



Aleatoriedade e Previsão

Modelagem

Modelagem busca construir uma abstração do fenômeno, a partir de dados existentes ou conhecimento tácito, identificando as variáveis relevantes, a representação do fenômeno em si e a formalização da relação entre eles.

Um modelo pode não funcionar por várias razões:

1. O comportamento expresso pelos dados coletados não se repete
2. A cobertura e/ou diversidade dos dados coletados é insuficiente
3. A função que expressa a relação entre as variáveis independentes e a variável dependente não captura a dinâmica do fenômeno
 - a. Funções em geral compreendem as premissas do modelo
4. O processo de construção do modelo é inadequado ou insuficiente
5. Os parâmetros e hiper-parâmetros utilizados não suportam a construção de um modelo de qualidade.

Avaliar um modelo é tão importante quanto construí-lo!



1 Simulando uma partida

Simulando uma partida

- Vamos simular uma partida usando um modelo simples: os gols ocorrem aleatoriamente;

```
#Length of match
match_minutes = 90
#Average goals per match
goals_per_match = 2.79
#Probability of a goal per minute
prob_per_minute = np.array(goals_per_match/match_minutes)
print('The probability of a goal per minute is %5.5f. \n' % prob_per_minute )
```

- Aqui consideramos que a duração do jogo é de 90 minutos (ou seja, ignoramos o tempo em que a bola não está sendo jogada).

Simulando uma partida

- Agora simulamos um único jogo. Os gols acontecem com a mesma probabilidade a cada minuto.

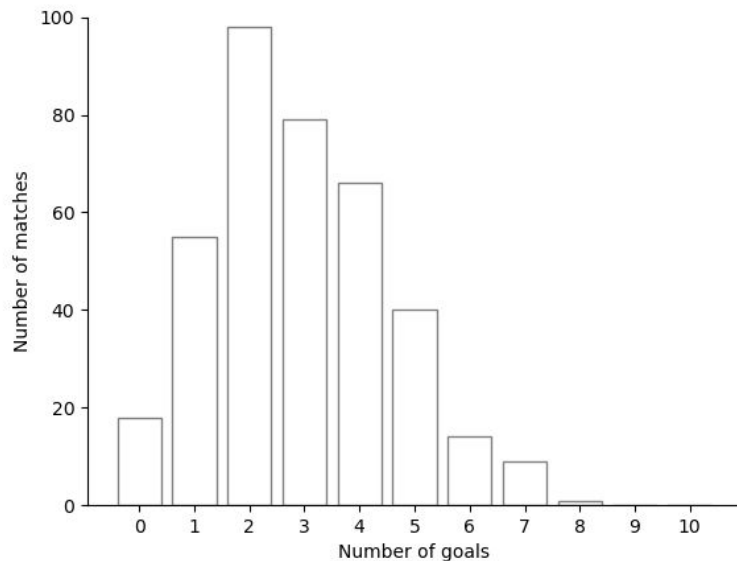
```
goals = 0 #Count of the number of goals
for minute in range(match_minutes):
    r = rnd.rand(1,1) #Generate a random number between 0 and 1.
    #Prints an X when there is a goal and a zero otherwise.
    if (r < probab_per_minute): #Goal - if the random number is less than the goal probability.
        print('X', end = ' ')
        goals = goals+1
        time.sleep(1) #Longer pause
    else:
        print('o', end = ' ')
        time.sleep(0.1) #Short pause
print("\nFinal whistle. \n \nThere were ' + str(goals) + ' goals.')
```

Simulando uma partida

- Se rodarmos o código anterior 10 vezes é provável que cerca de 2 ou 3 de suas simulações tenham terminado com 3 gols.
- A maioria das partidas tem entre 0 e 5 gols.
- É provável que você tenha visto no máximo uma partida com 0 gols.

Simulando gols em uma temporada

- Vamos simular 380 jogos de uma temporada de futebol;
- Veremos o quão bem nosso modelo prevê a distribuição do número de gols;
- O código do slide seguinte faz essa previsão:
 - Fazemos um loop de 380 partidas;
 - Armazenamos o número de gols para cada partida em um array;
 - Plotamos um histograma do número de gols.



Simulando gols em uma temporada

```
def simulateMatch(n, p):
    # n - number of time units; p - probability per time unit of a goal
    goals = 0
    for minute in range(n):
        r = rnd.rand(0, 1) # Generate a random number between 0 and 1
        if (r < p): goals = goals + 1
    return goals

num_matches = 380 # Number of matches
goals = np.zeros(num_matches) # Loop over all the matches and print the number of goals.
for i in range(num_matches):
    goals[i] = simulateMatch(match_minutes, probab_per_minute)

# Create a histogram
fig, ax = plt.subplots(num=1)
histogram_range = np.arange(-0.5, 10.51, 1)
histogram_goals = np.histogram(goals, histogram_range)
ax.bar(histogram_goals[1][:-1] + 0.5, histogram_goals[0], color='white', edgecolor='black', linestyle='-', alpha=0.5)
ax.set_ylim(0, 100)
ax.spines['top'].set_visible(False)
ax.spines['right'].set_visible(False)
ax.set_xticks(np.arange(0, 11, step=1))
ax.set_yticks(np.arange(0, 101, step=20))
ax.set_xlabel('Number of goals')
ax.set_ylabel('Number of matches')
plt.show()
```




2 Aleatoriedade inevitável

Probabilidade de Marcar

O que torna o futebol e outros esportes coletivos emocionantes é sua imprevisibilidade. Se você estiver assistindo a uma partida e desviar o olhar por alguns segundos, pode perder uma jogada importante e um gol repentino.

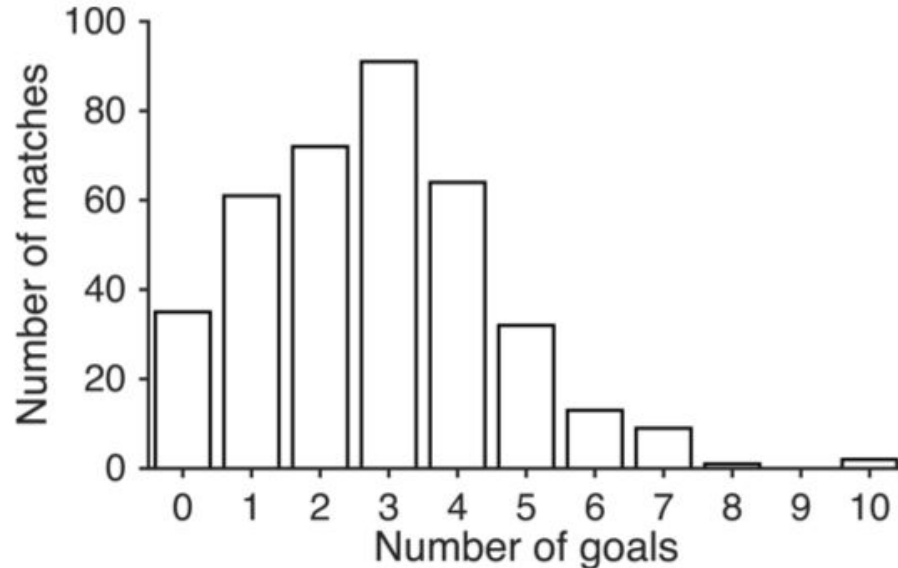


Probabilidade de Marcar

- **É provável que ocorra um gol a qualquer momento durante a partida**
 - Embora existam todos os tipos de fatores que determinam a taxa com que as equipes marcam, o momento dos gols é mais ou menos aleatório.
- Com uma média de 2,79 gols por partida, a probabilidade de haver um gol em um minuto é de $2,79/90=0,031$ (ignoramos os acréscimos neste modelo)
 - Isso significa que a chance de vermos um gol em qualquer minuto escolhido aleatoriamente é de cerca de 1 em 32.

Histograma de gols em uma temporada

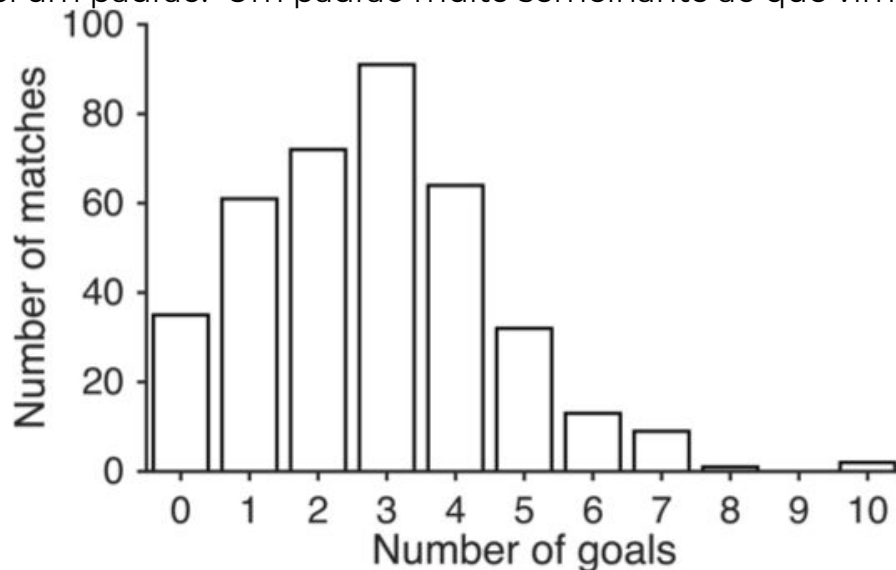
- Vamos analisar a temporada 2012/13 da Premier League.
- A figura é um histograma do número de gols marcados em todos os jogos da temporada.
- A quantidade média de gols marcados foi de 2,79 (como no nosso modelo da seção anterior).



