A Glimpse into the English Premier League

Travis Richardson*

April 16, 2019

Abstract

The English Premier League is the most popular league in the entire world, with a TV viewing audience of about 4.7 billion people. A better understanding of such a prominent league would lead yo better coaching decisions, better articles and write-ups, and most importantly (to certain people) better gambling odds [**?**]. In this paper, a statistically significant (R-squared: 0.9188), 33-variable multiple regression model, as well as a statistically significant (R-squared: 0.9237), 34-variable multiple regression model are used to predict goals and wins in a single English Premier League season. Additionally, for better understanding of how the teams in the English Premier League differentiate, a linear regression is ran in order to see the differences in the top teams, the middle of the pack teams, and the bottom dwellers of the past 12 seasons of the English Premier League. Additional visuals are presented for better perspective on important measures such as: correlation between goals and wins and then how long each of the 39 teams in the data set of been in the EPL for the last 12 years. Full results will later be added to the abstract

*Department of Health and Exercise Science, University of Oklahoma. E-mail address: twrichardson@ou.edu

1 Introduction

As mentioned, the English Premier League is one of the most popular leagues in the world with over 4.7 billion views each year and an average of 12 million viewers per game. With the introduction of data analytics in sports, for years now people have been trying to predict scores and outcomes of games using machine learning and multiple regression models. However, in 2010 the famous Paul the Octopus put all models to shame when it predicted a perfect eight out of eight matches correctly. Whether Paul has dumb luck or is a soccer genius, the feeling of knowing less about the beautiful game that an octopus left a sour taste in many analysts mouths. For this reason, the first part of this paper discusses the prediction of goals and wins for single English Premier League seasons. ”Predicting” is being used in a loose sense, however, the games have already been played and the data is set so the predictive models in this research are comparing fit to see how the models match up to the actual outcomes. Having a model with high correlation and fit will allow for future research to lead into predictive modeling in order to truly predict future seasons goal and win totals. Predictive modeling can lead to better coaching tactics, proper journalism, and more accurate betting odds. Coaches can benefit from predictive modeling by having an understanding of their team and looking into why they are predicted to only score a certain amount of goals or only win a certain amount of games. What is the team lacking or what do they need to focus on that better teams are doing, are questions managers could look at. Every sports fan gets excited for the predictions for each season, typically to be let down when their team finishes in 4th place (like always). Predictive modeling can help journalism in ways as to have proper and accurate assumptions leading to fans trusting that specific news source. The journal with the most accurate predictions would quickly become a fan favorite. This leads to predictive modeling providing more accurate betting odds. If a company giving out betting odds has extremely accurate goal and outcome predictions for single games or seasons, they would end up having to pay out less money. Vice versa, with better prediction modeling, gamblers will feel as they have better chances of winning big.

The next part of this paper will discuss the differentiation between top teams, middle of the pack teams, and lower tier English Premier League teams. In this research, top tier teams are defined as any team that has been in the English Premier League for at least 10 of the last 12 seasons (12). Middle of the pack teams are defined as being in between 9 and 4 seasons (15). Then lower tier teams are defined as having from 3 down to only 1 year in the Premier League (12). In European sports, quite the opposite of American sports, there are actually consequences to performing poorly during the season and rewards for performing well. The bottom three teams each year get relegated to a lower division and the top four teams qualify for the most prestigious tournament in the world, the UEFA Champions League. Being relegated or in the champions league determines a very big difference in income for a team [**?**]. Understanding the differences in these three categories of teams could allow for teams to realize possible stats and play types they are not performing properly and be able to adapt and become a top tier team.

