Analysis of Home Advantage in the English Premier League

Vivaan Bhaskar

Table of Contents

Abstract

Home game advantage in soccer is described to be the benefits the home team has over the visiting team through increased goals and wins. A significant explanatory variable for the occurrence of this phenomenon is through the presence of home crowds which spurs enthusiasm and motivation in support of the home team. As English Premier League teams played several games with crowds absent during the height of the pandemic, this instance provided the opportunity to analyze the magnitude of home crowds on home game advantage in soccer. Using exploratory statistical techniques in combination with several data sources, we found there to be significant evidence of home game advantage in the EPL through increased home goals, less yellow and red cards for the home team, and an overall higher home win rate. Furthermore, this effect was exacerbated for "top" EPL teams who typically host larger crowds. We used the same methods to discover that absent crowds during the COVID-season resulted in less home goals, less home conversions and ultimately, lower home win rates. Lastly, we used multivariate regression modeling to find a positive correlation between home winning rate and the number of attendees in the crowd and highlighted the distinctly higher winning rates of home teams prior to the COVID-season. Overall, our thorough analysis explores the presence of home game advantage in the EPL while highlighting the specific impact of crowds on this well-known sports phenomenon.

Overview and Research Question

In team sports, home game advantage is a phenomenon which describes the advantages the home team has over the visiting team through increased goals/points which ultimately leads to a home team victory. This added advantage is caused by a combination of various factors such as referee bias towards the home team, travel fatigue of the visiting team, the home team's familiarity with the home stadium, and more.

Posing the following research question: **"Does the presence of crowds influence home game advantage in the English Premier League?"**, we will analyze the specific impact home crowds have on home team performance within one of the most prominent soccer leagues. Based on prior knowledge, we know that supportive crowds are one of the strongest motivational forces for home team players and performance. The occurrence of the pandemic provides us the unique opportunity to test the presence of crowds on home game advantage throughout EPL seasons, as several games were conducted with zero attendees during the COVID season.

Throughout this research paper, we will explain the data and methodology behind our research, highlight key visualizations and models from our results, and finally summarize our findings on the impact of crowds on home advantage.

Data

This section introduces the datasets used to conduct our analysis of Home Advantage. We focus our analysis on the English Premier League, and source data from the *Football Data UK website* (1). There is data available for almost the entirety of the Premier League (the competition's name changed from Premiership in 1992). Data for the early years included only results for each game. From the 2000s, it started offering in-depth statistics per game for the entire season. The timeframe for our analysis is from *2009/10 to 2020/21* season. The 2020/21 season was played during capacity restrictions, i.e., limited capacity was allowed for some games, while the majority of games were played behind closed doors.

From the varied statistics available, we choose data for the following - Home and Away team goals, the full-time result, number of shots and shots on target per team, the number of yellow and card given, and lastly, fouls committed by each team. There is one observation per match (2), so we clean the data to make sure of two instances, one for each team per match. Here, an indicator variable is introduced to identify whether the team played at home or away, *HomeAway*. Instead of a result column, we introduce three indicator variables to indicate whether the match resulted in *Win, Loss, or Draw.*

For our analysis, we create additional statistics. The metrics used further in the analysis are explained below:

1. *Cards* = Yellow cards + Red cards * 10 → This variable is created to capture the implications of a red card. (3) and (4) discuss the implications of a red card in a game. The former's analysis is done for the 2018/19 season and found 59% of games were lost after a red card was given.

2. *Conversion* = Goals per game / Shots per game → This variable captures the efficiency of shots taken in a game. A higher ratio implies more efficient shot taking.

3. *Cards to Foul* = Cards / Fouls Committed → This is a proxy for referee bias. It captures the ratio of cards given per foul. A higher ratio implies that there is a higher chance for a card per foul.

The modified data is shown here (5). We also standardize the new statistics calculated for easier comparison.

Further, we look to add attendance data to understand the influence of crowds in home advantage. For this, we sourced data from Kaggle (6). The dataset includes average attendance data for all teams in a premier league season from 1949-2019. Data for remaining years is taken from the World Football website (7). There are 36 teams that featured in the premier league in the season concerned, with only seven teams featuring in all 12 seasons. Here, we create three additional dummy variables, *Top, Mid, and Low*, which categorizes the teams in the three

categories. They are dependent on the number of premier league appearances in the last 12 seasons. The cut off for mid teams was at least being part of 6 of the 12 seasons considered. (8) discusses how teams that get relegated have major downfalls in terms of media rights, crowd sizes and others.