Histograma de gols em uma temporada

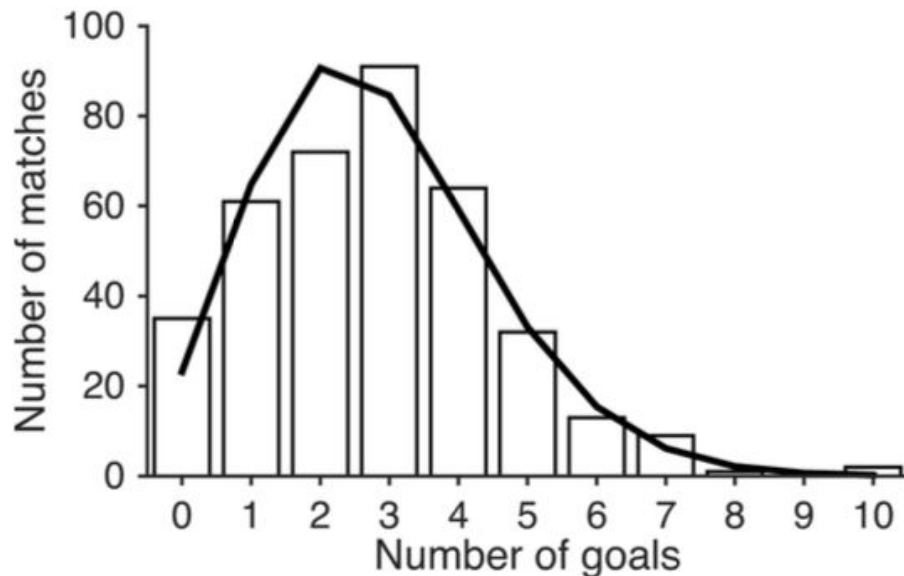
O histograma mostra a frequência com que vários placares ocorreram.


- Houve 35 empates em 0–0;
- O número de gols mais comum era três e, na maioria desses jogos, o placar final era 2–1;
- Já é perceptível um padrão. Um padrão muito semelhante ao que vimos em nossa simulação.



Histograma de gols em uma temporada

O gráfico abaixo compara a simulação da temporada média (linha sólida) com a distribuição de gols (histograma) na Premier League 2012-13.





A correspondência entre modelo e realidade é muito boa. Lembre-se de toda a complexidade em jogo aqui. Todos os gritos do técnico da linha lateral. Os torcedores tentando apoiar seu time ou, na maioria das vezes, dizendo a eles como eles são inúteis. Os pensamentos na cabeça dos jogadores quando eles dizem a si mesmos que agora é sua chance de marcar. Nenhum destes fatores parece afetar a distribuição dos gols marcados. Pelo contrário, são todos esses fatores agindo em conjunto que geram o tipo de aleatoriedade assumido no modelo. Quanto mais fatores envolvidos, maior a aleatoriedade nos gols e melhor a correspondência do nosso histograma simulado com a realidade.

- David Sumpter, Soccermetrics

Histograma de gols em uma temporada

- Assumimos na simulação que nem o número de gols marcados até agora, nem a quantidade de tempo jogado, influenciam a probabilidade de outro gol ser marcado. A distribuição resultante é notavelmente bem-sucedida em capturar a forma geral do histograma de objetivo.
- As previsões do modelo, no entanto, **não são perfeitas**.

Histograma de gols em uma temporada

- Assumimos na simulação que nem o número de gols marcados até agora, nem a quantidade de tempo jogado, influenciam a probabilidade de outro gol ser marcado. A distribuição resultante é notavelmente bem-sucedida em capturar a forma geral do histograma de objetivo.
- As previsões do modelo, no entanto, **não são perfeitas**. Isso é explicado por:
 1. Variações de Modelo
 2. Erros de Modelo

Histograma de gols em uma temporada

1. Variações de Modelo:

- Sempre esperamos variação entre as temporadas.
 - Se você executar uma simulação de temporada algumas vezes, às vezes houve mais partidas com 3 gols do que 2, outras vezes foi o contrário.
 - Às vezes não houve partidas com 8 ou 9 gols, outras vezes sim.
- A segunda maneira na qual o modelo varia da realidade é para partidas de 10 gols.
 - Estes são extremamente improváveis no modelo, mas ocorreram algumas vezes durante a temporada PL de 2012-13.

Histograma de gols em uma temporada

2. Erros de Modelo:

- Podemos perceber mais empates em 0-0 do que o modelo prevê.
 - Nosso modelo é baseado em uma suposição muito simples: que os gols são igualmente prováveis de ocorrer a qualquer momento da partida.
 - Mas esse resultado sugere que a *suposição está errada* em duas situações específicas:
 1. As equipes que empatam em 0 a 0 parecem se contentar com o empate.
 2. Às vezes, as partidas, como a última partida de Alex Ferguson, ficam um pouco loucas e mais gols do que o esperado são marcados.



Histograma de gols em uma temporada

- Observe que o modelo pode fornecer ainda mais insights quando está errado!
- Se estivéssemos criando um modelo de apostas em futebol, essas discrepâncias seriam importantes e gostaríamos de mudar nossas suposições de modelo. Mas para muitas aplicações, basta estar ciente delas.

Distribuição Binomial

Vamos chamar a probabilidade, por minuto, de marcar de p e o número de minutos em uma partida de n . Para nossa simulação de futebol, $p=2,7/95=0,028$ e $n = 95$ minutos.

Para melhor entender essa distribuição vamos responder 3 perguntas:

1. Qual é a probabilidade de não haver gol em n minutos?
2. Qual é a probabilidade de ter exatamente um gol nos primeiros n minutos?
3. De quantas maneiras podemos ter k gols durante 95 minutos?

Distribuição Binomial - Questão 1

Vamos simplificar: vamos imaginar que perdemos os primeiros 5 minutos de uma partida.

Qual é a probabilidade de não haver gol nesse tempo? Bem, a probabilidade de não haver gol no primeiro minuto é **$1-p$** , ou seja, é de **$1-0,028=0,972$** .

A probabilidade de não haver gols nos primeiros dois minutos é então, seguindo a lógica acima,

$$(1 - p) \cdot (1 - p) = (1 - p)^2$$

E, seguindo a mesma regra, a probabilidade de não haver gol após 5 minutos é

$$(1 - p) \cdot (1 - p) \cdot (1 - p) \cdot (1 - p) \cdot (1 - p) = (1 - p)^5$$

No caso do futebol, isso é $(0,972)^5 \approx 0,868$. A probabilidade de você não ter perdido nenhum gol se ligar a TV cinco minutos após o início de uma partida é de 86,8%.

Distribuição Binomial - Questão 1

Para saber a probabilidade de não haver gols em todo o jogo, basta continuar a multiplicar. São 95 vezes, para os 95 minutos. Isto é,

$$\underbrace{(1 - p) \cdot (1 - p) \cdot \dots \cdot (1 - p) \cdot (1 - p)}_{95 \text{ vezes}} = (1 - p)^{95}$$

que é $(0,972)^{95} \approx 0,067$. Há 6,7% de probabilidade de que não haja gol de nenhuma das equipes durante a partida.

Distribuição Binomial - Questão 2

Novamente, vamos simplificar o problema: Qual é a probabilidade de haver exatamente um gol nos primeiros 5 minutos? Usando a notação 'o' para nenhum gol durante um minuto de jogo e 'X' para um gol, existem 5 maneiras diferentes pelas quais um único gol pode ocorrer:

1. Xoooo (gol no primeiro minuto)
2. oXooo (gol no segundo minuto)
3. ooXoo (gol no terceiro minuto)
4. oooXo (gol no quarto minuto)
5. ooooX (gol no quinto minuto)

Distribuição Binomial - Questão 2

Assim, podemos escrever a probabilidade de ocorrer um único gol nos primeiros 5 minutos como:

$$\begin{aligned} & p \cdot (1 - p) \cdot (1 - p) \cdot (1 - p) \cdot (1 - p) + \\ & (1 - p) \cdot p \cdot (1 - p) \cdot (1 - p) \cdot (1 - p) + \\ & (1 - p) \cdot (1 - p) \cdot p \cdot (1 - p) \cdot (1 - p) + \\ & (1 - p) \cdot (1 - p) \cdot (1 - p) \cdot p \cdot (1 - p) + \\ & (1 - p) \cdot (1 - p) \cdot (1 - p) \cdot (1 - p) \cdot p \end{aligned}$$

A expressão acima pode ser simplificada para,

$$5 \cdot p \cdot (1 - p)^4$$

Distribuição Binomial - Questão 2

Em geral, para n minutos e probabilidade p de um gol por minuto, a probabilidade de exatamente um gol é

$$n \cdot p \cdot (1 - p)^{n-1}$$

Aplicando isso ao exemplo da Premier League,

$$95 \cdot 0.028 \cdot (1 - 0.028)^{94} \approx 0.184$$

18,4% das partidas terminam com apenas um gol.

Distribuição Binomial - Questão 3

Para entender isso, primeiro imagine que você recebeu uma pasta com uma correspondência de 95 minutos dividida em segmentos de 1 minuto. Quantas maneiras diferentes haveria de ordenar esses arquivos na pasta?

$$95 \cdot 94 \cdot 93 \cdot 92 \cdot \dots \cdot 3 \cdot 2 \cdot 1$$

maneiras de organizar esses arquivos. Agora imagine que $k=4$ gols ocorreram na partida em quatro momentos diferentes da partida. De quantas maneiras esses 4 gols podem ser organizados dentro da pasta? Isso é

$$4 \cdot 3 \cdot 2 \cdot 1$$

Para um valor geral de k isto é

$$k \cdot (k - 1) \cdot \dots \cdot 2 \cdot 1$$

Distribuição Binomial - Questão 3

E, finalmente, o que dizer dos 86 minutos que não foram gols. Quantas maneiras existem para organizá-los? Bem, isso é

$$(95 - k) \cdot (95 - k - 1) \cdot (95 - k - 2) \cdot (95 - k - 3) \cdots 3 \cdot 2 \cdot 1$$

Para encontrar todas as maneiras pelas quais 4 gols podem ocorrer em uma partida de 95 minutos, dividimos as maneiras de organizar todos os minutos pelas maneiras de organizar os gols e as maneiras de organizar os não-gols. Aquilo é,

$$\frac{95 \cdot 89 \cdot 88 \cdot 87 \cdots 3 \cdot 2 \cdot 1}{4 \cdot 3 \cdot 2 \cdot 1 \cdot 91 \cdot 90 \cdot 89 \cdot 88 \cdots 3 \cdot 2 \cdot 1}$$

A abreviação para escrever isso (para o caso de ***k*** gols) é:

$$\frac{95!}{(95 - k)!k!}$$

Distribuição Binomial - Questão 3

Isso pode ser usado para encontrar a probabilidade de ***k*** gols em uma partida de

$$\frac{95!}{(95 - k)!k!} \cdot (2.7/95)^k \cdot (1 - 2.7/95)^{95-k}$$

Usando os termos ***p*** e ***n***:

$$\frac{n!}{(n - k)!k!} \cdot (p)^k \cdot (1 - p)^{n-k}$$

Esta é a distribuição binomial. Ela dá a probabilidade de haver ***k*** gols assumindo que apenas um gol pode acontecer por minuto e que os gols ocorram em momentos aleatórios durante a partida.

