An Analysis of Soccer Wins Based on Two Variables in the 2008-2016 seasons

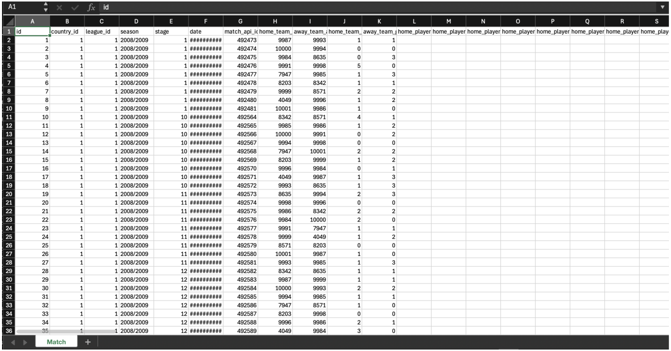
Thomas Xiao

March 21, 2025

Data Set Modeled with CDPF

The data set used in this analysis is the European Soccer Database by Hugo Mathien, which is an extensive resource for soccer data analysis and machine learning. It includes information on more than 25,000 matches and over 10,000 players from 11 European countries, covering the seasons from 2008 to 2016. More specifically, in each match it includes detailed events such as goals, possession, corners, crosses, fouls, and red and yellow card frequencies. However, this analysis will only focus on two match details, which are the number of goals scored and the number of shots aimed towards the goal of every team, by using the CDPF, or the Cobb-Douglas production function, wich is an economic model that estimates US total production. This analysis will use the CDPF and extend it into soccer by using the data set. Unfortunately, due to the size of the csv file, I am unable to directly put a table into this editing software. However, I have the download to the file here, where you can open it on Excel locally by selecting "Download" on the file preview page.

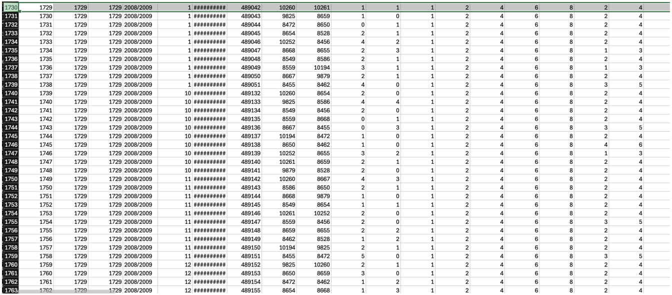
Data Set cont.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	id	country_id	league_id	season	stage	date	match_api_ic	home_team_	away_team_a	home_team_	away_team_g	home_player	home_player	home_player	home_player	home_player	home_player	home_player	home_playe
2	1	1	1	2008/2009	1	##########	492473	9987	9993	1	1								
3	2	1	1	2008/2009	1	##########	492474	10000	9994	0	0								
4	3	1	1	2008/2009	1	##########	492475	9984	8635	0	3								
5	4	1	1	2008/2009	1	##########	492476	9991	9998	5	0								
6	5	1	1	2008/2009	1	##########	492477	7947	9985	1	3								
7	6	1	1	2008/2009	1	##########	492478	8203	8342	1	1								
8	7	1	1	2008/2009	1	##########	492479	9999	8571	2	2								
9	8	1	1	2008/2009	1	##########	492480	4049	9996	1	2								
10	9	1	1	2008/2009	1	##########	492481	10001	9986	1	0								
11	10	1	1	2008/2009	10	##########	492564	8342	8571	4	1								
12	11	1	1	2008/2009	10	##########	492565	9985	9986	1	2								
13	12	1	1	2008/2009	10	##########	492566	10000	9991	0	2								
14	13	1	1	2008/2009	10	##########	492567	9994	9998	0	0								
15	14	1	1	2008/2009	10	##########	492568	7947	10001	2	2								
16	15	1	1	2008/2009	10	##########	492569	8203	9999	1	2								
17	16	1	1	2008/2009	10	##########	492570	9996	9984	0	1								
18	17	1	1	2008/2009	10	##########	492571	4049	9987	1	3								
19	18	1	1	2008/2009	10	##########	492572	9993	8635	1	3								
20	19	1	1	2008/2009	11	##########	492573	8635	9994	2	3								
21	20	1	1	2008/2009	11	##########	492574	9998	9996	0	0								
22	21	1	1	2008/2009	11	##########	492575	9986	8342	2	2								
23	22	1	1	2008/2009	11	##########	492576	9984	10000	2	0								
24	23	1	1	2008/2009	11	##########	492577	9991	7947	1	1								
25	24	1	1	2008/2009	11	##########	492578	9999	4049	1	2								
26	25	1	1	2008/2009	11	##########	492579	8571	8203	0	0								
27	26	1	1	2008/2009	11	##########	492580	10001	9987	1	0								
28	27	1	1	2008/2009	11	##########	492581	9993	9985	1	3								
29	28	1	1	2008/2009	12	##########	492582	8342	8635	1	1								
30	29	1	1	2008/2009	12	##########	492583	9987	9999	1	1								
31	30	1	1	2008/2009	12	##########	492584	10000	9993	2	2								
32	31	1	1	2008/2009	12	##########	492585	9994	9985	1	1								
33	32	1	1	2008/2009	12	##########	492586	7947	8571	1	0								
34	33	1	1	2008/2009	12	##########	492587	8203	9998	0	0								
35	34	1	1	2008/2009	12	##########	492588	9996	9986	2	1								
36	35	1	1	2008/2009	12	##########	492589	4049	9984	3	0								