Methodology

This section introduces our approach taken to understand the influence of home advantage in the English Premier League and whether crowds influence home advantage. The aim is to understand differences in performance of home and away teams. We now introduce our three-facet approach.

The first part of our analysis is exploratory. We focus on the main dataset with performance variables. Here, we look to find empirical evidence of differences in performance by comparing the win rate, goals scored and conceded, the manipulated 'cards' variable, efficiency in shots taken (conversion) and cards per foul (proxy for referee bias).

First, performance is compared across home and away teams for each observation in the data frame (approximately 500 instances). The differences in mean values are calculated for each variable discussed above for comparison. Next, data is grouped by team category. The categories are top, mid, and low, which are introduced in the earlier section. This is done to understand differences in performance of teams that are considered more experienced with larger stadiums and better squads. Team performance can be understood inside each category through the difference metric and across categories to further our understanding of the influence. Finally, we look to compare performance across seasons to understand differences in performance across other seasons that were not affected by capacity restrictions with the 2020/21 season.

The second part of our analysis focuses on building multivariate regression models. The first model we built to understand the factors that contribute to home advantage. We club data by season and team and include the attendance data for this part and include data for only home games. There are 233 observations. Win% of a team in each season is taken as the independent variable, and the same five variables discussed above and include 'Attendance'. All variables are standardized. The next model includes controls for team category, which provides additional insight into the effect team size and experience play in the creation of home advantage.

The last part explores the effect of crowds on home advantage. This is done to understand what drives teams playing at home. Several questions are raised - whether players are used to the dimensions of the pitch, and playing on the particular ground, or is it the crowds that represent the fans cheering them on, taking on the role of the '12th man'. The 2020/21 season provides an interesting scenario to test this hypothesis. The games are played on the

same ground as earlier games were played but with no crowds. Hence, we compare performance of teams at home pre-restrictions against performance in the 2020/21.

Visualizations

To better understand the home advantage and explanatory variables behind it, we performed descriptive statistics shown through three visualizations [(9), (10), (11)] which describe the differences between home and away team performances overall, by team category (low, mid, high tier), and by season, respectively. We further standardized all the visualizations so that the results would have consistency.

In visualization (9), we found that the average goal difference (shown by the blue bar) between a home and away team playing at the home team's stadium was approximately 0.4 higher standard deviations for the home team, a figure which is quite significant. Similarly, the average goals conceded (shown by the orange bar) was approximately 0.4 standard deviations lower for the home team, which further supports the higher average goal difference result. The most significant result from this visualization is the average win difference between a home and away team (shown by the purple bar) being approximately 1.5 standard deviations higher, which is a strong indicator for the presence of home game advantage. It is also very important to note the number of cards shown to home relative to away teams (shown by the grey bar), which shows that when a team travels for an away game, the home team has a slight advantage likely due to familiarity of the ground and playing conditions apart from the crowds. As a result, away teams play a more aggressive game hoping to aim for a draw to gain a point rather than receiving no points if they lose. This ultimately results in higher cards on average being shown to away teams, as away teams have approximately 0.5 standard deviations more cards than the home team. This statistic includes receiving a red card meaning players are sent off for the remainder of the match which can have a severe negative effect on the match result.

In visualization (10), we categorized the teams in the EPL as a top, mid, and low-level team to analyze the impacts of potential home advantages based on team performance. We see that the average goal difference (shown by the first set of bars) between a top home team relative to any away team is much higher - at approximately 0.5 more standard deviations - than the difference between a mid or low-level home team. From the first set of bars, we also see that the average win difference between top home teams is higher, although not drastically different from mid and low level home team victories. However, looking at the 8th set of bars which depicts average cards received by home teams, we see a very significant result showing home teams receiving much fewer cards relative to mid and low home teams. This highlights the fact that a low or mid-level away team which travels to a top home team for a match, would probably play a very aggressive game to try and finish with a draw which can result in more cards being handed to that away team for potential fouls committed as a result. These results help in further understanding the impact of home advantage for the strength of a team being a top, mid, or low-level team which from the results shows that top teams tend to have the highest positive impact from home advantages which is not surprising.

