A Replication Study of Civil War Exposure and Violence Using Soccer Data

ECON 148 Project 3

Aneesh Patel, Weiqian Peng, Boyu Fan, Bennett Fees

April 9th, 2024

Abstract

This paper replicates the seminal work of Miguel, Saiegh, and Satyanath (2011), who examined the relationship between exposure to civil conflict and aggressive behavior in professional soccer players. Leveraging disciplinary records from European soccer leagues spanning 1980 to 2005, our replication effort corroborates the original study's conclusion of a statistically significant association between civil war exposure and the likelihood of incurring yellow card penalties on the field. Employing Python, we meticulously reproduce the original analysis, ensuring computational reproducibility while expanding accessibility. Despite minor variations, our results affirm the robustness of Miguel et al.'s findings, underscoring the enduring impact of civil conflict on individual behavior.

KEYWORDS: Civil war exposure, professional soccer, violent behavior, replication study, computational reproducibility

1 Introduction

In their exploration of the behavioral consequences of civil war exposure, Miguel et al. (2011) investigate the link between civil conflict in a player's home country and their propensity to incur penalties in professional soccer leagues. This paper utilizes the same processed data source—the disciplinary records from six major European soccer leagues, encompassing players from over 70 countries with varying experiences of civil conflict. The primary data span from 1980 to 2005, focusing on yellow and red cards as proxies for aggressive behavior.

The research design applies a comparative analysis method, isolating the impact of civil war exposure by controlling for multiple player and country characteristics. Miguel et al (2011) use the institutional penalty system in international soccer to estimate changes in behavior across individuals from countries that have and haven't experienced civil war finding a "strong correlation between the number of years of civil conflict in a player's native country and his likelihood of earning yellow and red cards in Europe" at a 99% confidence interval (Miguel et al., 2011, p. 60). This approach moves past estimation techniques that rely on surveys and instead use the amount of penalties the player incurs to estimate violent behavior. Their regressions support their argument that civil war exposure significantly influences individual behavior in a structured, rule-bound environment like professional soccer.

In our replication attempt, we accessed the original data and computational codes directly from the corresponding author's repository. This replication effort is aimed at both verifying these findings and expanding the scope of analysis through the translation of the original Stata code into Python, which enables broader accessibility and flexibility in the analytical process. By re-coding the study's original methodology, we engage in computational reproducibility, striving to duplicate the original study's results using identical datasets and analytical procedures albeit in a different programming environment. While slight variations in coefficient values were noted, they did not materially affect the direction or significance of the original

findings. These variations are explored further in the results section of this paper, suggesting robustness in the original study's conclusions.

2 Reproducibility

In replicating the original study, which scrutinizes the association between civil war exposure and aggressive behavior in soccer using disciplinary actions as a proxy, we have closely followed the detailed instructions given. The study's approach to collect and analyze data made our endeavor of trying to recreate the same results straightforward, with the goal of precisely matching the original tables and figures.

Upon our examination, we noticed a minor discrepancy in the coding results. The replication diverged from the published outcomes in the case of the coefficient for players in German leagues. In the original paper, this coefficient was reported as 0.313, whereas in our replication, it was slightly higher at 0.317. Correspondingly, the standard error exhibited a slight deviation, from 6.37 in the original study to 6.58 in our replication. Despite these minor variances, the overall statistical significance remains unaffected, confirming the robustness of the original conclusions.

While looking into the supplementary materials, our findings were consistent with the updated empirical results delineated in the 'Table 1: Updated Empirical Results for Table 2' found in the readme.pdf within the replication folder. This reaffirms the integrity of the original results despite the aforementioned minor deviations.

In addition, a noteworthy complication arose when attempting to replicate a table in the appendix of the original study that involved the 'Log years of civil war.' The issue encountered was that taking the logarithm of zero years of civil war—which applies to some countries—yields an undefined value. The original paper does not explicitly address this issue, thus creating a hurdle in the replication process.

Besides the primary results and conclusions, we have also explored an intriguing aspect of the study's discussion. The authors note that in some cases, the primary

data source did not provide specific information regarding players' places of birth or indicated dual nationality. In these instances, the players' participation in national teams was used to infer nationality. This, as the authors conjecture, likely leads to an underestimation of the impact of civil war exposure on violent conduct in soccer. To verify this assumption, the authors executed robustness checks using alternative data sources; however, the specific methodology and results of these checks were not delineated in the study.

Addressing this omission, we propose to conduct a robustness check by creating a binary variable indicating players for whom place of birth information is lacking. We hypothesize, in line with the authors' note, that including this variable in our core specifications would result in an increased point estimate for the influence of exposure to civil war on soccer violence. This would suggest that the original coding criteria, while not ideal, likely do not exaggerate the effect of civil war exposure. Such an analysis would offer an additional layer of validation for the original study's findings and ensure that our replication captures the nuances of the data and its implications.

3 Replication

3.1 Regression model

We utilized a **Negative Binomial Regression Model** for analyzing the count data, specifically the number of yellow cards issued in soccer games. This model choice is ideal for count data exhibiting over-dispersion where the variance significantly exceeds the mean.

3.1.1 Regression Model Description The dependent variable:

- *yellow_cards*: A discrete variable representing the total count of yellow cards received by individual players across matches.

The independent variables span several domains, encapsulating both player-specific attributes and broader country characteristics, which are hypothesized to

influence on-field conduct:

- **Country characteristics:**
  - *Years of civil war*: A continuous measure reflecting the accumulated years of civil strife experienced in a player's home country.
  - *Civil war years postbirth* and *prebirth*: These variables measure the exposure to civil conflict after and before a player's birth, providing a temporal dimension to potential psychosocial impacts.
  - *Log GNI per capita*: Logarithm of the Gross National Income per capita, serving as a proxy for the economic conditions of the player's country.
  - *Rule of Law*: An index measuring the adherence to law and order within a player's nation, influencing societal norms.
- **Player characteristics:**
  - *Age*: The player's age, a potential indicator of maturity and experience.
  - *Log transfer fee*: The natural logarithm of the transfer fee, indicating the market value and possibly the skill level of the player.
  - *Games Started* and *Substitute*: These variables account for the player's participation as a starter or a substitute, which might correlate with playing style and aggression.
  - *Defender*, *Forward*, *Midfield*, *Goalie*: Positional dummies to control for the role-specific behaviors inherent to different playing positions.
  - *Goals*: The number of goals scored by the player, inversely related to defensive aggression in some interpretations.
- **League Membership**:
  - League-specific dummy variables (*Italian League*, *European Champions League*, etc.) account for the competitive environment and refereeing rigor, which may vary across leagues.

- **Regional Fixed Effects**: Dummy variables for each region (*africa*, *asia*, *lac*, *east_europe*) to control for unobserved heterogeneity linked to geographical and cultural contexts.

Estimation Procedure:

The regression employs robust standard errors clustered by the *nation* to account for intra-national correlation—players from the same country may exhibit similar behavior due to shared cultural and socio-economic influences. This clustering is crucial in yielding standard errors that are robust to such within-group correlation.

Interpretation of Coefficients:

The coefficients derived from the negative binomial regression are interpreted as the logarithm change in the expected count of the dependent variable for a one-unit change in the predictor variable, holding all other factors constant. A positive coefficient suggests an increase in the expected count, while a negative coefficient indicates a decrease.

Model Diagnostics and Fit:

The model's goodness of fit will be assessed using pseudo R^2 measures, likelihood ratio tests, and examination of over-dispersion statistics. Observations for the analysis are conditioned on league participation and the representation threshold of countries, ensuring a sample that is conducive to a rigorous empirical inquiry.

3.1.2 Rationale for Using Negative Binomial Regression The negative binomial regression is selected for its capability to handle variance greater than the mean, which is common in sports analytics. This model:

- Offers **flexibility** in dealing with varied data distributions that do not conform to the Poisson assumption of equal mean and variance.
- Includes an **adjustable dispersion parameter** that models the extra variability observed in the data.

In conclusion, the regression model aims to uncover nuanced insights into the determinants of disciplinary actions within the realm of professional soccer,

providing empirical grounding for theoretical postulations on the influence of socio-economic factors and individual attributes on player behavior.

3.2 Visualizations

In this section, we discuss the process of reproducing the visualizations that were present in the original paper. Visualizations play a crucial role in conveying the complex relationships and patterns within the data, providing insights that may not be readily apparent from textual descriptions alone. By replicating the original visualizations, we seek to validate the findings of the study and assess the robustness of its conclusions. Through this process, we aim to contribute to the transparency and reproducibility of the research, facilitating a deeper understanding of the relationship between civil war exposure and violent behavior among soccer players.

3.2.1 Original Visualizations The original paper includes four key visualizations that provide insights into the relationship between civil war exposure and violent behavior among soccer players. The first visualization, *Figure 1*, is a bar graph depicting the percentage of yellow cards issued based on the type of offense in Italy's Serie A and the UEFA Champions League. This visualization aims to elucidate the primary causes of yellow cards, with a focus on violent acts such as assault or extreme unsporting behavior, which constitute the vast majority of infractions across both leagues.

The remaining three visualizations consist of scatter plots representing the results of regression analyses predicting the average number of yellow cards per player season in relation to the average years of civil war since 1980 per country. Each point on the scatter plot represents a different country, with the x-axis showing the residuals of the years of civil war since 1980 and the y-axis displaying the residuals of the average yellow cards per player season. Additionally, the size of each point corresponds to the number of players from the respective country, and a line of best fit is included to illustrate the association between these variables.

The second visualization, *Figure 2*, presents the scatter plot for all countries

in the dataset, highlighting a strong positive trend between years of civil war and the number of yellow cards. The third visualization, *Figure 3*, focuses on non-OECD countries exclusively and demonstrates a consistent positive trend similar to the previous plot. Finally, the fourth visualization, *Figure 4*, also focuses on non-OECD countries but excludes Colombia, Iran, Israel, Peru, and Turkey, revealing a slightly weaker but still positive association between civil war exposure and violent behavior in soccer.