Distribuição de Poisson

- Dividir a partida em blocos de 90 minutos é algo arbitrário.
- Imagine, em vez disso, que o dividimos em n intervalos de tempo discretos, dentro de cada um dos quais um gol pode ocorrer. Agora a probabilidade de k gols nesses slots n é

$$\frac{n!}{(n-k)!k!} \cdot (2.7/n)^k \cdot (1 - 2.7/n)^{n-k}$$

- A distribuição de Poisson é obtida primeiro reorganizando esta equação para obter

$$\frac{n \cdot (n-1) \cdots (k+1)}{k!} \cdot (2.7/n)^k \cdot (1 - 2.7/n)^{n-k}$$

Distribuição de Poisson

$$\frac{n \cdot (n-1) \cdots (k+1)}{k!} \cdot (2.7/n)^k \cdot (1 - 2.7/n)^{n-k}$$

$$\frac{\frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{k+1}{n}}{k!} \cdot (2.7)^k \cdot (1 - 2.7/n)^{n-k}$$

$$\frac{(1 - \frac{1}{n}) \cdot (1 - \frac{2}{n}) \cdots (1 - \frac{k-1}{n})}{k!} \cdot (2.7)^k \cdot (1 - 2.7/n)^n \cdot (1 - 2.7/n)^{-k}$$

Quando tomamos o limite $n \rightarrow \infty$ obtemos

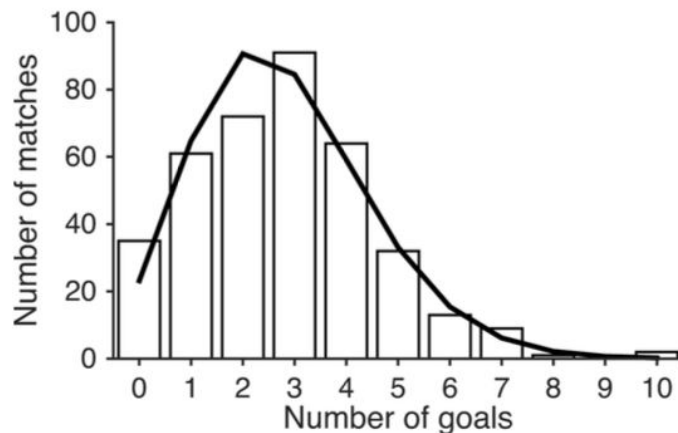
$$\frac{1}{k!} \cdot (2.7)^k \cdot \exp(-2.7) \cdot 1$$

que dá uma probabilidade de k gols em uma partida de

$$\frac{(2.7)^k \exp(-2.7)}{k!}$$

Distribuição de Poisson

Este gráfico mostra a linha comparada com a distribuição de gols da Premier League.



A equação - de apenas um parâmetro (o número médio de gols por partida de 2,7) - captura a **curva de distribuição completa** de gols por partida. Esta é uma observação muito poderosa, porque nos permite atribuir probabilidades aos resultados das partidas, e também (através da regressão de Poisson) avaliar as equipes e como várias ações contribuem para um futebol eficaz.



3 Simulando resultados

Modelos de Regressão Linear

Tem por objetivo modelar o comportamento de uma variável de interesse por meio da combinação de variáveis explicativas.

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \epsilon$$

Coeficiente de determinação é uma medida do ajuste de um modelo estatístico linear:

$$SQ_{\text{tot}} = \sum_{i=1}^n (y_i - \bar{y})^2, \quad SQ_{\text{res}} = \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad SQ_{\text{exp}} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2,$$

$$SQ_{\text{tot}} = SQ_{\text{exp}} + SQ_{\text{res}}.$$

$$R^2 = \frac{SQ_{\text{exp}}}{SQ_{\text{tot}}} = 1 - \frac{SQ_{\text{res}}}{SQ_{\text{tot}}}.$$

Modelos de Regressão Linear

1. **Relacionamento linear:** Deve existir uma relação linear entre as variáveis independentes e dependentes.
2. **Independência residual:** Um resíduo é a diferença entre os dados observados e o valor previsto. Resíduos não devem ter um padrão identificável entre eles.
3. **Normalidade:** Os resíduos devem ser normalmente distribuídos. Os resíduos devem se encaixar ao longo de uma linha diagonal no centro do gráfico Q-Q.
4. **Homocedasticidade:** A homocedasticidade supõe que os resíduos tenham uma variância constante ou desvio padrão da média para cada valor de x .

Modelos Lineares Generalizados

Os MLGs (Modelos Lineares Generalizados) estendem os modelos de regressão simples e múltipla.

Eles possibilitam utilizar outras distribuições para os erros e uma função de ligação relacionando a média da variável resposta à combinação linear das variáveis explicativas.

Três componentes

1. Comportamento (distribuição) da variável resposta
2. Variáveis explicativas
3. Função de ligação entre as variáveis explicativas e variável resposta

Com os modelos lineares generalizados é possível modelar variáveis de interesse que assumem a forma de contagem, contínuas simétricas e assimétricas, binárias e categóricas.

Uma das limitações dos MLGs é a exigência de que os erros sejam independentes.

Modelos Lineares Generalizados

Tipo de regressão	utilização	exemplo de uso
Poisson	usada para modelar dados de contagem.	número de mortes em determinada região ou o número de consumidores que entram em um estabelecimento comercial.
Bernoulli	utilizada na modelagem de fenômenos que podem ser resumidos em uma variável binária, ou seja, se ocorreu ou não um evento.	modelos de concessão de crédito ou em pesquisas clínicas que tem como objetivo verificar os fatores de influência na ocorrência ou não de uma determinada doença.
Gama	usada para modelar dados positivos e assimétricos. A regressão Gama modela variáveis contínuas.	estudo dos fatores que influenciam no valor de um imóvel ou ainda os fatores que influenciam na demanda de produtos em diferentes centros de distribuição.

Simulando resultados

```
# importing the tools required for the Poisson regression model
import statsmodels.api as sm
import statsmodels.formula.api as smf
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn
from scipy.stats import poisson,skellam
```

Simulando resultados

```
epl = pd.read_csv("https://www.football-data.co.uk/mmz4281/2122/E0.csv")
ep = epl[['HomeTeam', 'AwayTeam', 'FTHG', 'FTAG']]
epl = epl.rename(columns={'FTHG': 'HomeGoals', 'FTAG': 'AwayGoals'})
epl.head()

epl = epl[:-10]
epl.mean()
```

```
HomeGoals    1.491892
AwayGoals    1.297297
HTHG         0.686486
HTAG         0.586486
HS           13.764865
...
PCAHA        1.976486
MaxCAHH      2.019946
MaxCAHA      2.054649
AvgCAHH      1.925973
AvgCAHA      1.959514
Length: 98, dtype: float64
```

Regressão

No ajuste, incluímos um parâmetro para vantagem de casa. Equipe e oponente são efeitos fixos.

```
goal_model_data = pd.concat([epi[['HomeTeam','AwayTeam','HomeGoals']].assign(home=1).rename(
    columns={'HomeTeam':'team', 'AwayTeam':'opponent','HomeGoals':'goals'}),
    epi[['AwayTeam','HomeTeam','AwayGoals']].assign(home=0).rename(
    columns={'AwayTeam':'team', 'HomeTeam':'opponent','AwayGoals':'goals'})])

poisson_model = smf.glm(formula="goals ~ home + team + opponent", data=goal_model_data,
    family=sm.families.Poisson()).fit()
poisson_model.summary()
```


Regressão

No ajuste, incluímos um parâmetro para vantagem de casa. Equipe e oponente são efeitos fixos.

Generalized Linear Model Regression Results

Dep. Variable: goals	No. Observations: 740
Model: GLM	Df Residuals: 700
Model Family: Poisson	Df Model: 39
Link Function: Log	Scale: 1.0000
Method: IRLS	Log-Likelihood: -1043.1
Date: Mon, 27 Feb 2023	Deviance: 776.72
Time: 18:22:23	Pearson chi2: 681.
No. Iterations: 5	Pseudo R-squ. (CS): 0.2498
Covariance Type: nonrobust	

Simulando uma partida

Vamos agora simular uma partida entre City e Arsenal

```
home_team='Man City'
away_team='Arsenal'

#Predict for Arsenal vs. Manchester City
home_score_rate=poisson_model.predict(pd.DataFrame(data={'team': home_team, 'opponent': away_team,'home':1},index=[1]))
away_score_rate=poisson_model.predict(pd.DataFrame(data={'team': away_team, 'opponent': home_team,'home':0},index=[1]))
print(home_team + ' against ' + away_team + ' expect to score: ' + str(home_score_rate))
print(away_team + ' against ' + home_team + ' expect to score: ' + str(away_score_rate))

#Lets just get a result
home_goals=np.random.poisson(home_score_rate)
away_goals=np.random.poisson(away_score_rate)
print(home_team + ' : ' + str(home_goals[0]))
print(away_team + ' : ' + str(away_goals[0]))
```