In visualization (11), we compared the home vs away team statistics across multiple seasons to assess whether such differences were consistent to further understand the impact of home advantage. This visualization shows the most striking results with respect to the effect of crowds on the home advantage for a team. If you concentrate on the last bar within the first set of bars, it shows that the average goal difference between a home and away team is almost negligible. This is fascinating since this bar represents the pandemic year when COVID-19 resulted in the absence of crowds at matches. When compared to the other bars, we see that every year before, a home team has a much higher average goal scoring rate relative to away teams which we assume to be caused by the presence of crowds driving home advantage. Another very interesting result is in the second set of results which shows the average win rate between home and away teams, we see in the last bar that home teams were more likely to lose on average against an away team in the year when there was an absence of crowds. This is relative to the other years when home teams had a much higher average win rate as compared to away teams. These significant results highlight the potential role that crowds play behind giving home teams an advantage.

Regression Model

In this multivariate regression model, we regressed *Home Team Winning Percentage* against variables *Attendance* (measures crowds), *Goals* (measures goals for home team), *Conceded* (measures goals against home team), *Conversion* (measures the goals per shot taken), and *Cards to Foul* (measures referee bias) to prove the home game advantage phenomenon. We also incorporated controlled variables, as mentioned earlier, referring to *Top, Mid,* and *Low* teams to capture the true *Home Team Winning Percentage*. We computed three different regression models with each model being limited to certain variables to help select the best fitting model and to improve interpretability.

Focusing on Model 6 (12), the output produced evidently depicts the impact each of the listed variables have on *Home Team Winning Percentage*. For example, holding all else constant, conceding a goal will cause *Home Team Winning Percentage* to fall by ~ 18 percentage points and scoring a goal will increase it by ~ 23 percentage points. Furthermore, if the home team falls under the *Top, Mid, or Low* categories then it will increase its winning percentage by ~ 9 percentage points, 7 percentage points, and 7 percentage points respectively. Another application of this model that we can realize is that if the home team were to convert an extra goal from an extra shot taken, their chances of winning would rise by ~ 48 percentage points. Nonetheless, not only are these coefficients statistically significant, but the R-squared is relatively high at ~ 88 percentage points which concludes that most of the variation is explained by the model.

Additionally, we created another regression model (13) however this time, included the *HomeAway* variable as a replacement for the dummy variables *Top, Mid,* and *Low* to account for the entire dataset rather than just home games. Through our analysis in Model 5, we discovered that when comparing two teams with the same statistics, the winning percentage for a home team is about 14 percentage points higher than that of an away team at a 1 percent significance level, further proving the home game advantage phenomenon. Moreover, in this regression model we omitted the other variables to see the true impact a home team/away team has on winning percentage.

Impact of Crowds

To test the impact of crowds on home advantage, we built three different regression models for pre-covid era and covid season (14). Model 1 includes all performance variables and excludes cards_to_foul and attendance variables. Model 2 excludes the performance variables and focuses on cards_to_foul and attendance variables. Model 3 includes all performance variables. Although all three models show high statistical significance, we focus on model 2 to analyze the impact of crowds as it excludes performance variables and only includes attendance and cards_to_foul variables, which are factors that are highly correlated and influential to home advantage. Further, we focus on the top teams because since they are the Big 6 of the EPL, they have the largest supporters; thus, intuitively, attendance will be above average and will possess higher influence on the game. Model 2's coefficient of 0.24, statistically significant at a 99% level, indicates a 24% increase in chance of the home team winning during the pre-covid era. The coefficient of 0.11, statistically significant at a 90% level, indicates an 11% increase in chance of the home team winning during the covid season. From these two coefficients, we can conclude that the probability of winning for home, top teams increase by 13% with crowds, which confirms our research question of the existence of impact of crowds on home advantage.

Summary

To explore the presence and magnitude of a home crowds' influence on home advantage in the English Premier League, we derived multiple regressions with contributing factors during pre-covid seasons and the covid season. Through our regressions, we found that home advantage is indeed present, notably at a larger magnitude for the Big 6, top teams. Through our analysis of the differences between home and away statistics, we concluded that home advantage increased the number of goals and decreased the number of cards from fouls, thus, increased the overall winning probability for home teams. Through our analysis, we saw that these magnitudes greatly decreased during covid season in which games were played in empty stadiums. Through further analysis, we found that the difference between these magnitudes between covid and pre-covid seasons were largely due to the absence of crowds. With high statistical significance, we concluded that the winning probability for home teams increased by 13% with crowds for top teams in the league. In sum, we concluded our research question that the presence of home crowds is indeed a strong driving force behind home advantage in the English Premier League. However, it is important to note that our methodology may be limited due to potential confounding variables, such as team performance during covid being affected by player illness rather than crowds, which is not captured in our model. Another limitation is that we took an average for crowd attendance rather than match-by-match attendance which could have led to lower model accuracy. Overall, while there are limits within our research methodology, our models still displayed high statistical significance and allowed us to answer our research question with confidence.