3.2.2 Replication Process To replicate the analysis conducted in the original paper, we utilized the datasets and Stata code snippets provided by the authors. Our replication efforts focused on recreating the original analysis using the Python programming language. We employed various libraries and packages to assist in this analysis, including Pandas for data cleaning and processing, Statsmodels for running the regression analysis, and Matplotlib to generate visualizations that closely resembled the original visualizations presented in the paper.

Throughout the replication process, we meticulously followed the steps outlined in the original State code, aiming to reproduce the results as faithfully as possible. However, we also exercised critical scrutiny, identifying any discrepancies or ambiguities in the code and making necessary adjustments to ensure accuracy and consistency. Our goal was to maintain the integrity of the original analysis while addressing any potential shortcomings or areas for improvement. By adhering closely to the original methodology while utilizing Python, we aimed to provide a comprehensive and transparent replication of the findings reported in the original paper.

3.2.3 Results of Replication The replication process yielded visualizations that closely resembled the original visualizations presented in the original paper. All replicated visualizations can be found in Section 5. These visualizations include a bar graph illustrating the percentage of yellow cards given according to the type of offense (Figure 1), as well as scatter plots representing the regression analyses

predicting the average number of yellow cards per player season and average years of civil war since 1980 (Figures 2, 3, and 4).

While the resulting visualizations closely matched the originals, it is important to note that some nuances in Stata-specific functions posed challenges during replication. Specifically, functions related to grouping and aggregating dataframes, as well as some specifics of plot customization, such as sizing the dots and adding a fit line, required more creative adaptations in Python. Despite these challenges, our replication efforts successfully produced visualizations that mirrored the original findings, validating the robustness of the analysis conducted in the original paper.

4 Conclusion

The study conducted by Miguel et al. (2011) delves into the behavioral implications of civil war exposure, particularly its correlation with penalties incurred by professional soccer players. Using disciplinary records from major European soccer leagues spanning over 70 countries, the research focuses on yellow and red cards as indicators of aggressive conduct. Through a comparative analysis method, the original authors isolate the impact of civil war exposure, revealing a significant correlation between years of civil conflict in a player's native country and their likelihood of receiving penalties in European leagues.

In our replication process, we accessed the original datasets, translated the original Stata code into Python, and successfully validated the results of the analysis carried out by the authors. The implications of these replication results underscore the importance of transparency and reproducibility in research. By meticulously documenting our replication process and addressing challenges encountered along the way, we contribute to the broader scientific endeavor of verifying and validating research findings. Additionally, our replication serves as a testament to the reliability and accuracy of the original analysis, providing confidence in the reported relationship between civil war exposure and violent behavior on the soccer field.

Looking ahead, future research endeavors could explore additional dimensions

of the relationship between civil war exposure and behavioral outcomes in professional sports. Further investigations might delve into the long-term effects of civil conflict on player performance and career trajectories. Additionally, incorporating qualitative methodologies, such as interviews or focus groups with athletes, could provide deeper insights into the subjective experiences and perceptions of individuals affected by civil war. Furthermore, expanding the analysis to encompass a broader range of sports and geographical regions could offer a more comprehensive understanding of the intersection between conflict exposure and sporting behavior. These avenues of inquiry hold promise for advancing our comprehension of the intricate interplay between socio-political contexts and individual behaviors in sports settings.

References

Miguel, E., Saiegh, S. M. and Satyanath, S.: 2011, Civil War Exposure and Violence, *Economics & Politics* **23**(1), 59–73.

Miguel et al. (2011)

5 Replicated Visualizations

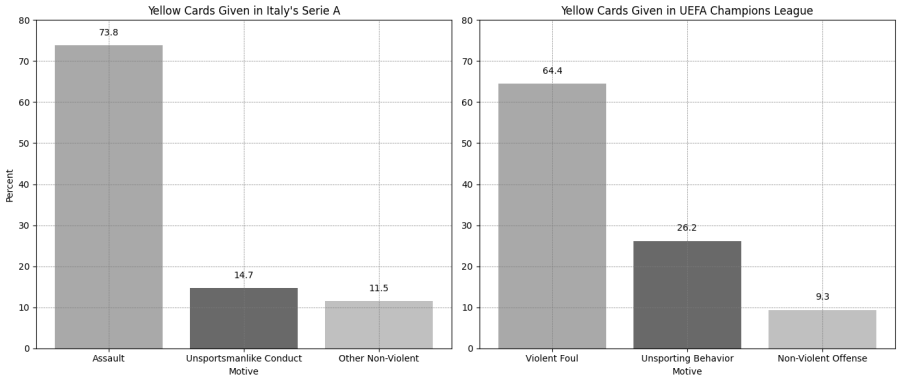


Figure 1: Yellow cards according to type of offense.

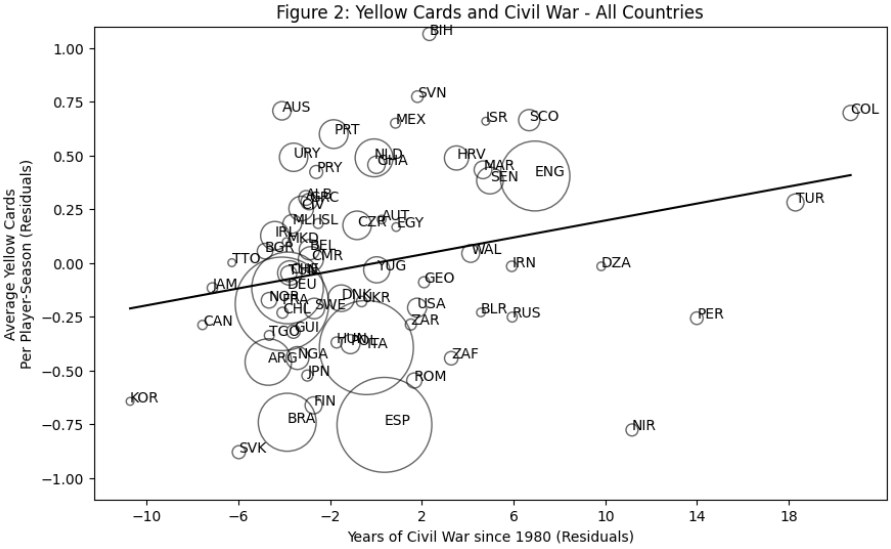


Figure 2: Yellow cards and civil war (conditional on control variables in Table 2, regression 1) – all countries.

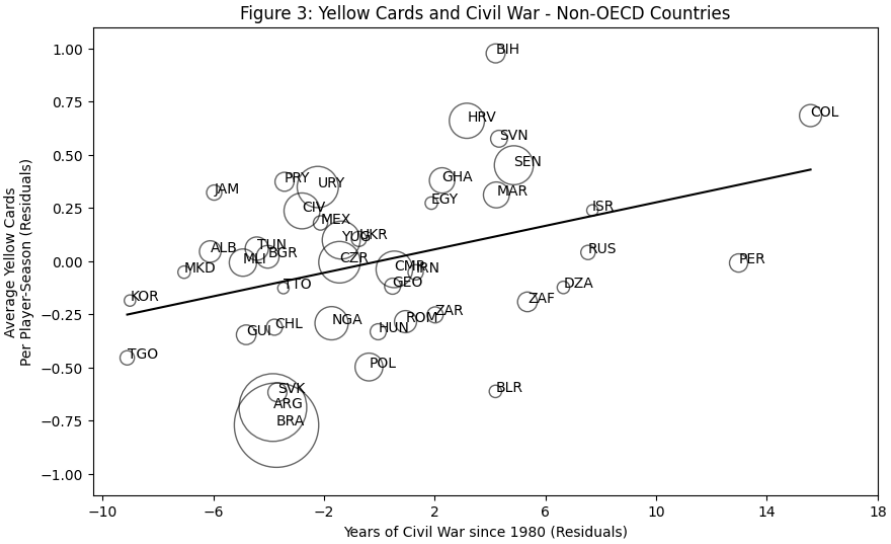


Figure 3: Yellow cards and civil war (conditional on control variables in Table 2, regression 1) – non-OECD countries.

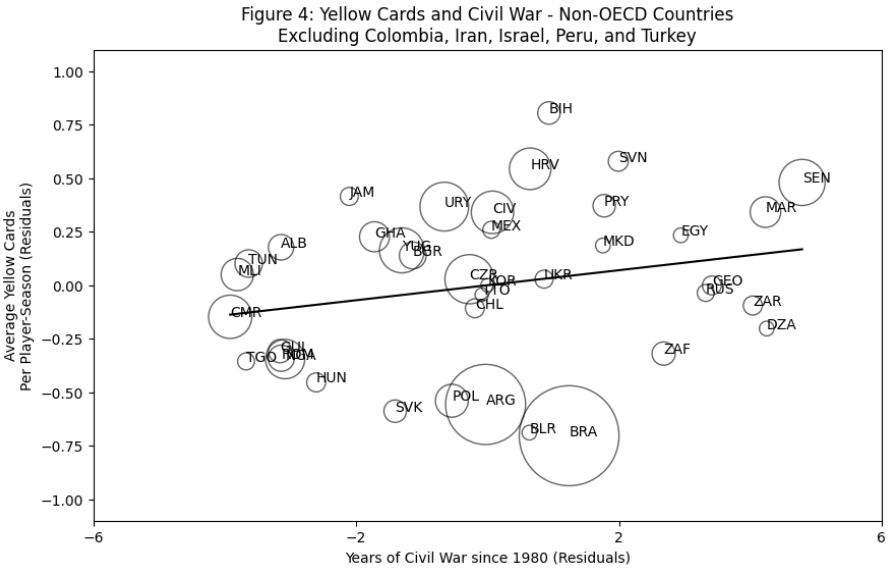


Figure 4: Yellow cards and civil war (conditional on control variables in Table 2, regression 1) – non-OECD countries, excluding Colombia, Iran, Israel, Peru, and Turkey.