Figure 1: The first few rows of the data set opened in Excel

Data Set cont.



1730	1729	1729	1729	2008/2009	1	##########	489042	10260	10261	1	1	1	2	4	6	8	2	4
1731	1730	1729	1729	2008/2009	1	##########	489043	9825	8659	1	0	1	2	4	6	8	2	4
1732	1731	1729	1729	2008/2009	1	##########	489044	8472	8650	0	1	1	2	4	6	8	2	4
1733	1732	1729	1729	2008/2009	1	##########	489045	8654	8528	2	1	1	2	4	6	8	2	4
1734	1733	1729	1729	2008/2009	1	##########	489046	10252	8456	4	2	1	2	4	6	8	2	4
1735	1734	1729	1729	2008/2009	1	##########	489047	8668	8655	2	3	1	2	4	6	8	1	3
1736	1735	1729	1729	2008/2009	1	##########	489048	8549	8586	2	1	1	2	4	6	8	2	4
1737	1736	1729	1729	2008/2009	1	##########	489049	8559	10194	3	1	1	2	4	6	8	1	3
1738	1737	1729	1729	2008/2009	1	##########	489050	8667	9879	2	1	1	2	4	6	8	2	4
1739	1738	1729	1729	2008/2009	1	##########	489051	8455	8462	4	0	1	2	4	6	8	3	5
1740	1739	1729	1729	2008/2009	10	##########	489132	10260	8654	2	0	1	2	4	6	8	2	4
1741	1740	1729	1729	2008/2009	10	##########	489133	9825	8586	4	4	1	2	4	6	8	2	4
1742	1741	1729	1729	2008/2009	10	##########	489134	8549	8456	2	0	1	2	4	6	8	2	4
1743	1742	1729	1729	2008/2009	10	##########	489135	8559	8668	0	1	1	2	4	6	8	2	4
1744	1743	1729	1729	2008/2009	10	##########	489136	8667	8455	0	3	1	2	4	6	8	3	5
1745	1744	1729	1729	2008/2009	10	##########	489137	10194	8472	1	0	1	2	4	6	8	2	4
1746	1745	1729	1729	2008/2009	10	##########	489138	8650	8462	1	0	1	2	4	6	8	4	6
1747	1746	1729	1729	2008/2009	10	##########	489139	10252	8655	3	2	1	2	4	6	8	1	3
1748	1747	1729	1729	2008/2009	10	##########	489140	10261	8659	2	1	1	2	4	6	8	2	4
1749	1748	1729	1729	2008/2009	10	##########	489141	9879	8528	2	0	1	2	4	6	8	2	4
1750	1749	1729	1729	2008/2009	11	##########	489142	10260	8667	4	3	1	2	4	6	8	2	4
1751	1750	1729	1729	2008/2009	11	##########	489143	8586	8650	2	1	1	2	4	6	8	2	4
1752	1751	1729	1729	2008/2009	11	##########	489144	8668	9879	1	0	1	2	4	6	8	2	4
1753	1752	1729	1729	2008/2009	11	##########	489145	8549	8654	1	1	1	2	4	6	8	2	4
1754	1753	1729	1729	2008/2009	11	##########	489146	10261	10252	2	0	1	2	4	6	8	2	4
1755	1754	1729	1729	2008/2009	11	##########	489147	8559	8456	2	0	1	2	4	6	8	3	5
1756	1755	1729	1729	2008/2009	11	##########	489148	8659	8655	2	2	1	2	4	6	8	2	4
1757	1756	1729	1729	2008/2009	11	##########	489149	8462	8528	1	2	1	2	4	6	8	2	4
1758	1757	1729	1729	2008/2009	11	##########	489150	10194	9825	2	1	1	2	4	6	8	2	4
1759	1758	1729	1729	2008/2009	11	##########	489151	8455	8472	5	0	1	2	4	6	8	3	5
1760	1759	1729	1729	2008/2009	12	##########	489152	9825	10260	2	1	1	2	4	6	8	2	4
1761	1760	1729	1729	2008/2009	12	##########	489153	8650	8659	3	0	1	2	4	6	8	2	4
1762	1761	1729	1729	2008/2009	12	##########	489154	8472	8462	1	2	1	2	4	6	8	2	4
1763	1762	1729	1729	2008/2009	12	##########	489155	8654	8668	1	3	1	2	4	6	8	2	4