References

(1)
https://www.football-data.co.uk/englandm.php

(2)

	Div	Day	Month	Year	Season	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HS	AS	HST	AST	HF	AF	HY	AY	HR	AR
0	E0	15	8	2009	2009/2010	Aston Villa	Wigan	0	2	A	11	14	5	7	15	14	2	2	0	0
1	E0	15	8	2009	2009/2010	Blackburn	Man City	0	2	A	17	8	9	5	12	9	2	1	0	0
2	E0	15	8	2009	2009/2010	Bolton	Sunderland	0	1	A	11	20	3	13	16	10	2	1	0	0
3	E0	15	8	2009	2009/2010	Chelsea	Hull	2	1	H	26	7	12	3	13	15	1	2	0	0
4	E0	15	8	2009	2009/2010	Everton	Arsenal FC	1	6	A	8	15	5	9	11	13	0	0	0	0

(3)
https://www.researchgate.net/publication/46554858_Estimating_the_Effect_of_the_Red_Card_in_Soccer_When_to_Commit_an_Offense_in_Exchange_for_Preventing_a_Goal_Opportunity

(4)
https://blog.innerdrive.co.uk/sports/impact-of-red-cards-on-footballers

(5)

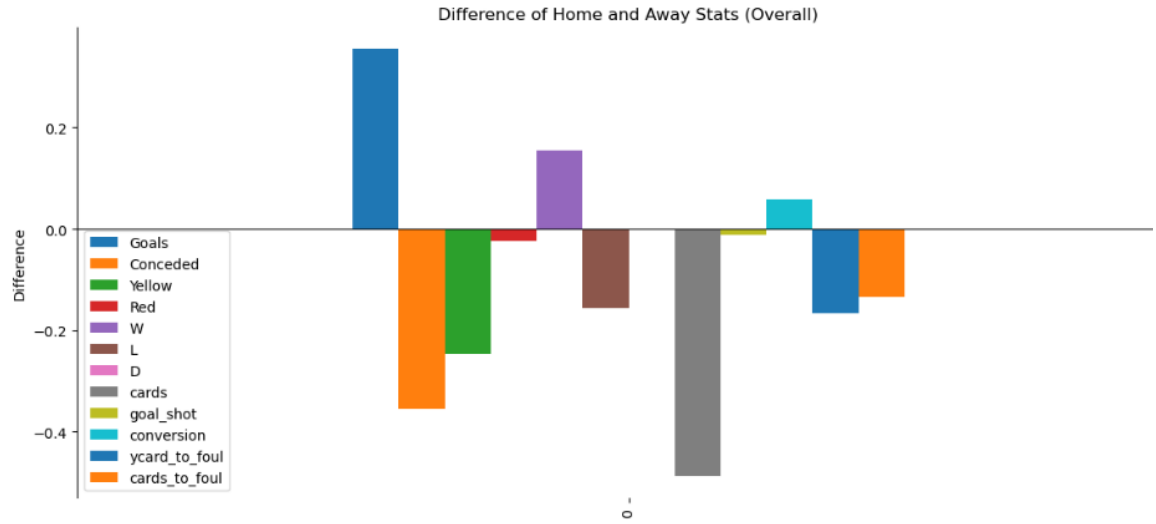
	Season	Team	HomeAway	Goals	Conceded	Yellow	Red	W	L	D	cards	goal_shot	conversion	ycard_to_foul	cards_to_foul
0	2009/2010	Aston Villa	1	0	2	2	0	0	1	0	2	0.246606	-1.038670	-0.142676	-0.276849
1	2009/2010	Blackburn	1	0	2	2	0	0	1	0	2	0.648618	-1.038670	0.135765	-0.165826
2	2009/2010	Bolton	1	0	1	2	0	0	1	0	2	-0.729707	-1.038670	-0.212287	-0.304605
3	2009/2010	Chelsea	1	2	1	1	0	1	0	0	1	0.284157	-0.339236	-0.613885	-0.464734
4	2009/2010	Everton	1	1	6	0	0	0	1	0	0	1.161900	0.097911	-1.256443	-0.720941