6 Replicated Tables

Table 1: Table 1: Descriptive Statistics

Variable	Mean	Standard deviation	Minimum	Maximum	Observations
Rule infractions					
yellow_card	2.43	2.73	0.00	16.00	5,035
red_card	0.16	0.42	0.00	3.00	5,035
Country characteristics					
Years of civil war	2.74	4.74	0.00	26.00	5,035
Rule of Law	0.85	0.89	-1.76	2.10	5,035
Log GNI per capita	9.97	0.82	6.58	10.70	4,965
Player characteristics					
Age	25.99	4.40	17.00	41.00	5,035
weekly_wage	23,991.38	27,014.38	0.00	190,000.00	5,034
Log transfer fee	15.15	1.08	8.01	18.17	5,035
Games Started	13.80	11.49	0.00	40.00	5,035
Substitute	3.13	3.90	0.00	29.00	5,035
Goalie	0.08	0.27	0.00	1.00	5,035
Defender	0.33	0.47	0.00	1.00	5,035
Forward	0.23	0.42	0.00	1.00	5,035
Midfield	0.36	0.48	0.00	1.00	5,035
Goals	1.65	3.13	0.00	31.00	5,035
Player region of origin					
africa	0.07	0.26	0.00	1.00	5,035
asia	0.00	0.06	0.00	1.00	5,035
lac	0.12	0.33	0.00	1.00	5,035
east_europe	0.07	0.25	0.00	1.00	5,035
oecd	0.74	0.44	0.00	1.00	5,035
Soccer leagues					
English League	0.17	0.38	0.00	1.00	5,035
European Champions League	0.19	0.39	0.00	1.00	5,035
French League	0.15	0.36	0.00	1.00	5,035
German League	0.14	0.35	0.00	1.00	5,035
Italian League	0.17	0.38	0.00	1.00	5,035
Spanish League	0.17	0.37	0.00	1.00	5,035

Notes: The source of the rule infraction, goals, player characteristics, player country of origin, and soccer leagues data are the ESPN Soccernet website. Weekly salaries and transfer fees are expressed in current U.S. dollars. Sources: Football Manager, 2005, and World Soccer Manager, 2006. The source of the civil war data is the PRIO/Uppsala Armed Conflict database, and the source of the rule of law variable is the WGI project. Income per capita is measured in purchasing power parities (PPP) (2006 dollars); World Bank's World Development Indicators (2007).

Table 2: Empirical Results: Civil War

	Yellow cards	Yellow cards	Yellow cards	Red cards	Goals scored	Yellow cards	Red cards
	(2)	(2)	(3)	(4)	(5)	(6)	(7)
Country characteristics							
Years of civil war	0.0076(2.63)***	0.0078(2.51)**	0.0075(2.59)***	0.0126(1.92)*	0.0001(0.02)		
Civil war years postbirth						0.0052(1.86)	0.014 (2.13)
Civil war years prebirth						0.0036(0.73)	0.004 (0.41)
Log GNI per capita		0.046 (1.06)					
Rule of law			0.019 (0.40)	−0.143 (1.46)	0.006 (0.15)		
Player characteristics							
Age	0.013 (5.65)***	0.013 (5.40)***	0.013 (5.64)***	0.013 (1.74)*	0.021 (3.20)***	0.013 (5.77)***	0.010 (1.54)
Log transfer fee	0.032 (2.33)***	0.031 (2.22)***	0.032 (2.33)***	0.063 (2.11)**	0.322 (11.88)***	0.032 (2.34)***	0.062 (2.08)***
Games started	0.067 (36.09)***	0.068 (37.78)***	0.067 (36.17)***	0.051 (18.30)***	0.087 (40.16)***	0.067 (36.08)***	0.051 (18.42)***
Substitute	0.041 (10.93)***	0.041 (10.83)***	0.041 (10.89)***	0.011 (0.89)	0.069 (13.65)***	0.041 (19.09)***	0.011 (0.89)
Defender	1.715 (14.73)***	1.713 (14.79)***	1.714 (14.71)***	1.113 (72.20)***		1.714 (14.70)***	1.119 (7.28)***
Forward	1.397 (11.06)***	1.399 (11.13)***	1.396 (11.05)***	0.720 (40.00)***	1.647 (21.26)***	1.396 (11.01)***	0.726 (4.06)***
Midfield	1.729 (12.67)***	1.728 (12.68)***	1.728 (12.66)***	0.889 (4.45)***	0.679 (11.31)***	1.728 (12.66)***	0.892 (4.50)***
Goalie					−18.216 (54.31)***		
Goals	−0.022 (5.81)***	−0.022 (6.29)***	−0.022 (5.83)***	−0.028 (3.37)***		−0.022 (5.84)***	−0.028 (3.27)***
Soccer leagues							
European Champions league	−0.031 (0.52)	−0.023 (0.38)	−0.036 (0.63)	−0.502 (2.43)**	0.211 (2.45)**	−0.028 (0.46)	−0.453 (2.23)**
league							
French league	0.264 (4.45)***	0.266 (4.27)***	0.259 (4.21)***	0.297 (2.62)***	0.078 (1.24)	0.263 (4.15)***	0.334 (2.93)***
German league	0.313 (6.37)***	0.321 (6.58)***	0.317 (6.58)***	0.098 (0.63)	0.244 (4.12)***	0.319 (6.33)***	0.112 (0.68)
Italian league	0.352 (6.46)***	0.355 (6.27)***	0.337 (5.11)***	0.629 (4.57)***	−0.012 (0.22)	0.353 (6.28)***	0.749 (7.10)***
Spanish league	0.544 (10.99)***	0.548 (10.75)***	0.534 (9.54)***	0.648 (6.15)***	0.002 (0.03)	0.551 (10.09)***	0.719 (6.70)***
Regional fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5035	4965	5035	5035	5035	5033	5033

Notes: The dependent variables are per player-season. Columns (1)–(7) contain the results of negative binomial specifications with disturbance terms clustered at the country level. The omitted categories in columns (1)–(4) and (6)–(7) are Goalie (for field position), OECD (for region), and the English Premier league (for league); in column (5), the baseline categories are defender (for field position), OECD (for region), and the English Premier league (for league). The region fixed effect results are not shown. Z-statistics are in parentheses. Statistical significance at ***90%, **95%, and *99% confidence levels.

7 APPENDIX

This section details the replication of the tables in the appendix.

7.0.1 Appendix Tables

Table A1: Countries and Players Represented in the Main Sample

Country	Observations	Yellow cards	Civil war years	Country	Observations	Yellow cards	Civil war years
Albania (ALB)	18	2.88	0	Macedonia (MKD)	6	4.16	1
Algeria (DZA)	6	1.50	15	Mali (MLI)	29	3.03	2
Argentina (ARG)	178	2.91	0	Mexico (MEX)	8	3.62	2
Australia (AUS)	28	2.57	0	Morocco (MAR)	26	3.15	10
Austria (AUT)	6	1.66	0	Netherlands (NLD)	118	2.06	0
Belarus (BLR)	6	1.50	0	Nigeria (NGA)	43	1.81	1
Belgium (BEL)	34	1.91	0	Northern Ireland (NIR)	12	1.00	13
Bosnia and Herzegovina (BIH)	14	2.92	4	Norway (NOR)	20	1.75	0
Brazil (BRA)	277	2.44	0	Paraguay (PRY)	14	2.42	1
Bulgaria (BGR)	20	2.55	0	Peru (PER)	13	1.38	19
Cameroon (CMR)	52	2.28	1	Poland (POL)	30	1.00	0
Canada (CAN)	7	3.71	0	Portugal (PRT)	68	3.02	0
Chile (CHL)	10	3.80	0	Romania (ROM)	19	1.21	1
Colombia (COL)	19	4.79	26	Russia (RUS)	8	1.75	13
Congo DR (ZAR)	10	2.50	6	Scotland (SCO)	37	2.16	13
Croatia (HRV)	48	2.37	3	Senegal (SEN)	59	2.25	10
Czech Republic (CZE)	67	2.24	0	Serbia (SRB)	8	1.75	3

Table A1 continued from previous page

Country	Observations	Yellow cards	Civil war years	Country	Observations	Yellow cards	Civil war years
Denmark (DNK)	58	1.84	0	Serbia and Montenegro (YUG)	48	2.83	3
Egypt (EGY)	6	1.00	6	Sierra Leone	5	2.00	10
England (GBR)	402	2.17	13	Slovak Republic (SVK)	14	0.92	0
Finland (FIN)	24	1.08	0	Slovenia (SVN)	11	1.63	0
France (FRA)	721	2.48	0	South Africa (ZAF)	15	1.06	9
Georgia (GEO)	10	3.20	4	South Korea (KOR)	5	1.00	0
Germany (DEU)	424	2.00	0	Spain (ESP)	742	2.91	5
Ghana (GHA)	25	2.40	2	Sweden (SWE)	35	1.77	0
Greece (GRC)	22	2.13	0	Switzerland (CHE)	49	2.40	0
Guinea (GIN)	15	2.33	2	Togo (TGO)	8	0.75	2
Hungary (HUN)	10	0.90	0	Trinidad and Tobago (TTO)	5	0.20	1
Iceland (ISL)	8	2.00	0	Tunisia (TUN)	21	2.33	1
Iran (IRN)	9	2.33	19	Turkey (TUR)	24	2.25	22
Ireland (IRL)	67	1.89	0	Ukraine (UKR)	9	1.44	0
Israel (ISR)	5	4.80	26	United States (USA)	30	0.96	4
Italy (ITA)	730	2.81	0	Uruguay (URY)	66	2.89	0
Ivory Coast (CIV)	49	3.26	3	Wales (WAL)	26	2.19	13
Jamaica (JAM)	9	1.77	0	Japan (JPN)	10	1.50	0

Notes: The source of this data is the ESPN Soccernet website. We include all countries with at least five player-seasons represented in the leagues for which we have data. The "Yellow Cards" column shows the average number of yellow cards per player/season by nationals of each respective country. The "Civil War Years" column shows the number of years of civil war since 1980 in the respective country.