Simulando uma partida

Vamos agora simular uma partida entre City e Arsenal

```
home_team='Man City'
away_team='Arsenal'

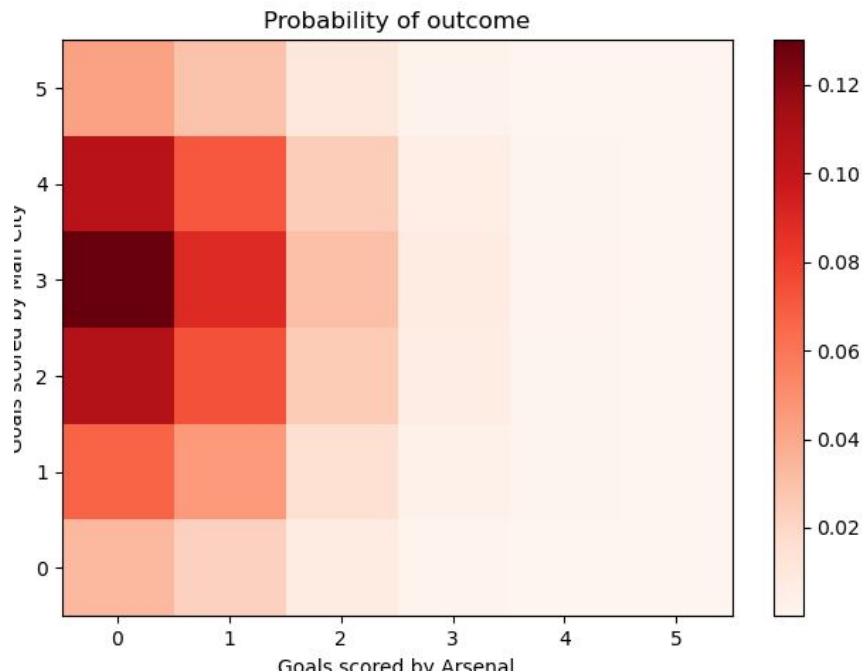
#Predict for Arsenal vs Man City
home_score_rate=poisson.pmf(1,2.481142)
away_score_rate=poisson.pmf(1,0.682894)
print(home_team + ' against Arsenal expect to score: 1 ' + str(home_score_rate))
print(away_team + ' against Man City expect to score: 1 ' + str(away_score_rate))

#Lets just get a result
home_goals=np.random.poisson(home_score_rate)
away_goals=np.random.poisson(away_score_rate)
print(home_team + ' : ' + str(home_goals[0]))
print(away_team + ' : ' + str(away_goals[0]))
```

Man City against Arsenal expect to score: 1 2.481142
dtype: float64
Arsenal against Man City expect to score: 1 0.682894
dtype: float64
Man City: 1
Arsenal: 1

Histograma bidimensional de gols

Isso dá a probabilidade de diferentes linhas de gols. [Código](#)



Times vs Arsenal

- Vamos aproveitar os resultados da regressão de Poisson para:
 - Determinar times com ataques **melhores** que o Arsenal;
 - Determinar times com ataques **piores** que o Arsenal.

Times vs Arsenal

	coef	std err	z	P> z	[0.025	0.975]
-----:-----:-----:-----:-----:-----:-----:						
Intercept	0.2470	0.203	1.214	0.225	-0.152	0.646
team[T.Aston Villa]	-0.1298	0.195	-0.666	0.506	-0.512	0.252
team[T.Brentford]	-0.1589	0.198	-0.802	0.423	-0.547	0.230
team[T.Brighton]	-0.3735	0.209	-1.788	0.074	-0.783	0.036
team[T.Burnley]	-0.5263	0.220	-2.395	0.017	-0.957	-0.096
team[T.Chelsea]	0.2714	0.177	1.529	0.126	-0.077	0.619
team[T.Crystal Palace]	-0.1370	0.196	-0.699	0.484	-0.521	0.247
team[T.Everton]	-0.2847	0.204	-1.394	0.163	-0.685	0.116
team[T.Leeds]	-0.3127	0.207	-1.507	0.132	-0.719	0.094
team[T.Leicester]	0.0481	0.188	0.256	0.798	-0.320	0.416
team[T.Liverpool]	0.4522	0.170	2.657	0.008	0.119	0.786
team[T.Man City]	0.5115	0.168	3.036	0.002	0.181	0.842
team[T.Man United]	0.0156	0.189	0.083	0.934	-0.354	0.385
team[T.Newcastle]	-0.2841	0.204	-1.389	0.165	-0.685	0.117
team[T.Norwich]	-0.8697	0.248	-3.507	0.000	-1.356	-0.384
team[T.Southampton]	-0.2770	0.205	-1.354	0.176	-0.678	0.124
team[T.Tottenham]	0.1294	0.183	0.706	0.480	-0.230	0.489
team[T.Watford]	-0.5181	0.220	-2.357	0.018	-0.949	-0.087
team[T.West Ham]	0.0398	0.187	0.213	0.831	-0.327	0.406
team[T.Wolves]	-0.4429	0.212	-2.088	0.037	-0.859	-0.027

Times vs Arsenal

coef > 0: Time melhor
que Arsenal

	coef	std err	z	P> z	[0.025	0.975]
-----:-----:-----:-----:-----:-----:-----:						
Intercept	0.2470	0.203	1.214	0.225	-0.152	0.646
team[T.Aston Villa]	-0.1298	0.195	-0.666	0.506	-0.512	0.252
team[T.Brentford]	-0.1589	0.198	-0.802	0.423	-0.547	0.230
team[T.Brighton]	-0.3735	0.209	-1.788	0.074	-0.783	0.036
team[T.Burnley]	-0.5263	0.220	-2.395	0.017	-0.957	-0.096
team[T.Chelsea]	0.2714	0.177	1.529	0.126	-0.077	0.619
team[T.Crystal Palace]	-0.1370	0.196	-0.699	0.484	-0.521	0.247
team[T.Everton]	-0.2847	0.204	-1.394	0.163	-0.685	0.116
team[T.Leeds]	-0.3127	0.207	-1.507	0.132	-0.719	0.094
team[T.Leicester]	0.0481	0.188	0.256	0.798	-0.320	0.416
team[T.Liverpool]	0.4522	0.170	2.657	0.008	0.119	0.786
team[T.Man City]	0.5115	0.168	3.036	0.002	0.181	0.842
team[T.Man United]	0.0156	0.189	0.083	0.934	-0.354	0.385
team[T.Newcastle]	-0.2841	0.204	-1.389	0.165	-0.685	0.117
team[T.Norwich]	-0.8697	0.248	-3.507	0.000	-1.356	-0.384
team[T.Southampton]	-0.2770	0.205	-1.354	0.176	-0.678	0.124
team[T.Tottenham]	0.1294	0.183	0.706	0.480	-0.230	0.489
team[T.Watford]	-0.5181	0.220	-2.357	0.018	-0.949	-0.087
team[T.West Ham]	0.0398	0.187	0.213	0.831	-0.327	0.406
team[T.Wolves]	-0.4429	0.212	-2.088	0.037	-0.859	-0.027

Times vs Arsenal

coef < 0: Time pior que Arsenal

	coef	std err	z	P> z	[0.025	0.975]
-----:-----:-----:-----:-----:-----:-----:						
Intercept	0.2470	0.203	1.214	0.225	-0.152	0.646
team[T.Aston Villa]	-0.1298	0.195	-0.666	0.506	-0.512	0.252
team[T.Brentford]	-0.1589	0.198	-0.802	0.423	-0.547	0.230
team[T.Brighton]	-0.3735	0.209	-1.788	0.074	-0.783	0.036
team[T.Burnley]	-0.5263	0.220	-2.395	0.017	-0.957	-0.096
team[T.Chelsea]	0.2714	0.177	1.529	0.126	-0.077	0.619
team[T.Crystal Palace]	-0.1370	0.196	-0.699	0.484	-0.521	0.247
team[T.Everton]	-0.2847	0.204	-1.394	0.163	-0.685	0.116
team[T.Leeds]	-0.3127	0.207	-1.507	0.132	-0.719	0.094
team[T.Leicester]	0.0481	0.188	0.256	0.798	-0.320	0.416
team[T.Liverpool]	0.4522	0.170	2.657	0.008	0.119	0.786
team[T.Man City]	0.5115	0.168	3.036	0.002	0.181	0.842
team[T.Man United]	0.0156	0.189	0.083	0.934	-0.354	0.385
team[T.Newcastle]	-0.2841	0.204	-1.389	0.165	-0.685	0.117
team[T.Norwich]	-0.8697	0.248	-3.507	0.000	-1.356	-0.384
team[T.Southampton]	-0.2770	0.205	-1.354	0.176	-0.678	0.124
team[T.Tottenham]	0.1294	0.183	0.706	0.480	-0.230	0.489
team[T.Watford]	-0.5181	0.220	-2.357	0.018	-0.949	-0.087
team[T.West Ham]	0.0398	0.187	0.213	0.831	-0.327	0.406
team[T.Wolves]	-0.4429	0.212	-2.088	0.037	-0.859	-0.027

Times vs Arsenal

p-value

Estatisticamente
significante se
 $p\text{-value} < 0.05$

	coef	std err	z	P> z	[0.025	0.975]
-----:-----:-----:-----:-----:-----:-----:						
Intercept	0.2470	0.203	1.214	0.225	-0.152	0.646
team[T.Aston Villa]	-0.1298	0.195	-0.666	0.506	-0.512	0.252
team[T.Brentford]	-0.1589	0.198	-0.802	0.423	-0.547	0.230
team[T.Brighton]	-0.3735	0.209	-1.788	0.074	-0.783	0.036
team[T.Burnley]	-0.5263	0.220	-2.395	0.017	-0.957	-0.096
team[T.Chelsea]	0.2714	0.177	1.529	0.126	-0.077	0.619
team[T.Crystal Palace]	-0.1370	0.196	-0.699	0.484	-0.521	0.247
team[T.Everton]	-0.2847	0.204	-1.394	0.163	-0.685	0.116
team[T.Leeds]	-0.3127	0.207	-1.507	0.132	-0.719	0.094
team[T.Leicester]	0.0481	0.188	0.256	0.798	-0.320	0.416
team[T.Liverpool]	0.4522	0.170	2.657	0.008	0.119	0.786
team[T.Man City]	0.5115	0.168	3.036	0.002	0.181	0.842
team[T.Man United]	0.0156	0.189	0.083	0.934	-0.354	0.385
team[T.Newcastle]	-0.2841	0.204	-1.389	0.165	-0.685	0.117
team[T.Norwich]	-0.8697	0.248	-3.507	0.000	-1.356	-0.384
team[T.Southampton]	-0.2770	0.205	-1.354	0.176	-0.678	0.124
team[T.Tottenham]	0.1294	0.183	0.706	0.480	-0.230	0.489
team[T.Watford]	-0.5181	0.220	-2.357	0.018	-0.949	-0.087
team[T.West Ham]	0.0398	0.187	0.213	0.831	-0.327	0.406
team[T.Wolves]	-0.4429	0.212	-2.088	0.037	-0.859	-0.027