Figure 2: Data set row 1730, as that is where the script starts observing based on lack of data from rows 1-1729

Writing the CDPF in another form: Exercise 81a

The logarithmic transformation of the CDPF is done to linearize the multiplicative relationship into an additive one, which simplifies analysis:

- Start with $P = bL^{\alpha}K^{1-\alpha}$.
- Take the natural logarithm of both sides: $\ln P = \ln(bL^{\alpha}K^{1-\alpha})$.
- Use properties of logarithms: $\ln(ab) = \ln a + \ln b$.
- Split the logarithms: $\ln P = \ln b + \ln L^{\alpha} + \ln K^{1-\alpha}$.
- Apply the power rule for logarithms: $\ln a^b = b \ln a$.
- Combine and rearrange terms: $\ln \frac{P}{K} = \ln b + \alpha \ln \frac{L}{K}$.

This transformation allows us to use linear regression techniques to estimate α and $\ln b$.

Exercise 81a cont.

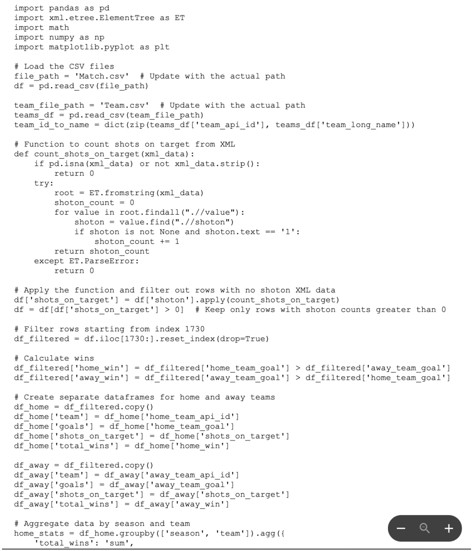
By taking logarithms, the CDPF function $P = bL^{\alpha}K^{1-\alpha}$ can be expressed as $\ln \frac{P}{K} = \ln b + \alpha \ln \frac{L}{K}$

$$\begin{aligned}
P &= bL^{\alpha}K^{1-\alpha} \\
\ln P &= \ln(bL^{\alpha}K^{1-\alpha}) \\
\ln P &= \ln b + \ln L^{\alpha} + \ln K^{1-\alpha} \\
\ln P &= \ln b + \alpha \ln L + (1-\alpha)\ln K \\
\ln P &= \ln b + \alpha \ln L + \ln K - \alpha \ln K \\
\ln P - \ln K &= \ln b + \alpha(\ln L - \ln K) \\
\ln \frac{P}{K} &= \ln b + \alpha \ln \frac{L}{K}
\end{aligned}$$

Consolidating Data and using Exercise 81a to Find α and b

The CDPF written in another form tells us that $y = \ln b + \alpha x$ where $y = \ln \frac{P}{K}$ and $x = \ln \frac{L}{K}$. P is defined as a team's total wins every year, L as total goals scored every year, and K as the number of shots on goal. All of these variables are collected from 2008-2016. The python script that consolidated the data into these numbers is linked here. The script works by finding each team's P, L, and K values summed together each year from the dataset and solving for each point $(\ln \frac{L}{K}, \ln \frac{P}{K})$, plotting them and then finding the least squares regression line through those points.