8 ACKNOWLEDGMENTS

We extend our sincere gratitude to our advisors Eric Van Dusen, Rohan Jha, and Peter Flo Grinde-Hollevik at the University of California, Berkeley, for their invaluable guidance and support throughout the course of Econ 148: Data Science for Economics. This course provided us with a rare and insightful opportunity to engage in the comprehensive process of replicating an academic paper from its foundational stages.

Our experience has been greatly enriched by the collaborative learning environment and the application of rigorous data science methodologies to the field of economics. We also wish to thank our peers in the class for their constructive feedback and stimulating discussions, which were instrumental in the successful completion of this project.

Finally, we take full responsibility for any errors in this replication. The process of critical analysis and the pursuit of accuracy have been a paramount part of our learning experience, and we welcome any further discussion and review of our work.

```
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt
import numpy as np
import statsmodels.formula.api as smf


# from stata2python import stata2python


# stata2python("corr points assists rebounds", "nba")


df = pd.read_stata('soccer_data.dta')
df.head()
```

	player_id	player_name	war_before	war_after	year	team	nationality	p
0	2726	Nelson de Jesus Dida	0.0	0.0	2004/05 Statistics	AC Milan	Brazil	
1	2726	Nelson de Jesus Dida	0.0	0.0	2004/05 Statistics	AC Milan	Brazil	
2	2741	Juliano Belletti	0.0	0.0	2004/05 Statistics	Barcelona	Brazil	
3	2741	Juliano Belletti	0.0	0.0	2004/05 Statistics	Barcelona	Brazil	
4	2749	Cris	0.0	0.0	2004/05 Statistics	Lyon	Brazil	

```
df.describe()
```

TABLE 1

```
import pandas as pd
import statsmodels.api as sm

df = pd.read_stata('soccer_data.dta')

x_region = ["africa", "asia", "lac", "east_europe"]

df.dropna(subset=['ln_contract'], inplace=True)

filter_condition = (
    (df['italian'] == 1) | (df['champions'] == 1) |
    (df['english'] == 1) | (df['french'] == 1) |
    (df['german'] == 1) | (df['spanish'] == 1)
) & (df['num_country'] >= 5)

filtered_df = df[filter_condition]

summary_stats = filtered_df[['yellow_card', 'red_card',
                             'civwar', 'r_law', 'income',
                             'age', 'weekly_wage', 'contract',
                             'games_start', 'games_sub',
                             'goalie', 'defender', 'forward', 'midfield',
                             'goals'] + x_region + ['oecd', 'english', 'champions', 'french', 'german', 'italian', 'spanish']].c

summary_stats.head()
```

	yellow_card	red_card	civwar	r_law	income	age	wee
count	5035.000000	5035.000000	5035.000000	5035.000000	4965.000000	5035.000000	5(
mean	2.434161	0.157498	2.741609	0.849827	26203.863041	25.992850	23
std	2.734036	0.416225	4.742952	0.886973	10923.413993	4.404231	27(
min	0.000000	0.000000	0.000000	-1.760000	720.000000	17.000000	
25%	0.000000	0.000000	0.000000	0.510000	21470.000000	23.000000	4

5 rows × 26 columns

TABLE 2

```
df = pd.read_stata('soccer_data.dta')

rename_dict = {
    'civwar': 'Years of civil war',
    'ln_income': 'Log GNI per capita',
    'r_law': 'Rule of Law',
    'age': 'Age',
    'ln_contract': 'Log transfer fee',
    'games_start': 'Games Started',
    'games_sub': 'Substitute',
    'defender': 'Defender',
    'forward': 'Forward',
    'midfield': 'Midfield',
    'goalie': 'Goalie',
    'goals': 'Goals',
    'champions': 'European Champions League',
    'french': 'French League',
    'german': 'German League',
    'italian': 'Italian League',
    'spanish': 'Spanish League',
    'english': 'English League',
    'war_after': 'Civil war years post-birth',
    'war_before': 'Civil war years pre-birth'
}

df.rename(columns=rename_dict, inplace=True)


league_columns = ['Italian League', 'European Champions League', 'English League', 'French League', 'German League', 'Spanish Le
df['league_membership'] = df[league_columns].sum(axis=1) >= 1
filtered_df = df[(df['league_membership']) & (df['num_country'] >= 5)]


formulas = {
    "column1": 'yellow_card ~ Q("Years of civil war") + Age + Q("Log transfer fee") + Q("Games Started") + Substitute + Defender
           'Goals + Q("Italian League") + Q("European Champions League") + Q("French League") + Q("German League") + Q("Spanish
           'africa + asia + lac + east_europe',
    "column2": 'yellow_card ~ Q("Years of civil war") + Q("Log GNI per capita") + Age + Q("Games Started") + Substitute + Defend
           'Goals + Q("Log transfer fee") + Q("Italian League") + Q("European Champions League") + Q("French League") + Q("Germa
           'africa + asia + lac + east_europe',
    "column3": 'yellow_card ~ Q("Years of civil war") +  Q("Rule of Law") + Age  + Q("Log transfer fee") + Q("Games Started") +
           'Goals + Q("Italian League") + Q("European Champions League") + Q("French League") + Q("German League") + Q("Spanish
           'africa + asia + lac + east_europe',
     "column4": 'red_card ~ Q("Years of civil war") +  Q("Rule of Law") + Age + Q("Log transfer fee") + Q("Games Started") + Sub
           'Goals + Q("Italian League") + Q("European Champions League") + Q("French League") + Q("German League") + Q("Spanish
           'africa + asia + lac + east_europe',
    "column5": 'Goals ~ Q("Years of civil war") +  Q("Rule of Law") + Age + Q("Log transfer fee") + Q("Games Started") + Substit
           'Goalie + Q("Italian League") + Q("European Champions League") + Q("French League") + Q("German League") + Q("Spanish
           'africa + asia + lac + east_europe',
    "column6": 'yellow_card ~ Q("Civil war years pre-birth") +  Q("Civil war years post-birth") + Age + Q("Log transfer fee") +
           'Goals + Q("Italian League") + Q("European Champions League") + Q("French League") + Q("German League") + Q("Spanish
           'africa + asia + lac + east_europe',
    "column6": 'red_card ~ Q("Civil war years pre-birth") +  Q("Civil war years post-birth") + Age + Q("Log transfer fee") + Q("
           'Goals + Q("Italian League") + Q("European Champions League") + Q("French League") + Q("German League") + Q("Spanish
           'africa + asia + lac + east_europe'
}
```

```
models = {}
for column, formula in formulas.items():
    model = smf.ols(formula, data=filtered_df).fit(cov_type='cluster', cov_kwds={'groups': filtered_df['nation']})
    models[column] = model
    print('')
    print(f"Results for {column}:")
    print(model.summary())
```

```
africa                          0.0200      0.247      0.081      0.935      -0.464       0.504
asia                           -0.6938      0.476     -1.457      0.145      -1.627       0.239
lac                             0.0463      0.278      0.167      0.868      -0.498       0.591
east_europe                     0.0848      0.213      0.398      0.691      -0.333       0.503
==============================================================================
Omnibus:                     2946.248   Durbin-Watson:                   1.883
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            46407.346
Skew:                           2.537   Prob(JB):                         0.00
Kurtosis:                      17.095   Cond. No.                          627.
==============================================================================

Notes:
[1] Standard Errors are robust to cluster correlation (cluster)

Results for column6:
                            OLS Regression Results
==============================================================================
Dep. Variable:               red_card   R-squared:                       0.095
Model:                           OLS   Adj. R-squared:                  0.092
Method:                Least Squares   F-statistic:                     88.96
Date:               Fri, 26 Apr 2024   Prob (F-statistic):           5.02e-40
Time:                       10:40:49   Log-Likelihood:                -2440.7
No. Observations:               4963   AIC:                             4921.
Df Residuals:                   4943   BIC:                             5052.
Df Model:                         19
Covariance Type:              cluster
================================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------------
Intercept                       -0.2319      0.054     -4.267      0.000      -0.338      -0.125
Q("Civil war years pre-birth")  -0.0003      0.001     -0.305      0.761      -0.003       0.002
Q("Civil war years post-birth")  0.0020      0.001      2.283      0.022       0.000       0.004
Age                              0.0006      0.001      0.752      0.452      -0.001       0.002
Q("Log transfer fee")            0.0056      0.004      1.561      0.119      -0.001       0.013
Q("Games Started")               0.0091      0.001      7.955      0.000       0.007       0.011
Substitute                      -0.0009      0.002     -0.468      0.640      -0.004       0.003
Defender                         0.1459      0.014     10.294      0.000       0.118       0.174
Forward                          0.0989      0.017      5.775      0.000       0.065       0.132
Midfield                         0.1074      0.019      5.532      0.000       0.069       0.145
Goals                           -0.0073      0.002     -4.265      0.000      -0.011      -0.004
Q("Italian League")              0.1373      0.016      8.476      0.000       0.106       0.169
Q("European Champions League")   0.0098      0.015      0.665      0.506      -0.019       0.039
Q("French League")               0.0438      0.017      2.618      0.009       0.011       0.077
Q("German League")               0.0183      0.018      1.033      0.301      -0.016       0.053
Q("Spanish League")              0.1284      0.019      6.871      0.000       0.092       0.165
africa                           0.0176      0.016      1.073      0.283      -0.015       0.050
asia                            -0.0576      0.042     -1.384      0.166      -0.139       0.024
lac                              0.0286      0.012      2.327      0.020       0.005       0.053
east_europe                      0.0046      0.025      0.182      0.856      -0.044       0.054
==============================================================================
Omnibus:                     2518.233   Durbin-Watson:                   1.964
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            15068.902
Skew:                           2.427   Prob(JB):                         0.00
Kurtosis:                      10.022   Cond. No.                          639.
==============================================================================

Notes:
[1] Standard Errors are robust to cluster correlation (cluster)
```

```
wb_codes = pd.read_csv('Wb_codes.csv')
wb_codes.head()
```