(6)
https://www.kaggle.com/datasets/joovasco/premier-league-attendance-from-1949-to-2019

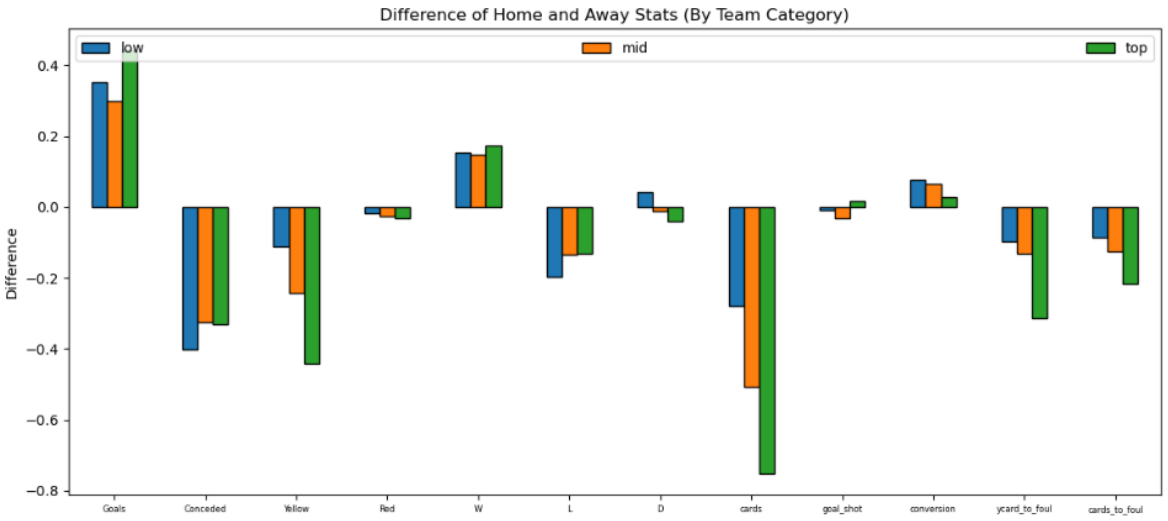
(7)
https://www.worldfootball.net/attendance/eng-premier-league-2020-2021/1/

(8)
https://bleacherreport.com/articles/1591059-the-cost-of-relegation-5-reasons-to-stay-in-the-epl

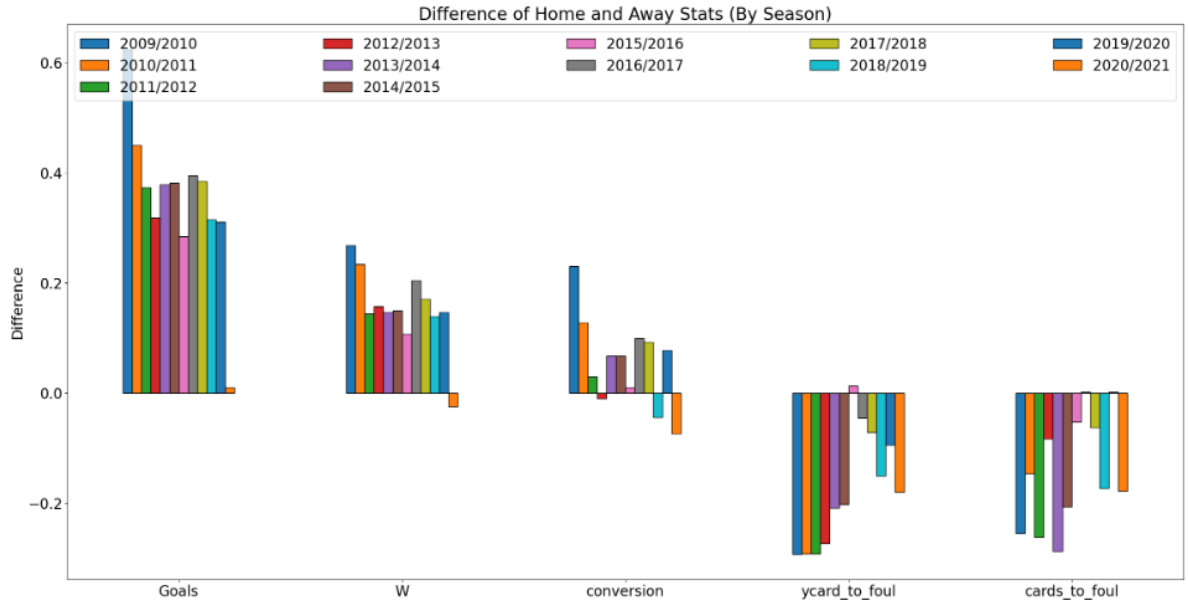
(9)