2 Literature Review

Although no analyst will ever be able to compare to the outstanding predictions of Paul the Octopus, there are still many predictive models and research papers done to try and perfect predicting scores and outcomes of games. Three known articles about predicting match results in the English Premier League and the World cup that will be used in this paper are from the works of Ben Ulmer Matthew Fernandez, Francisco Louzada, Adriano K. Suzuki Luis E. B. Salasar, and Andreas Groll, Gunther Schauberger Gerhard Tutz. Ulmer and Fernandez predict the results of English Premier League games using artificial intelligence and machine learning algorithms. They look into historical data and then the current teams form while using five different classifiers to predict win, loss or draw. However, from those five classifiers, their best error rates were only .48 and .50. This article emphasizes the importance of predictive modeling for gambling and coaching improvements [**?**].

In the next paper being examined, Louzda, Suzuki, and Salasar predict outcomes of the English Premier League as well, but with a focus on predicting who the champion will be. They are also interested in looking into ”the end result of a match, the chance of a team to be qualified for a specific tournament, the chance of being relegated, the best attack, the best defense,” [**?**]. To predict these outcomes, the authors are estimating average goals scored by assuming goals scored by each team follows a univariate Poisson distribution. The estimation of goals in Louzda’s paper correlates with the prediction of goals for this paper at hand, however, the model in this paper uses 34 variables, while Louzda only uses 5 covariates; which could make a difference.

The final paper being examined in regard to prediction of matches, by Groll, Schauberger, and Tutz looks into predicting World Cup 2014 matches using Poisson regression. Although not about the English Premier League, the authors try to predict World Cup success based on previous World Cup performances, which is similar to predictions made in this paper. The final predictions in both the models after being fitted and investigated favored Germany as the winner; which this ended up becoming true [**?**].

The other two articles being examined are from the works of Joel Oberstone and Jaime Araya Paul Larkin. Joel Oberstone’s article ”Differentiating the Top English Premier League Football Clubs from the Rest of the Pack: Identifying the Keys to Success” was a large foundation of why this paper came about. The questions of what distinguishes the best teams from all the rest was a thought that had been in the air for awhile and I wanted to help answer it. Oberstone looks into 24 ”pitch actions” collected during the 2007/2008 season [**?**]. The author separates the teams from that season into 3 groups, the top four teams, the middle twelve, and the bottom four teams. First, a regression was done to determine six statistically significant factors that led to the club’s ultimate success in that season. The six factors were: Percent Goals scored outside box, Percent of goals to shots, Number of yellow cards, Ratio of short to long passes, Number of total crosses, and Average goals conceded per game. Secondly, the 24 variables used in the one-way ANOVA determine that their are thirteen actions that are statistically significant that separate the three groups from one another.

The final paper to be examined was ”Key performance variables between the top 10 and bottom

10 teams in the English Premier League 2012/13 season" by Jaime Araya and Paul Larkin. As the title suggests, Araya and LArkin are examining what variables are different between the teams that finish in the top 10 of the English Premier League and then the teams that finsih in the bottom 10. They state that, "research has indicated that possession, shots at goal, and goals scored are key performance indicators for successful football teams. There is however, a lack of understanding of other potential attacking and defensive performance variables that may contribute to successful performance" [**?**]. The authors look at in-game statistics from 380 games from the 2012/2013 season. Their findings show that top tier teams tend to have more possession (p¿.01), have a lot more shots (p¿.01), score more from inside the 18-yard box (p=.02), have shorter passes, and then score more in open play. Basically, to be a top 10 team in the English Premier League, Araya and Larkin discovered teams need to "keep possession of the ball via short passes, with the attempt to penetrate the opposition defence to have a shot at goal from inside the 18-yard box" [**?**].