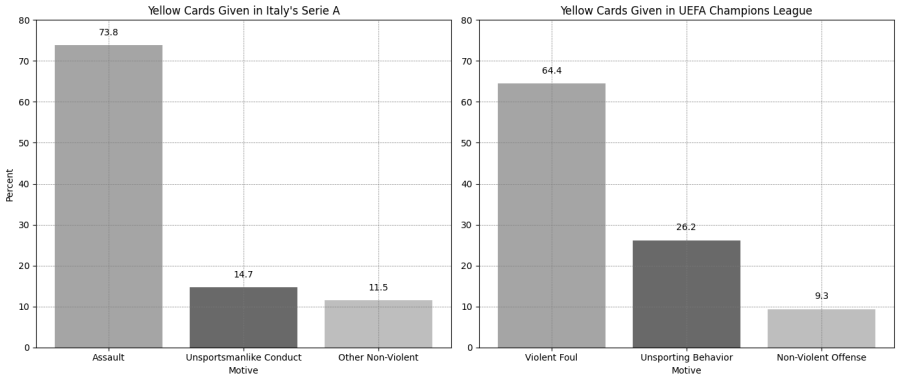
	id	wb_code
0	1	ALB
1	2	DZA
2	3	AGO
3	4	ARG
4	5	ARM

```
import pandas as pd

fig_1 = {
    'xf': ['Serie A TIM', 'Serie A TIM', 'Serie A TIM', 'Serie A TIM'],
    'cause': ['Assault', 'Unsportsmanlike Conduct', 'Other Non-Violent', 'Total'],
    '2005-2006': [1299, 207, 174, 1680],
    '2006-2007': [1357, 320, 235, 1912],
    '2007-2008': [1413, 281, 225, 1919],
    'Average': [1356.333333, 269.3333333, 211.3333333, 1837],
    'Percent': [73.83414999, 14.66158592, 11.5042642, 100]
}

fig_1_data = pd.DataFrame(fig_1)



fig_2 = {
    'cause': ['Violent Foul', 'Unsporting Behavior', 'Non-Violent Offense', 'Grand Total'],
    '2004-2005': [446, 177, 56, 679],
    '2005-2006': [464, 193, 76, 733],
    'Sum': [910, 370, 132, 1412],
    'Percent': [64.44759207, 26.20396601, 9.348441926, 100]
}

fig_2_data = pd.DataFrame(fig_2)
fig_2_data
```

	cause	2004-2005	2005-2006	Sum	Percent
0	Violent Foul	446	464	910	64.447592
1	Unsporting Behavior	177	193	370	26.203966
2	Non-Violent Offense	56	76	132	9.348442
3	Grand Total	679	733	1412	100.000000

Figure 1A/1B

```
categories_1 = fig_1_data['cause'][:-1]
percentages_1 = fig_1_data['Percent'][:-1]

categories_2 = fig_2_data['cause'][:-1]
percentages_2 = fig_2_data['Percent'][:-1]

colors = ['darkgrey', 'dimgray', 'silver']
patterns = ['', '', '']

fig, axes = plt.subplots(1, 2, figsize=(14, 6), tight_layout=True)

bars1 = axes[0].bar(categories_1, percentages_1, color=colors)

axes[0].set_title("Yellow Cards Given in Italy's Serie A")
axes[0].set_xlabel("Motive")
axes[0].set_ylabel("Percent")
axes[0].set_ylim(0, 80)
axes[0].grid(color='gray', linestyle='--', linewidth=0.5)

for bar in bars1:
    yval = bar.get_height()
    axes[0].text(bar.get_x() + bar.get_width()/2, yval + 2, round(yval, 1),
                 ha='center', va='bottom')

bars2 = axes[1].bar(categories_2, percentages_2, color=colors)

axes[1].set_title("Yellow Cards Given in UEFA Champions League")
axes[1].set_xlabel("Motive")
axes[1].set_ylim(0, 80)
axes[1].grid(color='gray', linestyle='--', linewidth=0.5)

for bar in bars2:
    yval = bar.get_height()
    axes[1].text(bar.get_x() + bar.get_width()/2, yval + 2, round(yval, 1),
                 ha='center', va='bottom')

plt.show()
```



Note: These are our initial attempts to replicate figure 2/3/4. Please see the next section for corrected visualizations.

Figure 2

```
fig_2_df = pd.read_stata('soccer_data.dta')

fig_2_df = df.dropna(subset=["contract"])

filter_vars = ['italian', 'champions', 'english', 'french', 'german', 'spanish']
fig_2_df = fig_2_df[fig_2_df[filter_vars].any(axis=1)]

fig_2_df = fig_2_df[fig_2_df["num_country"] >= 5]

wb_codes = pd.read_csv("Wb_codes.csv")
wb_codes = wb_codes.rename(columns={'id': 'nation'})
fig_2_df = fig_2_df.merge(wb_codes, on='nation')

indep_vars = ["age", "games_start", "games_sub", "defender", "goalie", "forward", "midfield",
              "goals", "ln_contract", "italian", "champions", "french", "german",
              "spanish", "africa", "asia", "lac", "east_europe"]
fig_2_X = sm.add_constant(fig_2_df[indep_vars])

yellow_card_y = fig_2_df["yellow_card"]
yellow_card_model = sm.OLS(yellow_card_y, fig_2_X).fit()
yellow_hat = yellow_card_model.predict(fig_2_X)
yellow_res = fig_2_df["yellow_card"] - yellow_hat
fig_2_df["yellow_res"] = yellow_res

civ_war_y = fig_2_df["civwar"]
civ_war_model = sm.OLS(civ_war_y, fig_2_X).fit()
war_hat = civ_war_model.predict(fig_2_X)
war_res = fig_2_df["civwar"] - war_hat
fig_2_df["war_res"] = war_res

fig_2_df = fig_2_df[["yellow_res", "war_res", "wb_code", "num_country"]].groupby("wb_code").mean().reset_index()
yellow_res = fig_2_df["yellow_res"]
war_res = fig_2_df["war_res"]
num_country = fig_2_df["num_country"]

plt.scatter(war_res, yellow_res, facecolors='none', edgecolors='b', alpha=0.5, s=num_country * 3)
plt.plot(np.unique(war_res), np.poly1d(np.polyfit(war_res, yellow_res, 1))(np.unique(war_res)), color='r')

plt.xlabel('Years of Civil War since 1980 (Residuals)')
plt.ylabel('Average Yellow Cards Per Player-Season (Residuals)')

plt.xlim(-10, 20)
plt.xticks(range(-10, 20, 4))
plt.ylim(-1, 1)

for i, wb in enumerate(fig_2_df["wb_code"]):
  plt.annotate(wb, (war_res[i], yellow_res[i]))

plt.show()
```

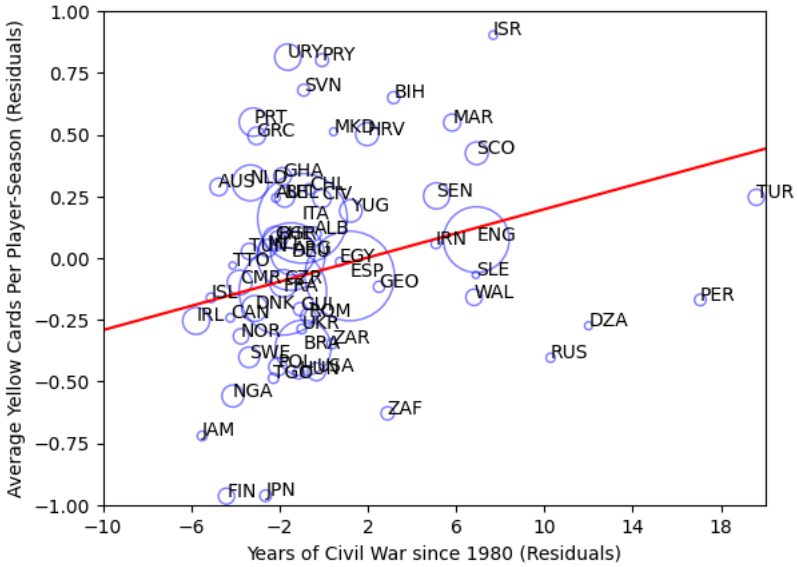

Average Yellow Cards Per Player-Season (Residuals)
Years of Civil War since 1980 (Residuals)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−10
−6
−2
2
6
10
14
18
ISR
URY
PRY
SVN
BIH
MAR
PRT
GRC
MKD
HRV
SCO
AUS
NLD
GHA
CHL
SEN
TUR
YUG
ITA
ALB
IRN
ENG
TTO
EGY
ESP
SLE
GEO
WAL
PER
ISL
CMR
DNK
IRL
CAN
DZA
NOR
UKR
ZAR
BRA
RUS
SWE
NGA
ZAF
JAM
FIN
JPN