Times vs Arsenal

	coef	std err	z	P> z	[0.025	0.975]	
-----:-----:-----:-----:-----:-----:-----:-----:							
Intercept	0.2470	0.203	1.214	0.225	-0.152	0.646	
team[T.Aston Villa]	-0.1298	0.195	-0.666	0.506	-0.512	0.252	
team[T.Brentford]	-0.1589	0.198	-0.802	0.423	-0.547	0.230	
team[T.Brighton]	-0.3735	0.209	-1.788	0.074	-0.783	0.036	
team[T.Burnley]	-0.5263	0.220	-2.395	0.017	-0.957	-0.096	
team[T.Chelsea]	0.2714	0.177	1.529	0.126	0.077	0.619	não-significante
team[T.Crystal Palace]	-0.1370	0.196	-0.699	0.484	-0.521	0.247	
team[T.Everton]	-0.2847	0.204	-1.394	0.163	-0.685	0.116	
team[T.Leeds]	-0.3127	0.207	-1.507	0.132	-0.719	0.094	
team[T.Leicester]	0.0481	0.188	0.256	0.798	-0.320	0.416	
team[T.Liverpool]	0.4522	0.170	2.657	0.008	0.119	0.786	significante
team[T.Man City]	0.5115	0.168	3.036	0.002	0.181	0.842	
team[T.Man United]	0.0156	0.189	0.083	0.934	-0.354	0.385	
team[T.Newcastle]	-0.2841	0.204	-1.389	0.165	-0.685	0.117	
team[T.Norwich]	-0.8697	0.248	-3.507	0.000	-1.356	-0.384	
team[T.Southampton]	-0.2770	0.205	-1.354	0.176	-0.678	0.124	
team[T.Tottenham]	0.1294	0.183	0.706	0.480	-0.230	0.489	
team[T.Watford]	-0.5181	0.220	-2.357	0.018	-0.949	-0.087	
team[T.West Ham]	0.0398	0.187	0.213	0.831	-0.327	0.406	
team[T.Wolves]	-0.4429	0.212	-2.088	0.037	-0.859	-0.027	

**Diferenças
estatisticamente
significativas são
extremamente difíceis de
encontrar no futebol!**

É por isso que nos concentramos nas
métricas de estilo de jogo...



4 Previsão de jogos

Sorte e aleatoriedade

- Usando a taxa de gols de cada clube durante a temporada PL 12/13, simulamos o que poderia acontecer durante a temporada 13/14.
- Simular a liga usando a hipótese de Poisson mostra que os resultados podem facilmente ser muito diferentes devido à natureza aleatória da forma como os gols acontecem.

Sorte e aleatoriedade

- Usando a taxa de gols de cada clube durante a temporada PL 12/13, simulamos o que poderia acontecer durante a temporada 13/14.
- Simular a liga usando a hipótese de Poisson mostra que os resultados podem facilmente ser muito diferentes devido à natureza aleatória da forma como os gols acontecem.

Simulação 1

	<i>Team</i>	<i>P</i>	<i>W</i>	<i>D</i>	<i>L</i>	<i>F</i>	<i>A</i>	Pts
1	Manchester City	38	22	7	9	71	42	73
2	Liverpool	38	22	5	11	64	43	71
3	Chelsea	38	21	5	12	74	51	68
7	Manchester United	38	19	7	12	61	45	64

Sorte e aleatoriedade

- Usando a taxa de gols de cada clube durante a temporada PL 12/13, simulamos o que poderia acontecer durante a temporada 13/14.
- Simular a liga usando a hipótese de Poisson mostra que os resultados podem facilmente ser muito diferentes devido à natureza aleatória da forma como os gols acontecem.

Simulação 2

	<i>Team</i>	<i>P</i>	<i>W</i>	<i>D</i>	<i>L</i>	<i>F</i>	<i>A</i>	Pts
2	Liverpool	38	23	7	8	68	37	76
3	Chelsea	38	22	8	8	75	52	74
7	Manchester United	38	22	5	11	72	43	71
1	Manchester City	38	19	8	11	64	42	65

Sorte e aleatoriedade

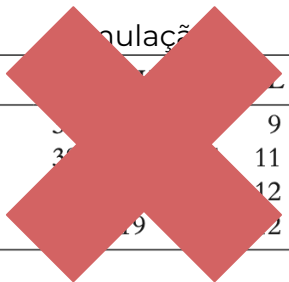
- Usando a taxa de gols de cada clube durante a temporada PL 12/13, simulamos o que poderia acontecer durante a temporada 13/14.
- Simular a liga usando a hipótese de Poisson mostra que os resultados podem facilmente ser muito diferentes devido à natureza aleatória da forma como os gols acontecem.

Simulação 1							
<i>Team</i>	<i>P</i>	<i>W</i>	<i>D</i>	<i>L</i>	<i>F</i>	<i>A</i>	Pts
Manchester City	38	22	7	9	71	42	73
Liverpool	38	22	5	11	64	43	71
Chelsea	38	21	5	12	74	51	68
Manchester United	38	19	7	12	61	45	64

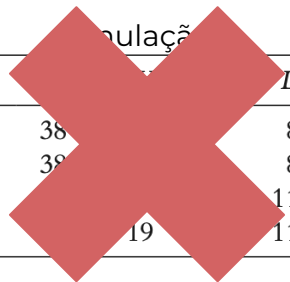
Simulação 2							
<i>Team</i>	<i>P</i>	<i>W</i>	<i>D</i>	<i>L</i>	<i>F</i>	<i>A</i>	Pts
Liverpool	38	23	7	8	68	37	76
Chelsea	38	22	8	8	75	52	74
Manchester United	38	22	5	11	72	43	71
Manchester City	38	19	8	11	64	42	65

Sorte e aleatoriedade

Por mais interessante que cada uma dessas duas realidades alternativas possa ser, **individualmente elas não são importantes.**



População					
Team		L	F	A	Pts
Manchester City	38	9	71	42	73
Liverpool	38	11	64	43	71
Chelsea	38	12	74	51	68
Manchester United	19	2	61	45	64



População					
Team		L	F	A	Pts
Liverpool	38	8	68	37	76
Chelsea	38	8	75	52	74
Manchester United	11	11	72	43	71
Manchester City	19	11	64	42	65

Sorte e aleatoriedade

- O importante é resumir o que acontece em **10.000 simulações**.
- Com que *frequência* times diferentes ganharam a liga?
 1. O Manchester United, campeão da PL 12/13, conquistou 26,2% das simulações
 2. Chelsea venceu 19,2%
 3. Arsenal 17,6%
 4. Manchester City 12,8%
 5. Liverpool 11,5%
 6. Tottenham Hotspur 6,0%.
- Os resultados mostram quanto do resultado da liga é determinado pela aleatoriedade.

Sorte e aleatoriedade

- Os modelos baseados em Poisson não são perfeitos.
- Eles tendem a **subestimar as melhores** equipes e **superestimar as piores** equipes.
- Durante a temporada, podemos querer usar um modelo baseado em xG em vez de um modelo baseado em gols para prever o resultado da temporada como um todo.

Limites de previsão

Podemos delinear dois tipos diferentes de previsões:

- **Baseados em modelos, como o modelo do Fivethirtyeight**
- Especialistas, que estão dizendo o que vai acontecer

FiveThirtyEight











English | Español | Português  



PREMIER LEAGUE
2022-23

England

Updated March 19, 2023, at 11:54 a.m.

TEAM	TEAM RATING		AVG. SIMULATED SEASON		END-OF-SEASON PROBABILITIES			
	SPI	OFF. DEF.	GOAL DIFF.	PTS.	EVERY POSITION	RELEGATED	QUALIFY FOR UCL	WIN PREMIER LEAGUE
 Arsenal 89 pts	85.5	2.5 0.5	+47	87		—	>99%	56%
 Man. City 61 pts	92.3	2.9 0.3	+57	85		—	>99%	44%
 Man. United 50 pts	79.6	2.3 0.7	+13	71		—	74%	<1%
 Newcastle 47 pts	80.3	2.1 0.5	+24	66		<1%	44%	<1%
 Liverpool 42 pts	84.4	2.6 0.6	+27	64		<1%	29%	<1%
 Tottenham 49 pts	75.8	2.2 0.8	+14	64		—	25%	<1%

Limites de previsão

Podemos delinear dois tipos diferentes de previsões:

- Baseados em modelos, como o modelo do Fivethirtyeight
- **Especialistas, que estão dizendo o que vai acontecer**



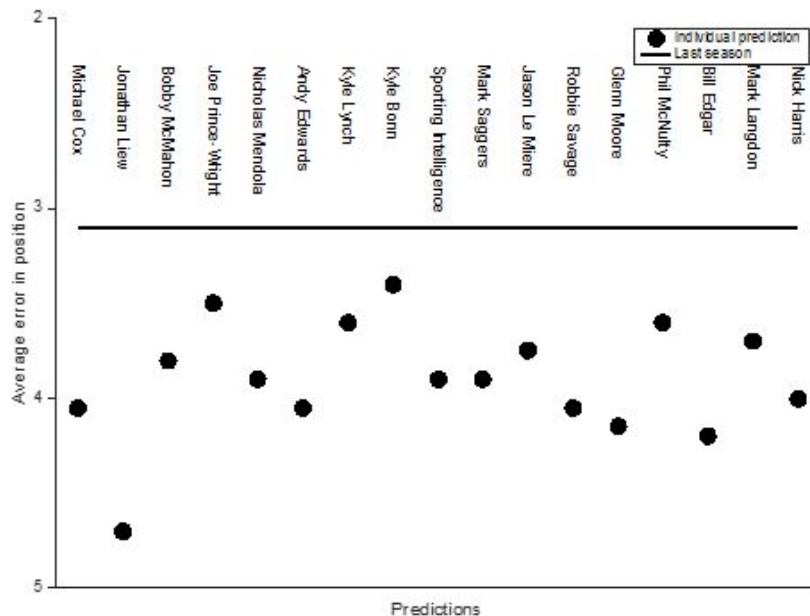
Limites de previsão

Podemos comparar esses dois tipos de previsão com um modelo simples:

- **Modelo simples:** copiar o que ocorreu no ano anterior, ou seja, repetir a colocação dos times na temporada passada.