Finding α and b cont.

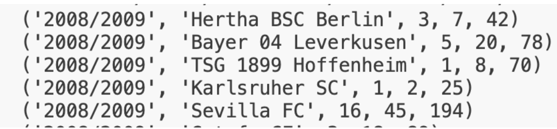


```
import pandas as pd
import xml.etree.ElementTree as ET
import math
import numpy as np
import matplotlib.pyplot as plt

# Load the CSV files
file_path = 'Match.csv'  # Update with the actual path
df = pd.read_csv(file_path)

team_file_path = 'Team.csv'  # Update with the actual path
teams_df = pd.read_csv(team_file_path)
team_id_to_name = dict(zip(teams_df['team_api_id'], teams_df['team_long_name']))

# Function to count shots on target from XML
def count_shots_on_target(xml_data):
    if pd.isna(xml_data) or not xml_data.strip():
        return 0
    try:
        root = ET.fromstring(xml_data)
        shoton_count = 0
        for value in root.findall(".//value"):
            shoton = value.find(".//shoton")
            if shoton is not None and shoton.text == '1':
                shoton_count += 1
        return shoton_count
    except ET.ParseError:
        return 0

# Apply the function and filter out rows with no shoton XML data
df['shots_on_target'] = df['shoton'].apply(count_shots_on_target)
df = df[df['shots_on_target'] > 0]  # Keep only rows with shoton counts greater than 0

# Filter rows starting from index 1730
df_filtered = df.iloc[1730:].reset_index(drop=True)

# Calculate wins
df_filtered['home_win'] = df_filtered['home_team_goal'] > df_filtered['away_team_goal']
df_filtered['away_win'] = df_filtered['away_team_goal'] > df_filtered['home_team_goal']

# Create separate dataframes for home and away teams
df_home = df_filtered.copy()
df_home['team'] = df_home['home_team_api_id']
df_home['goals'] = df_home['home_team_goal']
df_home['shots_on_target'] = df_home['shots_on_target']
df_home['total_wins'] = df_home['home_win']

df_away = df_filtered.copy()
df_away['team'] = df_away['away_team_api_id']
df_away['goals'] = df_away['away_team_goal']
df_away['shots_on_target'] = df_away['shots_on_target']
df_away['total_wins'] = df_away['away_win']

# Aggregate data by season and team
home_stats = df_home.groupby(['season', 'team']).agg({
    'total_wins': 'sum',
```

Figure 3: The first few lines of the script

Finding α and b cont.

Here is a demonstration using the first five teams and their P, L, K values.



```
('2008/2009', 'Hertha BSC Berlin', 3, 7, 42)
('2008/2009', 'Bayer 04 Leverkusen', 5, 20, 78)
('2008/2009', 'TSG 1899 Hoffenheim', 1, 8, 70)
('2008/2009', 'Karlsruher SC', 1, 2, 25)
('2008/2009', 'Sevilla FC', 16, 45, 194)
```

Figure 4: The first five teams outputted with their P, L, and K values in order from left to right

$$P_1\colon\ (\ln \tfrac{L}{K}, \ln \tfrac{P}{K}) = (\ln \tfrac{7}{42}, \ln \tfrac{3}{42}) \approx (-1.7918, -2.6391)$$

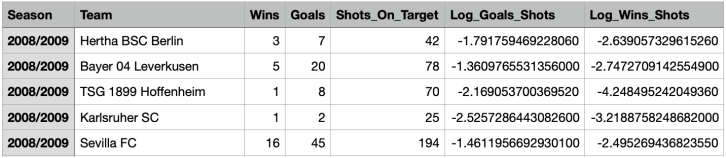
$$P_2\colon\ (\ln \tfrac{L}{K}, \ln \tfrac{P}{K}) = (\ln \tfrac{20}{78}, \ln \tfrac{5}{78}) \approx (-1.361, -2.7473)$$

Finding α and b cont.

