care    0.28924    0.07246   3.992 7.20e-05 ***
## Build_Up    -0.34199    0.06796  -5.032 6.07e-07 ***
## Progression -0.23816    0.05174  -4.603 4.88e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.242 on 754 degrees of freedom
## Multiple R-squared:  0.06218,    Adjusted R-squared:  0.05844
## F-statistic: 16.66 on 3 and 754 DF,  p-value: 1.722e-10
```

Apenas por curiosidade, rodei as regressões relacionando as métricas selecionadas pelo stepwise com Gols feitos e Gols concedidos, ambos em substituição aos pontos. No **ataque**, as métricas de posse mantiveram relação semelhante à obtida com os pontos (**aproximadamente 6%**), enquanto na **defesa** a posse previu **pouco mais de 1%** do sucesso.

```
##
## Call:
## lm(formula = Goals ~ Ball_Care + Build_Up + Progression, data = Metrics_Regression)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8105 -0.8716 -0.1277  0.6902  3.9320
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.10812    0.03691  30.019  < 2e-16 ***
## Ball_Care    0.28667    0.05927   4.836 1.60e-06 ***
## Build_Up    -0.21537    0.05559  -3.874 0.000116 ***
## Progression -0.25898    0.04233  -6.119 1.51e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.016 on 754 degrees of freedom
## Multiple R-squared:  0.06518,    Adjusted R-squared:  0.06146
## F-statistic: 17.53 on 3 and 754 DF,  p-value: 5.247e-11
```

```
##
## Call:
## lm(formula = Goals_Conceded ~ Ball_Care + Build_Up + Progression,
##     data = Metrics_Regression)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4710 -1.0024 -0.1242  0.8110  3.9744
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.10818    0.03788  29.253  < 2e-16 ***
## Ball_Care   -0.08019    0.06083  -1.318  0.18780
## Build_Up     0.18147    0.05705   3.181  0.00153 **
## Progression -0.01221    0.04343  -0.281  0.77878
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.043 on 754 degrees of freedom
## Multiple R-squared:  0.01553,    Adjusted R-squared:  0.01161
## F-statistic: 3.965 on 3 and 754 DF,  p-value: 0.008067
```

Regressão de Bolas Paradas

As bolas paradas, embora relevantes na análise do modelo de jogo adversário, principalmente em vídeos (identificando padrões defensivos e ofensivos), estatisticamente **não representaram nenhuma relevância na projeção de pontos conquistados**. O Multiple R-Squared ficou em apenas 0.03488 (**menos de 4%** dos dados de bolas paradas explicam o sucesso das equipes), e o stepwise **recomendou não ser utilizada esta fórmula**.

Vale lembrar que na **Análise Fatorial Exploratória** o teste **também não sugeriu** que os poucos dados de bolas paradas disponíveis fossem utilizados.

```
##
## Call:
## lm(formula = Points ~ Set_Piece_shot + Set_Pieces_Conceded, data = Metrics_Regression)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5850 -1.3039 -0.3526  1.5987  1.8547
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.33380    0.12585  10.598   <2e-16 ***
## Set_Piece_shot       0.03062    0.02615   1.171    0.242
## Set_Pieces_Conceded -0.02435    0.02615  -0.931    0.352
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.28 on 755 degrees of freedom
## Multiple R-squared:  0.003488,   Adjusted R-squared:  0.0008481
## F-statistic: 1.321 on 2 and 755 DF,  p-value: 0.2674

## Start:  AIC=377.01
## Points ~ Set_Piece_shot + Set_Pieces_Conceded
##
##                        Df Sum of Sq    RSS    AIC
## - Set_Pieces_Conceded  1    1.4204 1238.0 375.88
```

```
## - Set_Piece_shot       1    2.2462 1238.9 376.38
## <none>                              1236.6 377.01
##
## Step:  AIC=375.88
## Points ~ Set_Piece_shot
##
##                  Df Sum of Sq  RSS    AIC
## - Set_Piece_shot  1    2.9078 1241 375.66
## <none>                        1238 375.88
##
## Step:  AIC=375.66
## Points ~ 1


##
## Call:
## lm(formula = Points ~ 1, data = Metrics_Regression)
##
## Coefficients:
## (Intercept)
##       1.352
```

Regressão Múltipla - modelo utilizado na Análise Fatorial Confirmatória

Considerando-se os **quatro fatores do modelo original** testado na análise confirmatória, a regressão múltipla alcançou Multiple R-Squared de 0.571 (ou seja, a fórmula **explica 57,1% dos pontos conquistados**). O método stepwise **não recomendou nenhuma alteração**.

É curioso como na regressão múltipla a **defesa alcançou o maior coeficiente** (0.19107, contra 0.1046 do ataque), enquanto a **posse teve relação inversamente proporcional aos pontos conquistados** (-0.10506).