3 Data

The primary data source for this research is English Premier League team statistics for the 2006/2007 season through the 2017/2018 seasons. This dataset was found on Kaggle. The data is set up, starting with the 2006/2007 season, by the place each team finished. Manchester United won the 2006/2007 season followed by Chelsea, Liverpool, and Arsenal; after these teams the rest of the league, by position, is ordered. Once the twenty teams from that season are listed, the next season starts with the league leader, and this carries on all the way through to the 20017/2018 season. There are a total of 35 variables used in the data that are also used in regressions for this paper, those variables include: wins, losses, goals, total.yel.card, total.red.card, total.scoring.att, ontarget.scoring.att, hit.woodwork, att.hd.goal, att.pen.goal, att.freekick.goal, att.ibox.goal, att.obox.goal, goal.fastbreak, total.offside, clean.sheet, goals.conceded, saves, outfielder.block, interception, total.tackle, last.man.tackle, total.clearance own.goals, penalty.conceded, pen.goals.conceded, total.pass, total.long.balls, total.cross, corner.taken, touches, clearance.off.line, penalty.save, total.high.claim, punches. For non-soccer fans, some of these variables might be rather confusing, so a few of the variables will be expanded upon. Total.scoring.attempt is a fancy way of saying total number of shots. Att.hd.goal is a header being scored, att.ibox.goal and att.obox.goal are scoring from inside or outside the box. Outfielder.block is when a field player (not the keeper) blocks a shot. Penalty.conceded is when the team fouls someone and gives up a penalty and then penalty.goal.conceded is when they are scored on after giving up the foul. Total.high.claim is when the keeper jumps up and takes the ball over top other players. Most other variables seem rather obvious, even for a non-soccer fan.

The original dataset was rather clean, so no true cleaning of the data had to be done. However, a few extra variables were added based on other given variables to help with correlation and regressions. Since teams are disbursed throughout separate seasons, an extra ”Team” variable was added to show the 39 unique teams that have been in the English Premier League in the last twelve seasons. From this list of teams, I represented how many years they have been in the league in the past twelve years with ”Years in EPL”. ”Total Goals” and ”Total Wins” were added to show the total amount of wins and goals they have had in the last twelve seasons. ”Avg WPS” represents the average amount of wins each team had per season and then ”Avg.goals” then represents the average amount of goals each team had per season.

For one package in R, Corrplot, to create the needed plot to represent correlation, an extra data source was created to make everything into numeric data. In the original dataset, most variables were numeric, however, when trying to change just a few, I was unable to do so. This led to making a new data set and deleting out the unwanted variables in order to run the specific correlation. However, this source was only used for the one package a correlation.

4 Empirical Methods

While my approach explores a number of different approaches, the primary empirical model can be depicted in the following equation:

$$Y_{it} = \alpha_0 + \alpha_1 Z_{it} + \alpha_2 X_{it} + \varepsilon, \tag{1}$$

where Y_{it} is a continuous outcome variable for unit i in year t, and Z_{it} are characteristics about the firm at which i is working, while X_{it} are characteristics about i. The parameter of interest is α_1.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

5 Research Findings

The main results are reported in Table 2.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus.

Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

6 Conclusion

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Figures and Tables

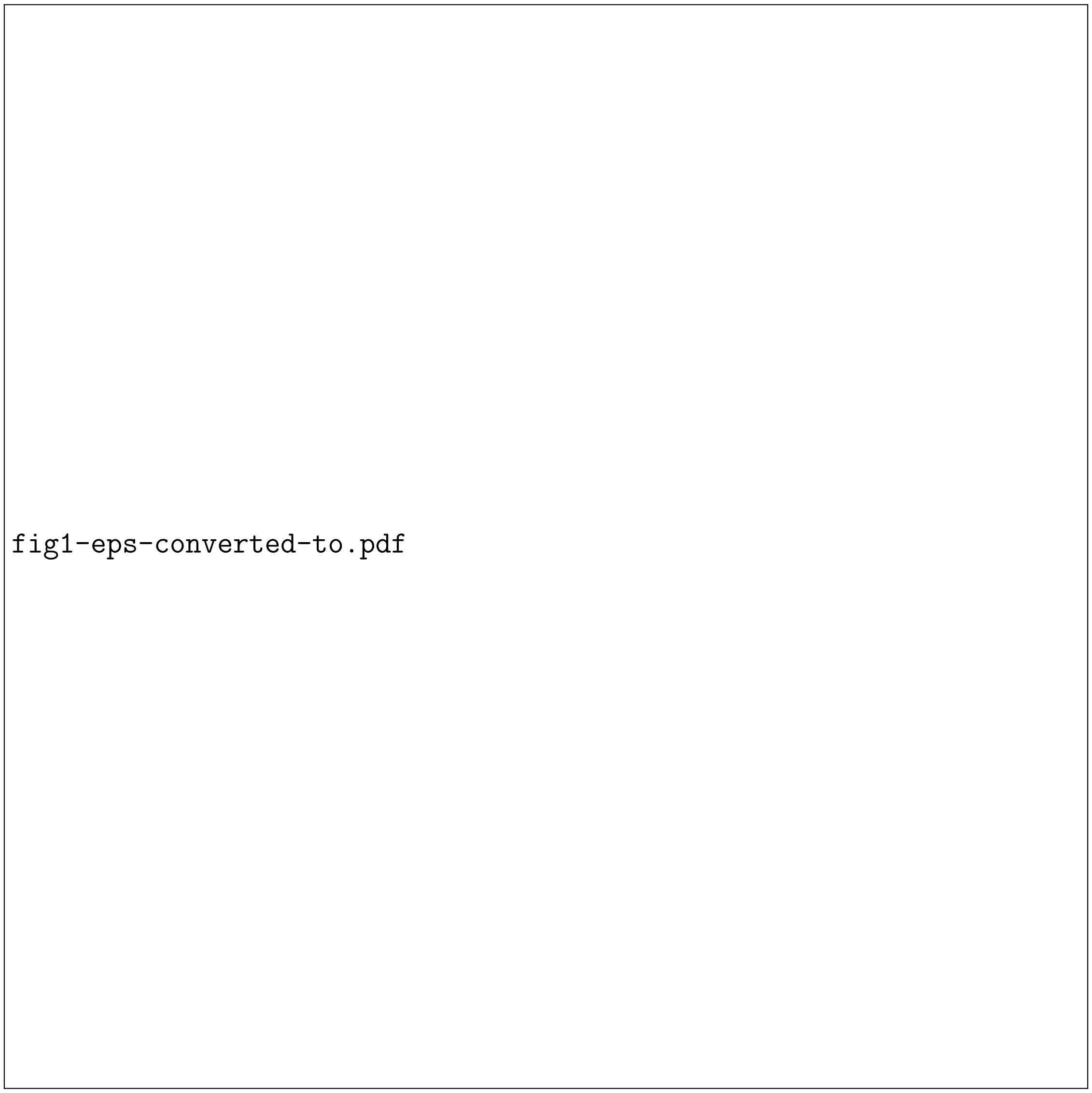

Figure 1: Figure caption goes here

Table 1: Summary Statistics of Variables of Interest

Panel A: Summary Statistics for Variables of Interest

	Mean	Std. Dev.	Min	Max
Outcome variable 1	4.127	1.709	0.000	8.516
Outcome variable 2	1.293	0.648	0.000	0.216
Policy variable	0.685	0.464	0.000	1.000
Control variable 1	0.451	0.497	0.000	1.000
Control variable 2	0.322	0.467	0.000	1.000

Panel B: Sample Means of Outcome Variables for Subgroups

	Group 1	Group 2	Group 3	Group 4
Outcome variable 1	1.782	2.181	3.749	4.127
Outcome variable 2	0.824	0.971	1.215	1.693
N	25,796	75,879	37,157	33,839

Notes: Put any notes about the table here. Sample size for all variables in Panel A is $N = 172,671$.

Table 2: Empirical estimates of parameter of interest

	Few Controls	Many Controls
Variable of interest	-1.977***	-0.536**
	(0.219)	(0.214)
Individual characteristics	✓	✓
Firm characteristics		✓
Location dummies		✓
N	172,671	172,671

Notes: Table notes here. Standard errors in parentheses. ***Significantly different from zero at the 1% level; **Significantly different from zero at the 5% level.