$$P_3\colon\ (\ln\tfrac{L}{K}, \ln\tfrac{P}{K}) = (\ln\tfrac{8}{70}, \ln\tfrac{1}{70}) \approx (-2.1691, -4.2485)$$

$$P_4\colon\ (\ln\tfrac{L}{K}, \ln\tfrac{P}{K}) = (\ln\tfrac{2}{25}, \ln\tfrac{1}{25}) \approx (-2.5257, -3.2189)$$

$$P_5\colon\ (\ln\tfrac{L}{K}, \ln\tfrac{P}{K}) = (\ln\tfrac{45}{194}, \ln\tfrac{16}{194}) \approx (-1.4612, -2.4953)$$



Season	Team	Wins	Goals	Shots_On_Target	Log_Goals_Shots	Log_Wins_Shots
2008/2009	Hertha BSC Berlin	3	7	42	-1.791759469228060	-2.639057329615260
2008/2009	Bayer 04 Leverkusen	5	20	78	-1.3609765531356000	-2.7472709142554900
2008/2009	TSG 1899 Hoffenheim	1	8	70	-2.169053700369520	-4.248495242049360
2008/2009	Karlsruher SC	1	2	25	-2.5257286443082600	-3.2188758248682000
2008/2009	Sevilla FC	16	45	194	-1.4611956692930100	-2.495269436823550

Figure 5: Script calculated values for each point, which match up with P_1 to P_5; full data can be found here.

Finding α and b cont.

The least squares regression line is given by the equation $y = mx + c$, where x and y are independant from the CDPF; using x and y here is just for convenience when graphing the line later on. m is given by $\frac{n\sum(x_i y_i) - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$ where n is the number of data points and each (x_i, y_i) is a point. In the context of this example, each (x_i, y_i) is a point calculated earlier; (x_1, y_1) is P_1, (x_2, y_2) is P_2, and so on until P_5. c is given by $\frac{\sum y_i - m\sum x_i}{n}$.

Finding α and b cont.

To find α and b, we use linear regression on the transformed data:

- $x = \ln \frac{L}{K}$ and $y = \ln \frac{P}{K}$.
- The least squares regression line is $y = mx + c$, where m is the slope and c is the intercept.
- Calculate the slope m:

$$m = \frac{n \sum (x_i y_i) - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

- Calculate the intercept c:

$$c = \frac{\sum y_i - m \sum x_i}{n}$$

- Identify α as the slope m and $\ln b$ as the intercept c.
- Exponentiate c to find b: $b = e^c$.

Finding α and b cont.

Using all five points, we can solve for each value.

$$
\begin{aligned}
m &= \frac{n\sum(x_i y_i) - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \\
n &= 5 \\
(x_i, y_i) &= P_i \\
\sum(x_i y_i) &= (-1.7918)(-2.6391)+ \\
&\quad (-1.361)(-2.7473)+ \\
&\quad (-2.1691)(-4.2485)+ \\
&\quad (-2.5257)(-3.2189)+ \\
&\quad (-1.4612)(-2.4953) \\
\sum(x_i y_i) &\approx 4.7350 + 3.7443 + 9.2070 + 8.1327 + 3.6447 \\
\sum(x_i y_i) &\approx 29.4637
\end{aligned}
$$

Finding α and b cont.

$$\begin{aligned}
\sum x_i \sum y_i &= (-1.7918 - 1.361 - 2.1691 - 2.5257 - 1.4612) \\
&\quad (-2.6391 - 2.7473 - 4.2485 - 3.2189 - 2.4953) \\
\sum x_i \sum y_i &= (-9.3096)(-15.3481) \\
\sum x_i \sum y_i &\approx 142.4604
\end{aligned}$$

$$\begin{aligned}
\sum x_i^2 &= (-1.7918)^2 + (-1.361)^2 + (-2.1691)^2 + (-2.5257)^2 + (-1.4612)^2 \\
\sum x_i^2 &\approx 18.2821
\end{aligned}$$

$$\begin{aligned}
(\sum x_i)^2 &= (-1.7918 - 1.361 - 2.1691 - 2.5257 - 1.4612)^2 \\
(\sum x_i)^2 &\approx 86.6538
\end{aligned}$$

Finding α and b cont.

$$m = \frac{5(29.4637) - 142.4604}{5(18.2821) - 86.6538}$$

$$m \approx 1.0213$$

$$b = \frac{\sum y_i - m \sum x_i}{n}$$

$$\sum y_i = -2.6391 - 2.7473 - 4.2485 - 3.2189 - 2.4953$$

$$\sum y_i = -15.3491$$

$$\sum x_i = -1.7918 - 1.361 - 2.1691 - 2.5257 - 1.4612$$

$$\sum x_i = -9.3088$$

$$b = \frac{-15.3491 - 1.0213(-9.3088)}{5}$$

$$b \approx -1.1685$$

$$y = 1.0213x - 1.1685$$

Finding α and b cont.