Limites de previsão

Especialistas não superam a média: eles não tendem a ter um desempenho superior se você apenas copiar o que aconteceu no ano anterior.



Limites de previsão

Mesmo modelos mais elaborados, como o modelo ELO, ainda não superam

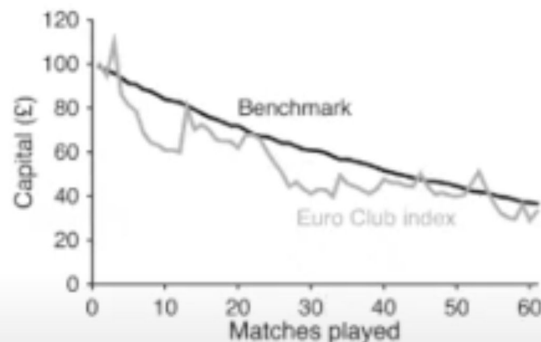


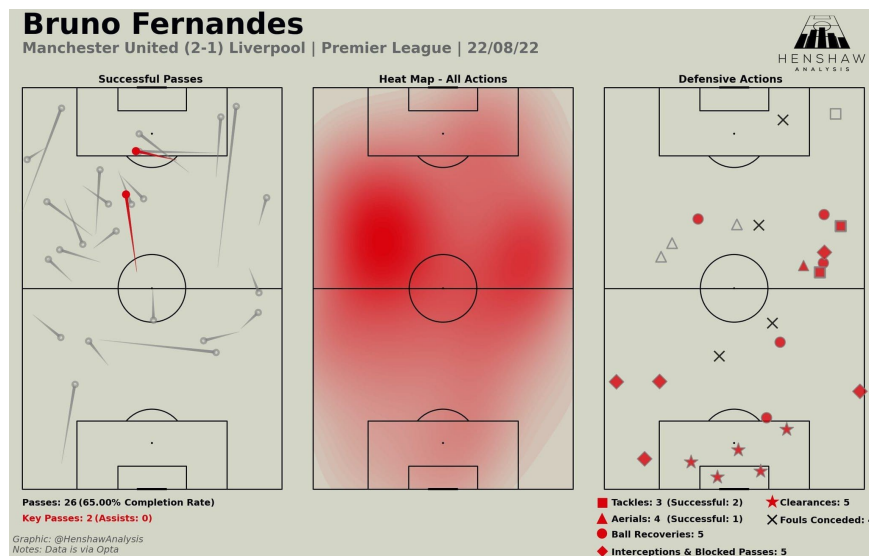
Figure 13.4 How the Euro Club strategy (grey) performed compared to a benchmark of random bets (black) with £100 initial capital over the first 60 matches of the 2015/16 Premier League.



5 Modelagem estatística

Modelagem estatística

Nesta seção, examinamos uma variedade de métodos para avaliar o grau em que o passe/posse leva a mais gols marcados e como podemos identificar as áreas mais fortes do campo para uma equipe.

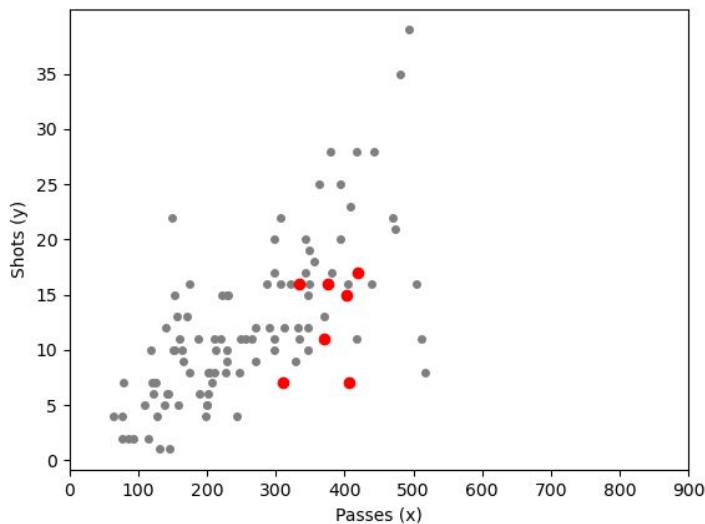


Copa do Mundo Feminina de 2019

- Analisaremos os jogos da Copa do Mundo Feminina de 2019
- Para nossa tarefa, precisamos de 2 dataframes diferentes.
 - `passshot_df`: Neste dataframe, gostaríamos de manter informações sobre o desempenho do time em todos os jogos que eles jogaram - índice de um jogo, nome de um time, número de chutes, número de gols e número de passes de perigo por esse time.
 - `hazard_passes_df`: dataframe de todos os passes de perigo durante o torneio.
- O código para a análise está [aqui](#).

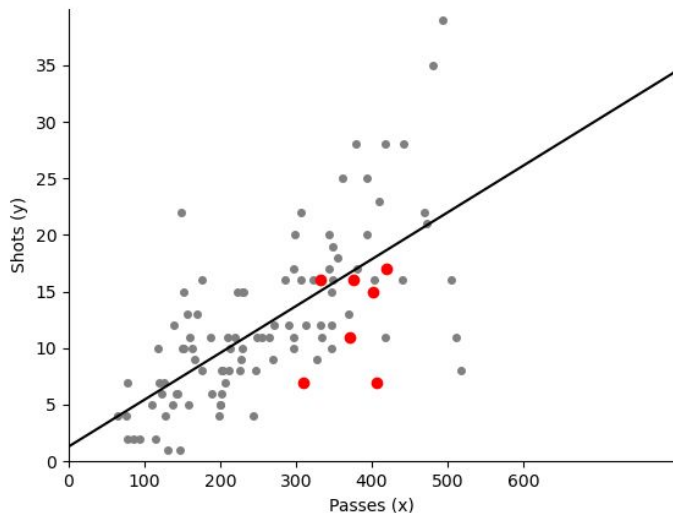
Copa do Mundo Feminina de 2019

- Gostaríamos de investigar se existe alguma relação entre o número de passes e o número de finalizações de uma equipe em um jogo.
- O gráfico de dispersão abaixo mostra isso. Em vermelho, temos o desempenho da seleção feminina da Inglaterra nessas 2 áreas.



Copa do Mundo Feminina de 2019

- Queremos investigar a relação linear entre o número de passes e o número de chutes.
- Para isso, ajustamos a regressão linear usando a biblioteca statsmodels.

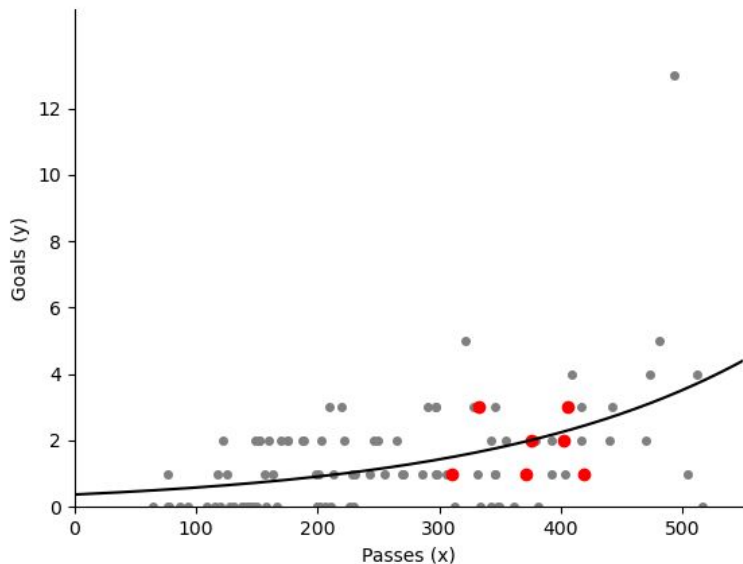


OLS Regression Results						
=====						
Dep. Variable:	Shots	R-squared:	0.466			
Model:	OLS	Adj. R-squared:	0.461			
Method:	Least Squares	F-statistic:	88.98			
Date:	Mon, 27 Feb 2023	Prob (F-statistic):	1.46e-15			
Time:	18:10:51	Log-Likelihood:	-318.20			
No. Observations:	104	AIC:	640.4			
Df Residuals:	102	BIC:	645.7			
Df Model:	1					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

Intercept	1.3098	1.268	1.033	0.304	-1.206	3.825
Passes	0.0414	0.004	9.433	0.000	0.033	0.050
=====						
Omnibus:	10.628	Durbin-Watson:	2.552			
Prob(Omnibus):	0.005	Jarque-Bera (JB):	13.317			
Skew:	0.543	Prob(JB):	0.00128			
Kurtosis:	4.376	Cond. No.	718.			
=====						
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.						

Copa do Mundo Feminina de 2019

- Queremos investigar a relação entre número de passes e número de gols.
- Para isso, ajustamos a regressão de Poisson usando a biblioteca statsmodels.
- É melhor usar a regressão de Poisson, pois os gols são pouco frequentes.