Figure 3

```
fig_3_df = pd.read_stata('soccer_data.dta')

fig_3_df = df.dropna(subset=["contract"])

filter_vars = ['italian', 'champions', 'english', 'french', 'german', 'spanish']
fig_3_df = fig_3_df[fig_3_df[filter_vars].any(axis=1)]

fig_3_df = fig_3_df[fig_3_df["oecd"] == 0]

fig_3_df = fig_3_df[fig_3_df["num_country"] >= 5]

wb_codes = pd.read_csv("Wb_codes.csv")
wb_codes = wb_codes.rename(columns={'id': 'nation'})
fig_3_df = fig_3_df.merge(wb_codes, on='nation')

indep_vars = ["age", "games_start", "games_sub", "defender", "forward", "midfield",
              "goals", "ln_contract", "italian", "champions", "french", "german",
              "spanish", "africa", "asia", "lac", "east_europe"]
fig_3_X = sm.add_constant(fig_3_df[indep_vars])

yellow_card_y = fig_3_df["yellow_card"]
yellow_card_model = sm.OLS(yellow_card_y, fig_3_X).fit()
yellow_hat = yellow_card_model.predict(fig_3_X)
yellow_res = fig_3_df["yellow_card"] - yellow_hat
fig_3_df["yellow_res"] = yellow_res

civ_war_y = fig_3_df["civwar"]
civ_war_model = sm.OLS(civ_war_y, fig_3_X).fit()
war_hat = civ_war_model.predict(fig_3_X)
war_res = fig_3_df["civwar"] - war_hat
fig_3_df["war_res"] = war_res

fig_3_df = fig_3_df[["yellow_res", "war_res", "wb_code", "num_country"]].groupby("wb_code").mean().reset_index()
yellow_res = fig_3_df["yellow_res"]
war_res = fig_3_df["war_res"]
num_country = fig_3_df["num_country"]

plt.scatter(war_res, yellow_res, facecolors='none', edgecolors='b', alpha=0.5, s=num_country * 3)
plt.plot(np.unique(war_res), np.poly1d(np.polyfit(war_res, yellow_res, 1))(np.unique(war_res)), color='r')

plt.xlabel('Years of Civil War since 1980 (Residuals)')
plt.ylabel('Average Yellow Cards Per Player-Season (Residuals)')

plt.xlim(-10, 18)
plt.xticks(range(-10, 20, 4))
plt.ylim(-1, 1)

for i, wb in enumerate(fig_3_df["wb_code"]):
  plt.annotate(wb, (war_res[i], yellow_res[i]))

plt.show()
```

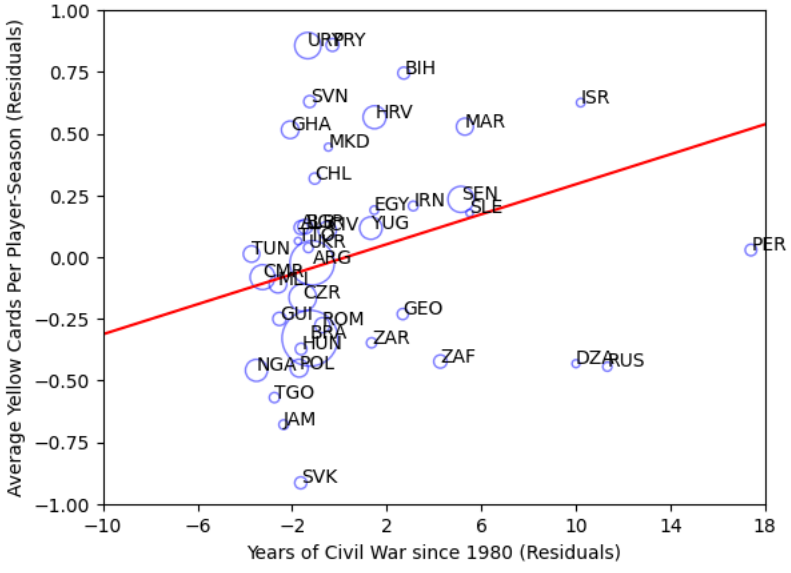
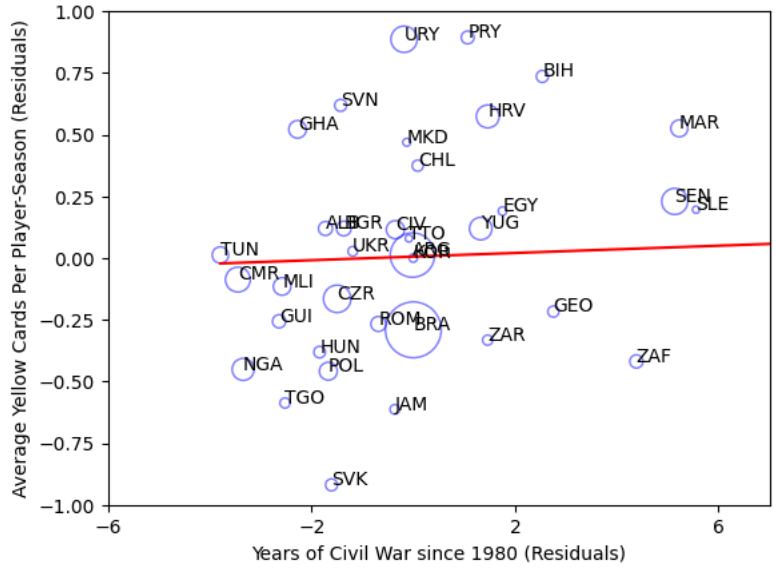


Figure 4 - Aneesh

```
fig_4_df = pd.read_stata('soccer_data.dta')

fig_4_df = df.dropna(subset=["contract"])

filter_vars = ['italian', 'champions', 'english', 'french', 'german', 'spanish']
fig_4_df = fig_4_df[fig_4_df[filter_vars].any(axis=1)]

fig_4_df = fig_4_df[fig_4_df["oecd"] == 0]

fig_4_df = fig_4_df[~fig_4_df["nation"].isin([23, 52, 54, 78, 101])]

fig_4_df = fig_4_df[fig_4_df["num_country"] >= 5]

wb_codes = pd.read_csv("Wb_codes.csv")
wb_codes = wb_codes.rename(columns={'id': 'nation'})
fig_4_df = fig_4_df.merge(wb_codes, on='nation')

indep_vars = ["age", "games_start", "games_sub", "defender", "forward", "midfield",
              "goals", "ln_contract", "italian", "champions", "french", "german",
              "spanish", "africa", "asia", "lac", "east_europe"]
fig_4_X = sm.add_constant(fig_4_df[indep_vars])

yellow_card_y = fig_4_df["yellow_card"]
yellow_card_model = sm.OLS(yellow_card_y, fig_4_X).fit()
yellow_hat = yellow_card_model.predict(fig_4_X)
yellow_res = fig_4_df["yellow_card"] - yellow_hat
fig_4_df["yellow_res"] = yellow_res

civ_war_y = fig_4_df["civwar"]
civ_war_model = sm.OLS(civ_war_y, fig_4_X).fit()
war_hat = civ_war_model.predict(fig_4_X)
war_res = fig_4_df["civwar"] - war_hat
fig_4_df["war_res"] = war_res

fig_4_df = fig_4_df[["yellow_res", "war_res", "wb_code", "num_country"]].groupby("wb_code").mean().reset_index()
yellow_res = fig_4_df["yellow_res"]
war_res = fig_4_df["war_res"]
num_country = fig_4_df["num_country"]

plt.scatter(war_res, yellow_res, facecolors='none', edgecolors='b', alpha=0.5, s=num_country * 3)
plt.plot(np.unique(war_res), np.poly1d(np.polyfit(war_res, yellow_res, 1))(np.unique(war_res)), color='r')

plt.xlabel('Years of Civil War since 1980 (Residuals)')
plt.ylabel('Average Yellow Cards Per Player-Season (Residuals)')

plt.xlim(-6, 7)
plt.xticks(range(-6, 8, 4))
plt.ylim(-1, 1)

for i, wb in enumerate(fig_4_df["wb_code"]):
  plt.annotate(wb, (war_res[i], yellow_res[i]))

plt.show()
```



Figures below are all corrected!

Revised Figure 2

```
fig_2_df = pd.read_stata('soccer_data.dta')
fig_2_df.drop('num_country', axis=1, inplace=True)
fig_2_df = df.dropna(subset=["contract"])
conditions = (fig_2_df['italian'] == 1) | (fig_2_df['champions'] == 1) | (fig_2_df['english'] == 1) | (fig_2_df['french'] == 1)
fig_2_df['num_country'] = fig_2_df[conditions].groupby('nationality')['player_id'].transform('count')

wb_codes = pd.read_csv("Wb_codes.csv")
wb_codes = wb_codes.rename(columns={'id': 'nation'})

fig_2_df = fig_2_df.merge(wb_codes, on='nation')
fig_2_df
```