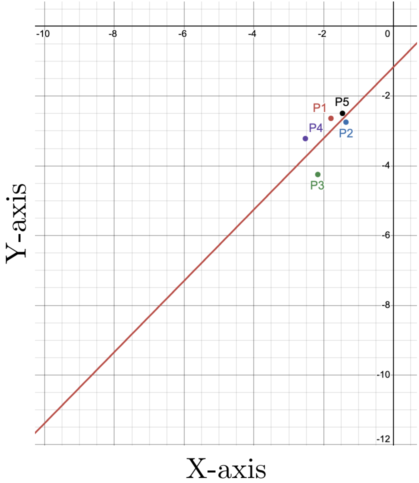


Figure 6: P_1 to P_5 plotted, along with the least squares regression line $y = 1.0213x - 1.1685$

Finding α and b cont.

Obviously, this line is only a good fit for these five points. Using the same approach, the script plots all 528 points and finds the line of best fit for all the points.

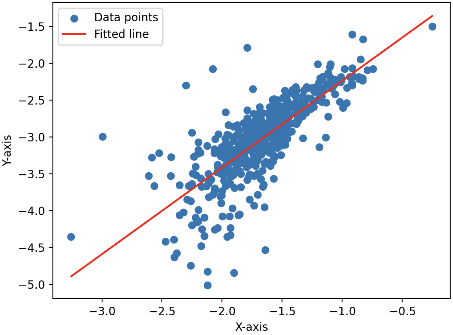


Figure 7: The least squares regression line through every point $(\ln \frac{L}{K}, \ln \frac{P}{K})$, which is $y = 1.1755x - 1.0643$

Finding α and b cont.

Since the line of best fit is $y = 1.1755x - 1.0643$, we can compare coefficients with $y = \alpha x + \ln b$ to get the following:

$$\begin{aligned} \alpha &= 1.1755 \\ \ln b &= -1.0643 \end{aligned}$$

Exponentiating both sides to solve for b gives the following:

$$\begin{aligned} e^{\ln b} &= e^{-1.0643} \\ b &= e^{-1.0643} \\ b &\approx 0.345 \end{aligned}$$

Writing the CDPF and Determining Accuracy

Therefore, the CDPF for this data set can be written as $P = 0.345L^{1.1755}K^{-0.1755}$. By evaluating each point using the CDPF, then comparing it to each true P value, error can be calculated as the average of the absolute differences between the predicted and actual values. The script calculates the mean absolute error to be 1.77, meaning that on average, the CDPF is off by abount 2 wins. When plotting each point (x, y) where x and y are actual wins and predicted wins respectively, a separate least regression calculation gives us the line of best fit as $y = 1.0361x + 0.0975$, as compared to $y = x$, where the CDPF predicted value would be the actual value for every data point.

Writing the CDPF and Determining Accuracy cont.

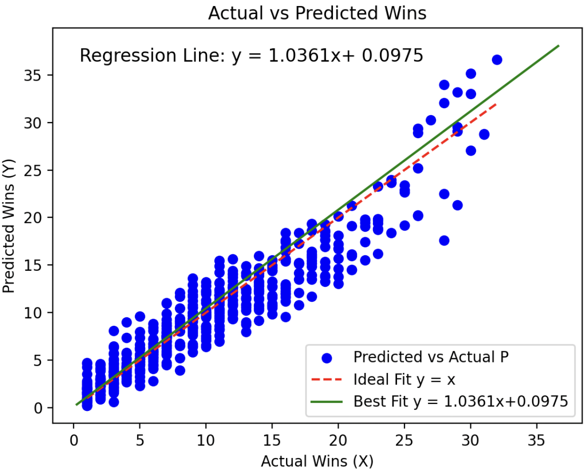


Figure 8: A graph of actual vs predicted wins evaluated by the CDPF using each L and K, with the line of best fit being $y = 1.0361x + 0.0975$, compared to the line of ideal fit $x = y$.