Generalized Linear Model Regression Results						
=====						
Dep. Variable:	Goals	No. Observations:	104			
Model:	GLM	Df Residuals:	102			
Model Family:	Poisson	Df Model:	1			
Link Function:	Log	Scale:	1.0000			
Method:	IRLS	Log-Likelihood:	-153.85			
Date:	Fri, 24 Mar 2023	Deviance:	130.89			
Time:	14:03:36	Pearson chi2:	121.			
No. Iterations:	5	Pseudo R-squ. (CS):	0.3275			
Covariance Type:		nonrobust				
=====						
	coef	std err	z	P> z	[0.025	0.975]

Intercept	-0.9924	0.246	-4.034	0.000	-1.475	-0.510
Passes	0.0045	0.001	6.358	0.000	0.003	0.006
=====						

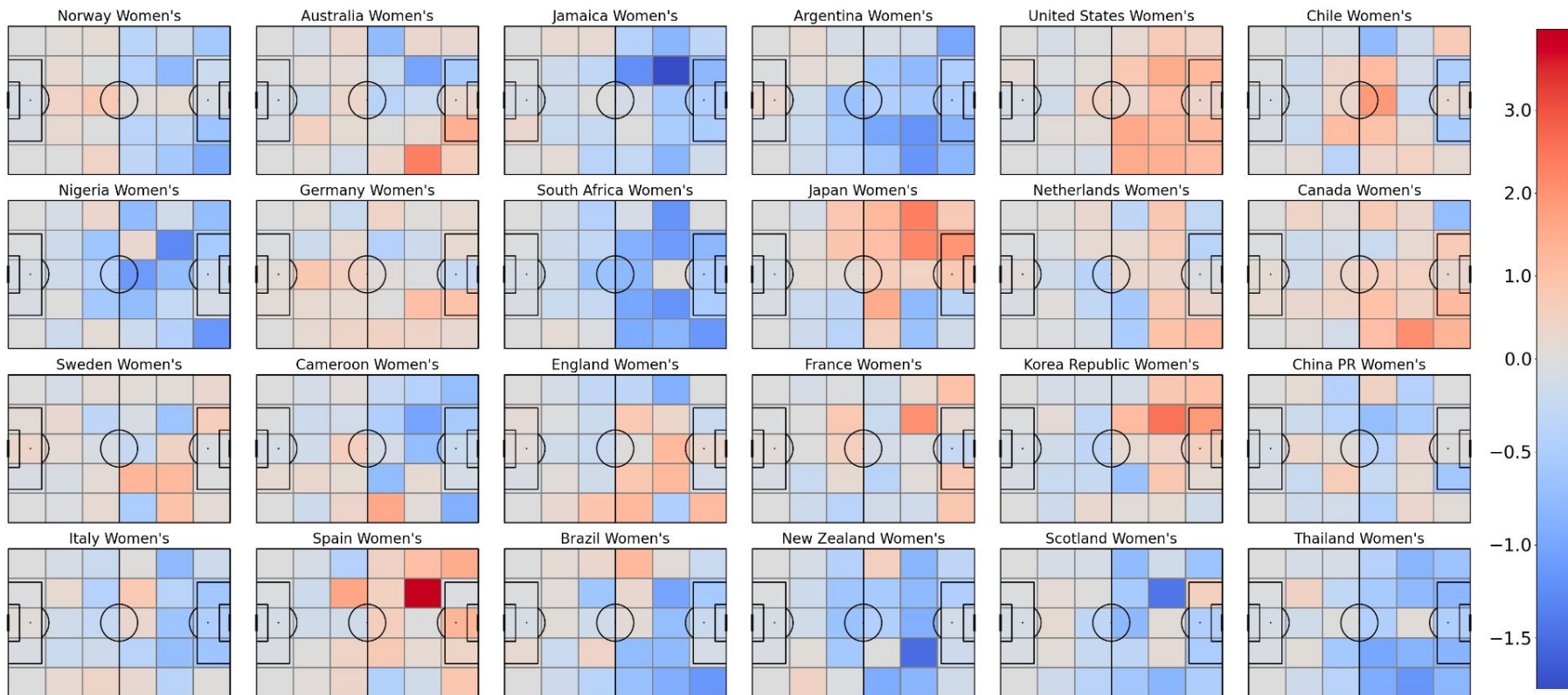
Copa do Mundo Feminina de 2019

Gostaríamos de saber qual time teve melhor e pior desempenho em relação ao número de passes de perigo em diferentes zonas.

Para cada time que jogou no WWC 2019:

1. Calculamos o número de passes em cada zona e normalizamos pelo número de jogos desse time.
2. Calculamos o número médio de passes perigosos por zona ao longo do torneio: μ
3. Subtraímos μ do número de passes de perigo em cada zona.
4. Traçamos o mapa de calor para cada equipe

Danger passes per game - performance above zone average





6 KPIs da equipe

KPIs - Indicadores de performance

- KPIs: usados por comissão técnica para análise
- Complementar a análise de vídeos
- Diferentes tipos de indicadores
- Quantitativos (mais simples de avaliar com dados e visualizações) x Qualitativos

KPIs - Indicadores de performance

Indicadores relacionados a um atleta

- Quantidade de desarmes
- Passes completos
- Passes incompletos
- Razão de passes completos / Incompletos
- Distância de passes (curto, médio, longo)
- Passes para determinada zona (p. ex., área)
- Jogador com quem mais interage

KPIs - Indicadores de performance

Indicadores relacionados a um atleta

- Finalizações no alvo
 - Razão finalização / gol
 - Razão finalização / alvo
 - Razão passe / finalização
 - Tipo de finalização (p. ex., cabeçada, chute)
 - Desarmes
 - Faltas
-
- Outros indicadores relacionados a um campeonato ou temporada.

KPIs - Indicadores de performance

Indicadores relacionados a uma equipe

- Finalizações no alvo
- contra-ataques realizados e concedidos
- recuperação + finalização com menos de X passes
- tempo médio e número de passes de cada ataque
- % posse de bola (total, campo de defesa/ataque)
- n e % de passes curtos, médios e longos por ataque
- finalizações concedidas (zona da perda da bola)

KPIs - Indicadores de performance

Indicadores relacionados a uma equipe

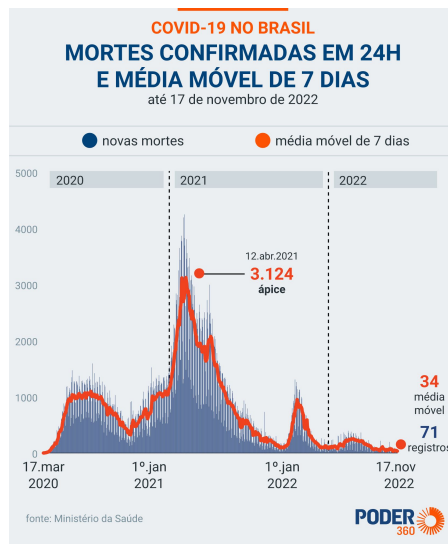
- chutão (ações de chutar a bola para fora ou para o campo adversário sem intenção de passe ou lançamento)
- origem e destino dos passes dos adversários em zona 1 (defensiva)
- Recuperação da bola (número e zonas)
- Motifs (sequências de passes que levam a chances de gol ou criam vantagens para avançar em uma zona do campo)



7 Média móvel de pontos

Média Móvel

- Em estatística, a média móvel é um recurso utilizado para se identificar a tendência de um conjunto de dados dispostos em uma série de tempo.
- Esse conceito ficou bem popular durante a pandemia.



Média Móvel do City de Guardiola

- Para nossa tarefa, vamos usar os dados do Football-Data com os resultados das partidas da Premier League inglesa desde que Pep Guardiola começou a treinar o Manchester City.
- Os códigos podem ser encontrados [aqui](#).

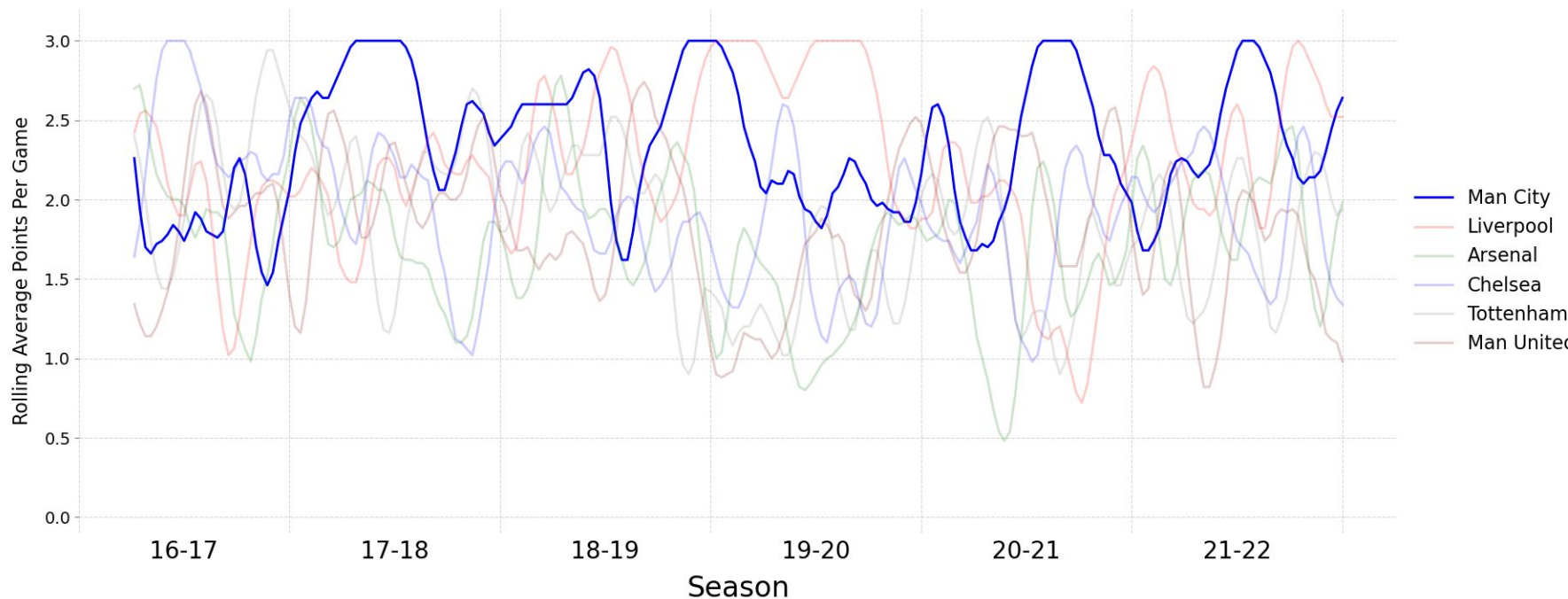


Média Móvel do City de Guardiola

- Queremos investigar o desempenho do City de Pep Guardiola em relação ao desempenho de outros clubes TOP 6 neste período.
- Para cada uma dessas equipes, pegamos os jogos disputados por eles e atribuímos o número de pontos que eles marcaram.
- Em seguida, calculamos a média móvel de 10 jogos.