Difference of Home and Away Stats (Overall)
Difference
0.2
0.0
−0.2
−0.4
Goals
Conceded
Yellow
Red
W
L
D
cards
goal_shot
conversion
ycard_to_foul
cards_to_foul
0

(10)


Difference of Home and Away Stats (By Team Category)
low
mid
top
Difference
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
Goals
Conceded
Yellow
Red
W
L
D
cards
goal_shot
conversion
ycard_to_foul
cards_to_foul

(11)



(12)

	Model 4	Model 5	Model 6
const	0.23***	0.34***	0.21***
	(0.02)	(0.02)	(0.03)
Att_10000		0.03***	0.01*
		(0.01)	(0.00)
Goals	0.21***		0.20***
	(0.02)		(0.02)
Conceded	-0.19***		-0.18***
	(0.02)		(0.02)
conversion	0.43*		0.48**
	(0.23)		(0.23)
cards_to_foul		-0.36***	0.03
		(0.13)	(0.07)
top	0.09***	0.26***	0.08***
	(0.01)	(0.02)	(0.01)
mid	0.07***	0.04***	0.06***
	(0.01)	(0.01)	(0.01)
low	0.07***	0.04***	0.06***
	(0.01)	(0.02)	(0.01)
R-squared	0.88	0.51	0.88
R-squared Adj.	0.88	0.50	0.88
R-squared	0.88	0.51	0.88
No. observations	233	233	233

Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01

(13)

	Model 4	Model 5	Model 6
const	0.27***	0.27***	0.27***
	(0.02)	(0.03)	(0.02)
Att_10000		0.04***	0.00
		(0.00)	(0.00)
Goals	0.24***		0.24***
	(0.01)		(0.01)
Conceded	-0.17***		-0.17***
	(0.01)		(0.01)
conversion	0.15		
	(0.16)		
cards_to_foul		-0.39***	
		(0.11)	
HomeAway	0.01	0.14***	0.01
	(0.01)	(0.02)	(0.01)
R-squared	0.88	0.32	0.88
R-squared Adj.	0.88	0.31	0.88
R-squared	0.88	0.32	0.88
No. observations	466	466	466

Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01

(14)

	Model 1(pre)	Model 1(covid)	Model 2(pre)	Model 2(covid)	Model 3(pre)	Model 3(covid)
const	0.25***	0.22**	0.28***	0.12*	0.23***	0.12
	(0.03)	(0.10)	(0.02)	(0.06)	(0.02)	(0.09)
Attendance	0.00*	-0.00	0.00***	0.00**	0.00*	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Goals	0.21***	0.20***			0.23***	0.22***
	(0.02)	(0.05)			(0.01)	(0.04)
Conceded	-0.18***	-0.13**			-0.18***	-0.10
	(0.02)	(0.06)			(0.02)	(0.06)
conversion	0.05	0.06				
	(0.03)	(0.05)				
cards_to_foul			-0.10**	-0.23	0.01	-0.10
			(0.04)	(0.14)	(0.02)	(0.09)
top	0.10***	0.09*	0.24***	0.11*	0.08***	0.04
	(0.02)	(0.05)	(0.02)	(0.05)	(0.01)	(0.04)
mid	0.08***	0.06	0.02	0.00	0.07***	0.04
	(0.01)	(0.04)	(0.01)	(0.03)	(0.01)	(0.04)
low	0.08***	0.07*	0.02	0.01	0.07***	0.04
	(0.01)	(0.04)	(0.01)	(0.03)	(0.01)	(0.03)
R-squared	0.88	0.89	0.51	0.65	0.88	0.89
R-squared Adj.	0.88	0.84	0.50	0.56	0.88	0.85
R-squared	0.88	0.89	0.51	0.65	0.88	0.89
No. observations	213	20	213	20	213	20

Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01