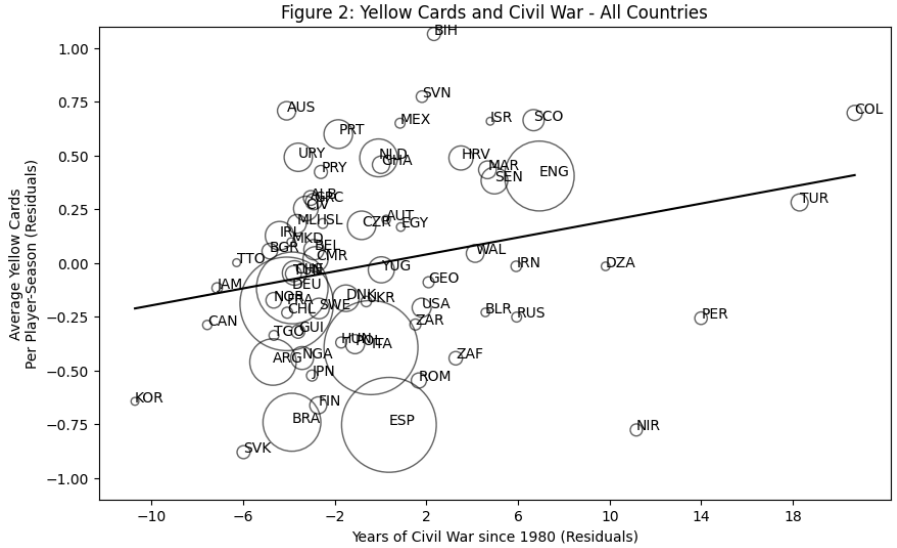
```
<ipython-input-35-81314a86954e>:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stab
  fig_2_df['num_country'] = fig_2_df[conditions].groupby('nationality')['player_
```

	player_id	player_name	war_before	war_after	year	team	nationality
0	2726	Nelson de Jesus Dida	0.0	0.0	2004/05 Statistics	AC Milan	Brazil
1	2726	Nelson de Jesus Dida	0.0	0.0	2004/05 Statistics	AC Milan	Brazil
2	2741	Juliano Belletti	0.0	0.0	2004/05 Statistics	Barcelona	Brazil
3	2741	Juliano Belletti	0.0	0.0	2004/05 Statistics	Barcelona	Brazil
4	2749	Cris	0.0	0.0	2004/05 Statistics	Lyon	Brazil
...	...	...	...	...	...	...	...
5073	68202	moumouni odagano	0.0	1.0	2005/06 Statistics	Sochaux	Burkina Faso
5074	12061	Juan Manuel Peña	3.0	0.0	2005/06 Statistics	Villarreal	Bolivia
5075	12061	Juan Manuel Peña	3.0	0.0	2005/06 Statistics	Villarreal	Bolivia
5076	27273	Chiguy Lucau	8.0	7.0	2005/06 Statistics	Le Mans	Democratic Republic of Congo
5077	69096	Jean jacque Pierre	0.0	3.0	2005/06 Statistics	Nantes	Haiti

5078 rows × 41 columns

```
conditions = (
    (fig_2_df['italian'] == 1) |
    (fig_2_df['champions'] == 1) |
    (fig_2_df['english'] == 1) |
    (fig_2_df['french'] == 1) |
    (fig_2_df['german'] == 1) |
    (fig_2_df['spanish'] == 1)
) & (fig_2_df['num_country'] >= 5)

filtered_df = fig_2_df[conditions]


aggregated_df = filtered_df.groupby('wb_code').agg({
    'yellow_card': 'mean',
    'civwar': 'mean',
    'nation': 'mean',
    'num_country': 'count',
    'age': 'mean',
    'games_start': 'mean',
    'games_sub': 'mean',
    'goalie': 'mean',
    'defender': 'mean',
    'forward': 'mean',
    'midfield': 'mean',
    'goals': 'mean',
    'ln_contract': 'mean',
    'italian': 'mean',
    'champions': 'mean',
    'english': 'mean',
    'french': 'mean',
    'german': 'mean',
    'spanish': 'mean',
    'africa': 'mean',
    'asia': 'mean',
    'lac': 'mean',
    'east_europe': 'mean',
    'oecd': 'mean'
}).reset_index()

fig_2_df = aggregated_df


X = fig_2_df[['age', 'games_start', 'games_sub', 'defender', 'forward', 'midfield', 'goals', 'ln_contract', 'italian', 'champion
X = sm.add_constant(X)

y = fig_2_df['yellow_card']

model_yellow = sm.OLS(y, X).fit()

# print(model_yellow.summary())

fig_2_df['yellowhat'] = model_yellow.predict(X)

fig_2_df['yellow_res'] = fig_2_df['yellow_card'] - fig_2_df['yellowhat']


model_civwar = sm.OLS(fig_2_df['civwar'], X).fit()

# Display the regression results
# print(model_civwar.summary())

# Predict values for civwar
fig_2_df['warhat'] = model_civwar.predict(X)

fig_2_df['war_res'] = fig_2_df['civwar'] - fig_2_df['warhat']
```

```
x = fig_2_df['yellow_res']
y = fig_2_df['war_res']


# Create the figure and axis
fig, ax = plt.subplots(figsize=(10, 6))

# Scatter plot of residuals
scatter = ax.scatter(x=y, y=x, s=fig_2_df['num_country']*6, facecolors='none', edgecolors='black', alpha=0.6, label='Residuals')



for i, wb in enumerate(fig_2_df["wb_code"]):
  plt.annotate(wb, (fig_2_df['war_res'][i], fig_2_df['yellow_res'][i]))


# Setting labels and titles
plt.plot(np.unique(fig_2_df['war_res']), np.poly1d(np.polyfit(fig_2_df['war_res'], fig_2_df['yellow_res'], 1))(np.unique(fig_2_c

plt.ylim(-1.1, 1.1)
plt.xticks(range(-10, 19, 4))

ax.set_ylabel('Average Yellow Cards\nPer Player-Season (Residuals)')
ax.set_xlabel('Years of Civil War since 1980 (Residuals)')
ax.set_title('Figure 2: Yellow Cards and Civil War - All Countries')

plt.show()
```

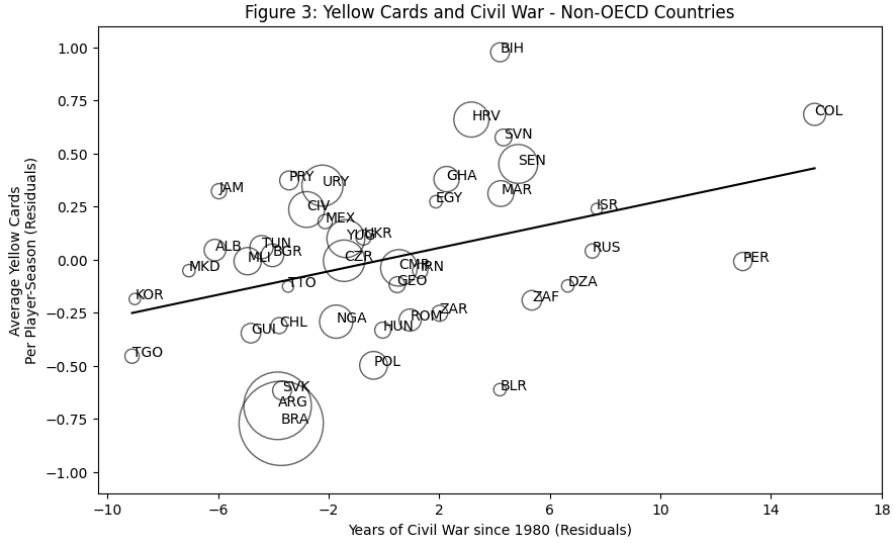


Revised Figure 3

```
fig_3_df = pd.read_stata('soccer_data.dta')
fig_3_df.drop('num_country', axis=1, inplace=True)
fig_3_df = df.dropna(subset=["contract"])
conditions = ((fig_3_df['italian'] == 1) | (fig_3_df['champions'] == 1) | (fig_3_df['english'] == 1) | (fig_3_df['french'] == 1)
fig_3_df['num_country'] = fig_3_df[conditions].groupby('nationality')['player_id'].transform('count')

wb_codes = pd.read_csv("Wb_codes.csv")
wb_codes = wb_codes.rename(columns={'id': 'nation'})

fig_3_df = fig_3_df.merge(wb_codes, on='nation')
```

```
conditions = (
    (fig_3_df['italian'] == 1) |
    (fig_3_df['champions'] == 1) |
    (fig_3_df['english'] == 1) |
    (fig_3_df['french'] == 1) |
    (fig_3_df['german'] == 1) |
    (fig_3_df['spanish'] == 1)
) & (fig_3_df['num_country'] >= 5) & (fig_3_df['oecd'] == 0)

filtered_df = fig_3_df[conditions]


aggregated_df = filtered_df.groupby('wb_code').agg({
    'yellow_card': 'mean',
    'civwar': 'mean',
    'nation': 'mean',
    'num_country': 'count',
    'age': 'mean',
    'games_start': 'mean',
    'games_sub': 'mean',
    'goalie': 'mean',
    'defender': 'mean',
    'forward': 'mean',
    'midfield': 'mean',
    'goals': 'mean',
    'ln_contract': 'mean',
    'italian': 'mean',
    'champions': 'mean',
    'english': 'mean',
    'french': 'mean',
    'german': 'mean',
    'spanish': 'mean',
    'africa': 'mean',
    'asia': 'mean',
    'lac': 'mean',
    'east_europe': 'mean',
    'oecd': 'mean'
}).reset_index()

fig_3_df = aggregated_df


# Set up regressions
X = fig_3_df[['age', 'games_start', 'games_sub', 'defender', 'forward', 'midfield', 'goals', 'ln_contract', 'italian', 'champion
X = sm.add_constant(X)

# Run regession on yellow cards
model_yellow = sm.OLS(fig_3_df['yellow_card'], X).fit()
fig_3_df['yellowhat'] = model_yellow.predict(X)
fig_3_df['yellow_res'] = fig_3_df['yellow_card'] - fig_3_df['yellowhat']

# Run regression on civil war
model_civwar = sm.OLS(fig_3_df['civwar'], X).fit()
fig_3_df['warhat'] = model_civwar.predict(X)
fig_3_df['war_res'] = fig_3_df['civwar'] - fig_3_df['warhat']
```