Analysis of $\frac{\partial P}{\partial L}$, $\frac{\partial P}{\partial K}$, and ∇P

The CDPF for this data set $P = 0.345L^{1.1755}K^{-0.1755}$ has two partial derivatives: $\frac{\partial P}{\partial L}$ and $\frac{\partial P}{\partial K}$. $\frac{\partial P}{\partial L}$ represents how the number of wins changes in response to a small change in the number of goals, while keeping the number of shots on goal constant. Conversely, $\frac{\partial P}{\partial K}$ represents how the number of wins changes based on a small change in the number of shots on goal, while keeping goals constant. When evaluated at a point in the domain of P, which represents a specific number of goals scored and shots on target for a team, the partial derivatives offer insight on which variable is most impactful towards an increase in wins. Finding the best way to increase wins, however, is done by taking both variables in consideration, which is what ∇P represents: the direction of the steepest increase of wins.

Analysis of $\frac{\partial P}{\partial L}$, $\frac{\partial P}{\partial K}$, and ∇P cont.

A team manager can use this information to prioritize strategies that are most effective in increasing wins. If $\frac{\partial P}{\partial L}$ is evaluated at a point (L, K) where that point represents a team's performance in a game and found to be signficantly larger than $\frac{\partial P}{\partial K}$ at the same point, then the manager might choose to prioritize training towards scoring more goals over shooting at the goal. The same is true for when $\frac{\partial P}{\partial K} > \frac{\partial P}{\partial L}$. This information can be demonstrated by doing an example with one of the teams analyzed earlier.

Hertha BSC Berlin won 3 times, scoring 7 goals and kicking 42 shots towards the goal in the 2008-2009 season. $\frac{\partial P}{\partial L}$ is given by $\frac{\partial}{\partial L}(0.345L^{1.1755}K^{-0.1755})$

Analysis of $\frac{\partial P}{\partial L}$, $\frac{\partial P}{\partial K}$, and ∇P cont.

- Start with the CDPF: $P = 0.345L^{1.1755}K^{-0.1755}$.
- Calculate $\frac{\partial P}{\partial L}$:

$$\frac{\partial P}{\partial L} = \frac{\partial}{\partial L}\left(0.345L^{1.1755}K^{-0.1755}\right) = 0.4055\left(\frac{L}{K}\right)^{0.1755}$$

- Evaluate at a specific point (L, K) to find the effect of one additional goal on wins.
- Calculate $\frac{\partial P}{\partial K}$:

$$\frac{\partial P}{\partial K} = \frac{\partial}{\partial K}\left(0.345L^{1.1755}K^{-0.1755}\right) = -0.0605\left(\frac{L}{K}\right)^{1.1755}$$

- Evaluate at a specific point (L, K) to find the effect of one additional shot on wins.

Analysis of $\frac{\partial P}{\partial L}$, $\frac{\partial P}{\partial K}$, and ∇P cont.

Now, we evaluate at the specific point (7,42)

$$\left.\frac{\partial P}{\partial L}\right|_{(7,42)} = 0.4055 \left(\frac{7}{42}\right)^{0.1755}$$

$$\left.\frac{\partial P}{\partial L}\right|_{(7,42)} \approx 0.2961$$

$$\left.\frac{\partial P}{\partial K}\right|_{(7,42)} = -0.0605 \left(\frac{7}{42}\right)^{1.1755}$$

$$\left.\frac{\partial P}{\partial K}\right|_{(7,42)} \approx -0.0073$$

$$\nabla P = \left(\frac{\partial P}{\partial L}, \frac{\partial P}{\partial K}\right)$$

$$\left.\nabla P\right|_{(7,42)} \approx (0.2961, -0.0073)$$

Analysis of $\frac{\partial P}{\partial L}$, $\frac{\partial P}{\partial K}$, and ∇P cont.

$\frac{\partial P}{\partial L}$ indicates that for each additional goal scored, the number of wins is expected to increase by approximately 0.297, assuming shots on target remain constant. $\frac{\partial P}{\partial K}$ indicates that an increase in shots on target has a very slight negative effect on wins. This result is intuitive as it aligns with practical observations in football. When the number of shots on target increases while the number of goals remains unchanged, it typically indicates that the team's efficiency in converting shots into goals is low. An increased number of unsuccessful shots not only fails to add to the goal count but also may result in lost possession, thereby potentially benefiting the opposing team. ∇P combines both values, indicating to the team manager that this specific team should focus more on shot accuracy, rather than shot quantity.