```
##
## Call:
## lm(formula = Points ~ Attack + Defense + Possession + Set_Piece,
##     data = KPI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6186 -0.6900 -0.1269  0.7626  2.9260
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.249905   0.085285  14.656  < 2e-16 ***
## Attack       0.104161   0.008062  12.920  < 2e-16 ***
## Defense      0.191072   0.011603  16.467  < 2e-16 ***
## Possession  -0.105063   0.013465  -7.802 2.03e-14 ***
## Set_Piece   -0.021941   0.014721  -1.490    0.137
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9943 on 753 degrees of freedom
## Multiple R-squared:  0.4002, Adjusted R-squared:  0.397
## F-statistic: 125.6 on 4 and 753 DF,  p-value: < 2.2e-16
```

```
## Start:  AIC=-3.75
## Points ~ Attack + Defense + Possession + Set_Piece
##
##              Df Sum of Sq     RSS     AIC
## <none>                     744.37  -3.754
## - Set_Piece   1     2.196  746.57  -3.521
## - Possession  1    60.179  804.55  53.176
## - Attack      1   165.022  909.39 146.026
## - Defense     1   268.068 1012.44 227.390

##
## Call:
## lm(formula = Points ~ Attack + Defense + Possession + Set_Piece,
##     data = KPI)
##
## Coefficients:
## (Intercept)       Attack      Defense   Possession    Set_Piece
##     1.24990      0.10416      0.19107     -0.10506     -0.02194
```

Se retirarmos de cada fator as métricas sugeridas nos métodos stepwise realizados durante o passo-a-passo anterior (considerando apenas gols em ataque, excluindo Pressing e PPDA na defesa, tempo de posse e velocidade de circulação na posse, e também excluindo o fator de bolas paradas), o Multiple R-Squared subiu para 0.705 (**70,5% dos pontos foram explicados pela fórmula**).

No entanto, nesta segunda versão **o ataque concentra toda a força preditiva** (mais uma vez a estatística encontra o dito popular *"futebol é bola na rede"*). O método stepwise sugere inclusive a exclusão do fator posse, resumindo todo o banco de dados - na relação com os pontos conquistados - praticamente a uma única variável (gols marcados).

```
##
## Call:
## lm(formula = Points ~ Attack + Defense + Possession, data = KPI_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8479 -0.5135 -0.1125  0.4820  1.9950
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.990686   0.041321  23.976   <2e-16 ***
## Attack       0.786413   0.024302  32.360   <2e-16 ***
## Defense      0.246740   0.009197  26.829   <2e-16 ***
## Possession  -0.014063   0.010306  -1.365    0.173
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6969 on 754 degrees of freedom
## Multiple R-squared:  0.705,  Adjusted R-squared:  0.7038
## F-statistic: 600.5 on 3 and 754 DF,  p-value: < 2.2e-16

## Start:  AIC=-543.57
## Points ~ Attack + Defense + Possession
```

```
##
##              Df Sum of Sq    RSS     AIC
## - Possession  1      0.90 367.05 -543.70
## <none>                    366.14 -543.57
## - Defense     1    349.54 715.69  -37.54
## - Attack      1    508.50 874.64  114.49
##
## Step:  AIC=-543.7
## Points ~ Attack + Defense
##
##           Df Sum of Sq    RSS     AIC
## <none>                 367.05 -543.70
## - Defense  1    358.77 725.81  -28.89
## - Attack   1    520.20 887.25  123.34

##
## Call:
## lm(formula = Points ~ Attack + Defense, data = KPI_2)
##
## Coefficients:
## (Intercept)       Attack      Defense
##      0.9896       0.7902       0.2482
```

Concluo que, com este banco de dados e com as métricas desenvolvidas a partir dele, o método de regressão - seja linear, seja múltipla - não se mostrou relevante para se utilizar como meio de análise do modelo de jogo. Não foi testado neste trabalho, mas acredito que seja uma técnica promissora para aplicar nos algoritmos de servidores de apostas esportivas, e também na indústria de games, porém no mercado dos clubes a Análise Fatorial Exploratória apresentou resultados mais relevantes, porque conseguiu extrair informação associada à maneira como as equipes jogam sem deixar de medir a eficiência na execução dos processos técnico-táticos, além de adaptar o banco de dados e os conceitos teóricos que baseiam a modelação ao contexto da competição.