Média Móvel do City de Guardiola

Man City since Guardiola's arrival - 10 game rolling average points comparing to TOP 6 clubs





8 Estudo de caso

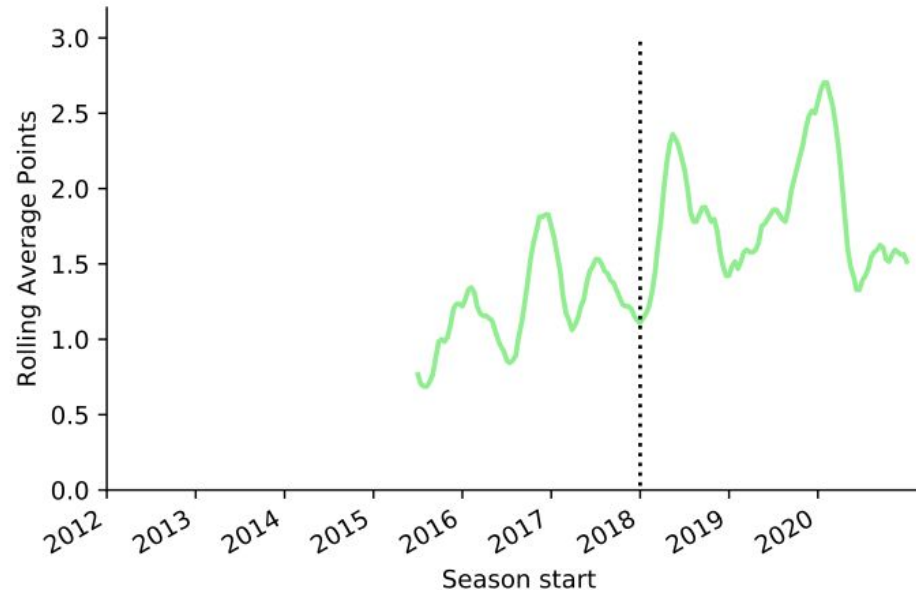
Tendência de longo prazo em Hammarby

Estudo de caso - Hammarby

- É importante que os KPIs reflitam esse trabalho em diferentes escalas de tempo.
- Para a diretoria de um clube, eles querem ver termos mais longos: o clube está indo na direção certa?
- Exemplo de relatório de análise (2021) - Hammarby

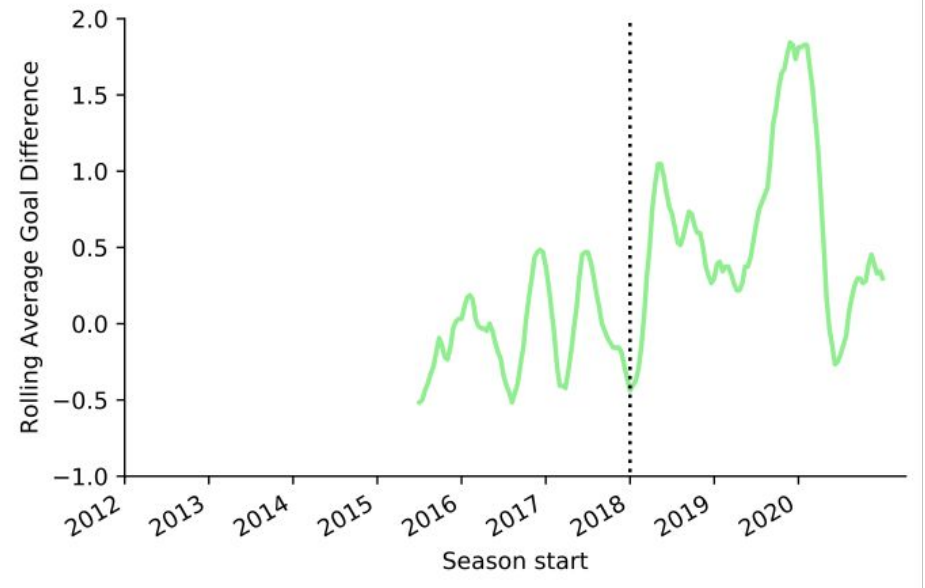
Estudo de caso - Hammarby

- O gráfico abaixo mostra uma média móvel de pontos do Hammarby antes e depois de Stefan Billborn assumir o cargo de técnico.



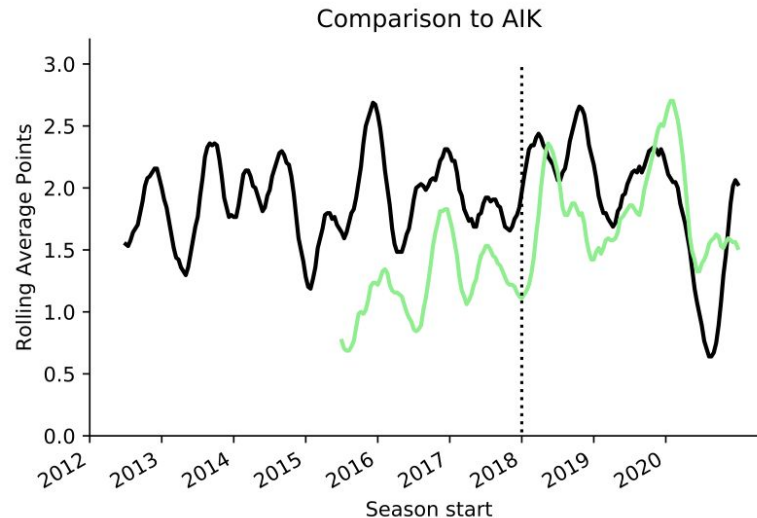
Estudo de caso - Hammarby

- Há uma clara e sustentada melhoria de pontos por jogo. Mesmo o nível inferior da equipe (ou seja, início da temporada de 2021 e final da temporada de 2018) é melhor do que agora do que em 2017. O mesmo é verdade se olharmos para o saldo de gols.



Estudo de caso - Hammarby

- Comparação com melhores equipes (comparação com AIK)
- A ambição do Hammarby é ser um time regular entre os três primeiros. Ao ter essa ambição, temos que estar cientes dos altos e baixos típicos de equipes que já estiveram nessa posição anteriormente.

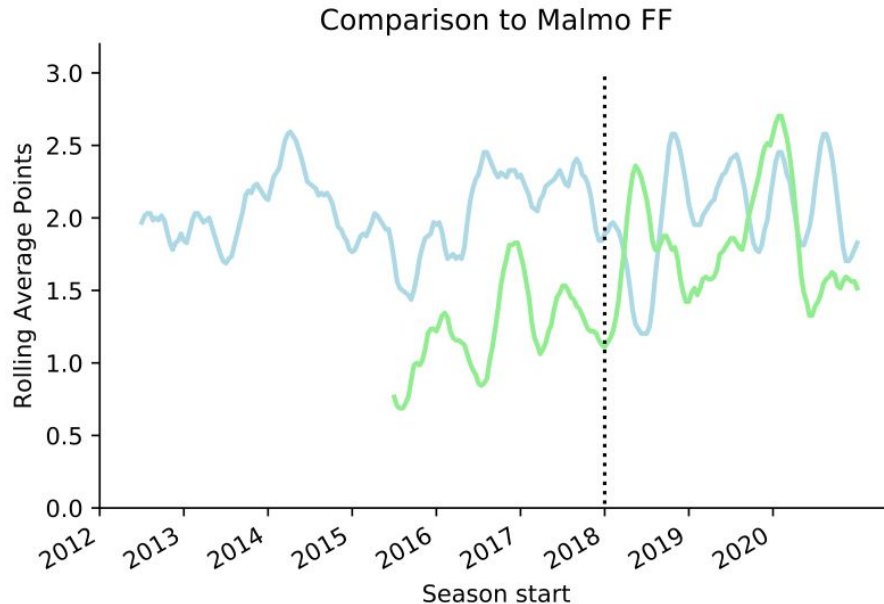


Estudo de caso - Hammarby

- A média móvel de pontos atual (Hammarby) é semelhante a um nível de AIK ao longo da última década.
- Os altos e baixos que Hammarby teve são bastante semelhantes aos experimentados pelo AIK nos últimos 10 anos.
- De 2011-2020, eles venceram o Allsvenskan uma vez, o segundo lugar quatro vezes e só não conseguiram terminar entre os quatro primeiros uma vez (em 2020).
- Que é, por definição, uma 'equipe regular de top 3'.

Estudo de caso - Hammarby

- Comparação com melhor time - Malmö (flutuações menores)
- Malmö teve um nível consistentemente alto por uma década.



Referências

- [Lesson 5 — Soccermatics](#)
- D. Sumpter, Soccermatics: Mathematical Adventures in the Beautiful Game. Bloomsbury Publishing Plc, 2016.

Obrigado!



luiza.chagas@dcc.ufmg.br



hugoriosneto@dcc.ufmg.br



adrianoc@dcc.ufmg.br



meira@dcc.ufmg.br