Budget Analysis

Suppose that a soccer team has a fixed budget of p dollars. Assuming that the training cost of producing L is $n > 0$ and the training cost of producing K is $m > 0$, then $p = Ln + Km$. Another constraint that is to be noted is that $L \leq K$, since each goal is also a shot on. Lagrange Multipliers can maximize the CDPF $P = 0.345L^{1.1755}K^{-0.1755}$ based on the constraint budget p.

Budget Analysis cont.

To maximize the CDPF $P = 0.345L^{1.1755}K^{-0.1755}$ under a budget constraint $p = Ln + Km$:

- Define the constraint function $g = Ln + Km - p$.
- Use Lagrange multipliers: $\nabla P = \lambda \nabla g$.
- Calculate ∇P:

$$\nabla P = \left(0.4055L^{0.1755}K^{-0.1755}, -0.0605L^{1.1755}K^{-1.1755}\right)$$

- Calculate ∇g:

$$\nabla g = (n, m)$$

- Set up the system of equations:

$$\begin{cases} \lambda n = 0.4055L^{0.1755}K^{-0.1755} \\ \lambda m = -0.0605L^{1.1755}K^{-1.1755} \end{cases}$$

- Solve for λ and equate the two expressions to find the relationship between L and K.
- Substitute back into the constraint to find the critical points.

Budget Analysis cont.

$$\begin{aligned}
\lambda n &= 0.4055L^{0.1755}K^{-0.1755} \\
\lambda m &= -0.0605L^{1.1755}K^{-1.1755} \\
\lambda &= \frac{0.4055}{n}L^{0.1755}K^{-0.1755} \\
\frac{0.4055m}{n}L^{0.1755}K^{-0.1755} &= -0.0605L^{1.1755}K^{-1.1755} \\
\frac{0.4055m}{n} &= \frac{-0.0605L}{K} \\
\frac{0.4055Km}{n} &= -0.0605L \\
L &= \frac{-6.7025Km}{n} \\
p &= Ln + Km \\
p &= -6.7025Km + Km
\end{aligned}$$

Budget Analysis cont.

$$\begin{aligned} Km &= \frac{p}{-5.7025} \\ K &= \frac{p}{-5.7025m} \end{aligned}$$

$$L = \frac{-6.7025p}{-5.7025n} = \frac{1.1754p}{n}$$

From the Lagrange Multiplier, we obtain that a critical point is $(\frac{1.1754p}{n}, \frac{p}{-5.7025m})$. However, since p, n, m are all positive numbers, this means that $\frac{1.1754p}{n} > 0 > \frac{p}{-5.7025m}$, or $L > K$, which is not in the constraint of $L \leq K$. Therefore, this critical point can be discarded, which only leaves the boundary points. Normally, there would be two points that occur when L or K are zero, but since L,$K > 0$ in accordance to the domain of $\ln \frac{L}{K}$ derived earlier, the only boundary point to be considered is when $L = K$.

Budget Analysis cont.

When $L = K = \frac{p}{n+m}$, $P = 0.345L = \frac{0.345p}{n+m}$, which is the maximum. This result makes since, as when the budget, p, grows larger, the number of wins grows larger. Conversely, as the cost to produce goals and shot ons grows larger, the number of wins grows smaller.

To maximize wins, the team should allocate its budget so that the number of goals equals the number of shots on target because this balance is derived from the Lagrange multiplier method. The critical point analysis shows that $L = K = \frac{p}{n+m}$ is optimal. This equal allocation ensures the maximum number of wins, as a team will be in a winning advantage if they are scoring in all of its attempts to shoot towards the goal. Of course, scoring in every attempt isn't very realistic for a team manager to consider, so a team manager's best option is to train shot accuracy, to try to score every shot the team takes, also as emphasized by ∇P.

Conclusion

In conclusion, after extensive analysis, it has been been determined that in terms of shots taken, one should emphasize quality over quantity. By analyzing every team's wins, total goals scored and total shots toward goal attempted, a least squares regression line was able to be calculated, which led to a function P, predicting a team's number of wins based on the two variables. From there, two groups of optimization were calculated; one with a set budget, one without; and both led to the same overall conclusion. A potential problem that arises with this model comes from the fact that it only analyzes wins based on two variables; a model with more variables takes into account more factors that lead to the winner of a match, which means it will be more accurate and have less error. Nevertheless, the CDPF was able to predict wins with relatively minimal error.