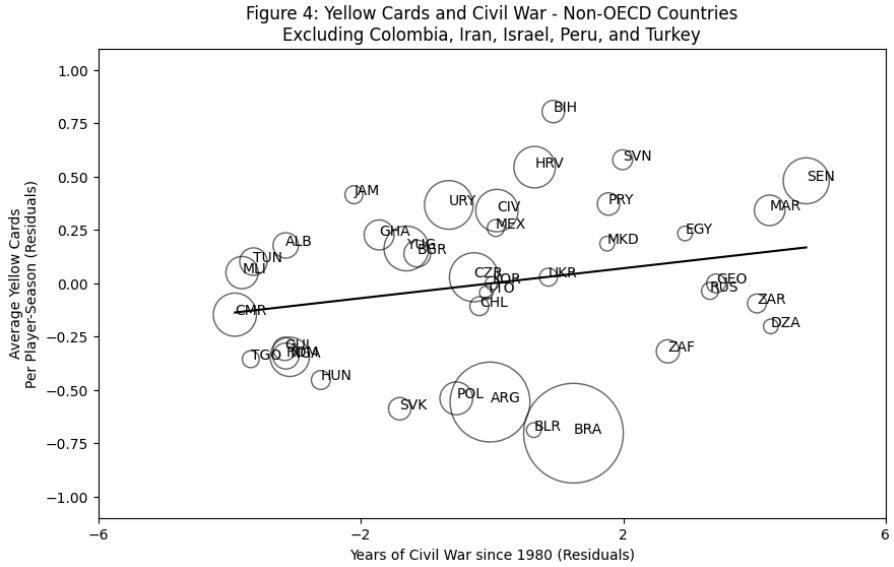
```
# Create the figure and axis
fig, ax = plt.subplots(figsize=(10, 6))

# Scatter plot of residuals
scatter = ax.scatter(x=fig_3_df['war_res'], y=fig_3_df['yellow_res'], s=fig_3_df['num_country']*13, facecolors='none', edgecolor


for i, wb in enumerate(fig_3_df["wb_code"]):
  plt.annotate(wb, (fig_3_df['war_res'][i], fig_3_df['yellow_res'][i]))


# Setting labels and titles
plt.plot(np.unique(fig_3_df['war_res']), np.poly1d(np.polyfit(fig_3_df['war_res'], fig_3_df['yellow_res'], 1))(np.unique(fig_3_c

plt.ylim(-1.1, 1.1)
plt.xticks(range(-10, 19, 4))

ax.set_ylabel('Average Yellow Cards\nPer Player-Season (Residuals)')
ax.set_xlabel('Years of Civil War since 1980 (Residuals)')
ax.set_title('Figure 3: Yellow Cards and Civil War - Non-OECD Countries')

plt.show()
```



Revised Figure 4

```
fig_4_df = pd.read_stata('soccer_data.dta')
fig_4_df.drop('num_country', axis=1, inplace=True)
fig_4_df = df.dropna(subset=["contract"])
conditions = ((fig_4_df['italian'] == 1) | (fig_4_df['champions'] == 1) | (fig_4_df['english'] == 1) | (fig_4_df['french'] == 1)
fig_4_df['num_country'] = fig_4_df[conditions].groupby('nationality')['player_id'].transform('count')

wb_codes = pd.read_csv("Wb_codes.csv")
wb_codes = wb_codes.rename(columns={'id': 'nation'})

fig_4_df = fig_4_df.merge(wb_codes, on='nation')
```

```
conditions = (
    (fig_4_df['italian'] == 1) |
    (fig_4_df['champions'] == 1) |
    (fig_4_df['english'] == 1) |
    (fig_4_df['french'] == 1) |
    (fig_4_df['german'] == 1) |
    (fig_4_df['spanish'] == 1)
) & (fig_4_df['num_country'] >= 5) & (fig_4_df['oecd'] == 0) & (~fig_4_df['nation'].isin([23, 52, 54, 78, 101]))

filtered_df = fig_4_df[conditions]


aggregated_df = filtered_df.groupby('wb_code').agg({
    'yellow_card': 'mean',
    'civwar': 'mean',
    'nation': 'mean',
    'num_country': 'count',
    'age': 'mean',
    'games_start': 'mean',
    'games_sub': 'mean',
    'goalie': 'mean',
    'defender': 'mean',
    'forward': 'mean',
    'midfield': 'mean',
    'goals': 'mean',
    'ln_contract': 'mean',
    'italian': 'mean',
    'champions': 'mean',
    'english': 'mean',
    'french': 'mean',
    'german': 'mean',
    'spanish': 'mean',
    'africa': 'mean',
    'asia': 'mean',
    'lac': 'mean',
    'east_europe': 'mean',
    'oecd': 'mean'
}).reset_index()

fig_4_df = aggregated_df


X = fig_4_df[['age', 'games_start', 'games_sub', 'defender', 'forward', 'midfield', 'goals', 'ln_contract', 'italian', 'champior
X = sm.add_constant(X)

model_yellow = sm.OLS(fig_4_df['yellow_card'], X).fit()
fig_4_df['yellowhat'] = model_yellow.predict(X)
fig_4_df['yellow_res'] = fig_4_df['yellow_card'] - fig_4_df['yellowhat']

model_civwar = sm.OLS(fig_4_df['civwar'], X).fit()
fig_4_df['warhat'] = model_civwar.predict(X)
fig_4_df['war_res'] = fig_4_df['civwar'] - fig_4_df['warhat']


fig, ax = plt.subplots(figsize=(10, 6))

scatter = ax.scatter(x=fig_4_df['war_res'], y=fig_4_df['yellow_res'], s=fig_4_df['num_country']*18, facecolors='none', edgecolor


for i, wb in enumerate(fig_4_df["wb_code"]):
  plt.annotate(wb, (fig_4_df['war_res'][i], fig_4_df['yellow_res'][i]))


plt.plot(np.unique(fig_4_df['war_res']), np.poly1d(np.polyfit(fig_4_df['war_res'], fig_4_df['yellow_res'], 1))(np.unique(fig_4_c

plt.ylim(-1.1, 1.1)
plt.xticks(range(-6, 7, 4))

ax.set_ylabel('Average Yellow Cards\nPer Player-Season (Residuals)')
ax.set_xlabel('Years of Civil War since 1980 (Residuals)')
ax.set_title('Figure 4: Yellow Cards and Civil War - Non-OECD Countries\nExcluding Colombia, Iran, Israel, Peru, and Turkey')

plt.show()
```



Appendix

TABLE1

```
filtered_df.columns
```

```
Index(['player_id', 'player_name', 'war_before', 'war_after', 'year', 'team',
       'nationality', 'position', 'age', 'league', 'games_start', 'games_sub',
       'goals', 'yellow_card', 'red_card', 'nation', 'defender', 'forward',
       'goalie', 'midfield', 'italian', 'champions', 'english', 'french',
       'german', 'spanish', 'r_law', 'civwar', 'africa', 'east_europe', 'oecd',
       'weekly_wage', 'contract', 'asia', 'lac', 'income', 'ln_wage',
       'ln_contract', 'num_country', 'ln_income'],
      dtype='object')
```

```
df = pd.read_stata('soccer_data.dta')

x_region = ["africa", "asia", "lac", "east_europe"]

df.dropna(subset=['ln_contract'], inplace=True)

filter_condition = (
    (df['italian'] == 1) | (df['champions'] == 1) |
    (df['english'] == 1) | (df['french'] == 1) |
    (df['german'] == 1) | (df['spanish'] == 1)
) & (df['num_country'] >= 5)

filtered_df = df[filter_condition]

aggregated_df = filtered_df.groupby('nationality').agg({
    'player_id': 'count',  # This assumes player_id can represent a player-season
    'yellow_card': 'mean',  # Calculate the average yellow cards
    'civwar': lambda x: (x >= 1980).sum(),  # Sum civil war years since 1980
}).rename(columns={
    'player_id': 'Observations',
    'yellow_card': 'Yellow cards',
    'civwar': 'Civil war years'
})

aggregated_df = aggregated_df[aggregated_df['Observations'] >= 5]

aggregated_df.reset_index(inplace=True)

aggregated_df.sort_values(by='nationality', inplace=True)

aggregated_df
```

	nationality	Observations	Yellow cards	Civil war years
0	Albania	18	2.888889	0
1	Algeria	6	1.500000	0
2	Argentina	178	2.915730	0
3	Australia	28	2.571429	0
4	Austria	6	1.666667	0
...	...	...	...	...
65	Turkey	24	2.250000	0
66	Ukraine	9	1.444444	0
67	United States	30	0.966667	0
68	Uruguay	66	2.893939	0
69	Wales	26	2.192308	0

70 rows × 4 columns

TABLE 2

```
df = pd.read_stata('soccer_data.dta')

rename_dict = {
    'civwar': 'Years of civil war',
    'ln_income': 'Log GNI per capita',
    'r_law': 'Rule of Law',
    'age': 'Age',
    'ln_contract': 'Log transfer fee',
    'games_start': 'Games Started',
    'games_sub': 'Substitute',
    'defender': 'Defender',
    'forward': 'Forward',
```

```
league_columns = ['Italian League', 'European Champions League', 'English League', 'French League', 'German League', 'Spanish Le
df['league_membership'] = df[league_columns].sum(axis=1) >= 1
filtered_df = df[(df['league_membership']) & (df['num_country'] >= 5)]
```

```
    'french': 'French League',
```

```
# print(subset['Log Years of civil war'].isna().sum())  # Check for missing values in 'nation'
```

```
    'english': 'English League'
```

```
import statsmodels.formula.api as smf
import numpy as np
import pandas as pd

try:
    subset.replace([np.inf, -np.inf], np.nan, inplace=True)  # Handle infs
    subset.dropna(inplace=True)  # Ensure no NaNs or infs
    model = smf.negativebinomial(formula, data=subset)
    fitted_model = model.fit(maxiter=100)
    print(fitted_model.summary())
except Exception as e:
    print(f"An error occurred: {e}")
```

```
 Optimization terminated successfully.
          Current function value: 1.735109
          Iterations: 50
```