The Effect of Player Salary on Home Attendance Using 96 Soccer Teams in Europe.



Bailey Thompson

04/25/2017

Econometrics 3112 – WF: 12:30

Dr. Stivender

# Part 1: Describing the Data

The research question I intend to answer is:*What is the marginal effect of average player salary on the percentage of home seats sold?*

* In other words, does spending more money on players have an effect on a team’s yearly attendance?
  + If so, is there a maximum amount soccer teams should pay their players?
* Perhaps performance has an effect on attendance?
  + If so, which performance factor has the greatest marginal effect?

Summary:

The data I have used in my regression is **team level, seasonal data, from over the course of 3 years (2012-2014).**

* I have chosen to include every team from four of the most popular and competitive leagues in Europe. These leagues are:
  + La Liga (Spain)
  + Bundesliga (Germany)
  + BPL (England)
  + Serie A (Italy)

I have included fourteen independent variables in my regression that I believe have an effect on team attendance. They are listed in the following table.

Independent Variables:

|  |  |  |
| --- | --- | --- |
| Variable Name | Description | Units |
| *player\_salary\_thou* | Team average player salary | Measured in thousands of dollars |
| *player\_salary\_thou2* | Team average player salary squared | Measured in thousands of dollars |
| *ln\_home\_city\_pop* | The team’s average home city population for season | Measured in percentage change |
| *team\_ytd\_age* | The team’s year-to-date age | Measured in number of years |
| *win\_percentage* | Team win percentage for season | Measured in percentage points |
| *losses* | Team losses for season | Measured in number of losses |
| *draws* | Team draws for season | Measured in number of draws |
| *goals\_for* | The number of goals team scored in season | Measured in number of goals |
| *goals\_against* | The number of goals team conceded in season | Measured in number of goals |
| *won\_championship* | Did the team win their league’s championship? | Binary Choice (Y=1,N=0) |
| *share\_stadium* | Does the team share their stadium? | (Y=1,N=0) |
| *spain* | Is the team located in Spain? | (Y=1,N=0) |
| *germany* | Is the team located in Germany? | (Y=1,N=0) (Omitted Variable) |
| *italy* | Is the team located in Italy? | (Y=1,N=0) |
| *england* | Is the team located in England? | (Y=1,N=0) |
|  |  |  |

My dependent variable, attendance, is described below:

Dependent Variable

|  |  |  |
| --- | --- | --- |
| Variable Name | Description | Units |
| *ln\_avg\_per\_of\_home\_seat\_sold* | The team’s average percentage of home seats sold in season | Measured in percentage change |

Variables generated in SATA

I generated log variables in STATA so that some of my interpretations make more sense. The commands I used are below.

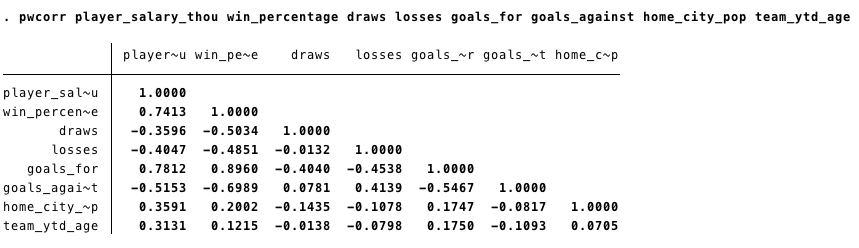
* Generating *ln\_avg\_per\_of\_home\_seat*:
* Generating *ln\_home\_city\_pop*:



Correlation Information:

To test for multicollinearity, I ran Pearson’s correlation test in STATA.

* The correlation test shows which direction and how correlated variables are.
* The command I used for Pearson’s correlation test and its results are below:



The results of the correlation test are clear, I have four issues with multicollinearity that I need to address.

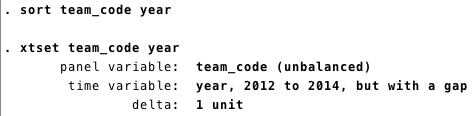
* *goals\_for* is highly and positively correlated with *player\_salary\_thou* (.7812) and *win\_percentage* (.8960).
  + I believe that *goals\_for* and *win\_percentage* may be accounting for too much of each others independent movement; since they are both measuring how well the team performed.
  + **Therefore, I have decided to eliminate *goals\_for* as one of my independent variables.**
* *goals\_against* is highly and negatively correlated with *win\_percentage* (-.6989)
  + This correlation **is a data-based issue.** It would be expected that *goals\_against* and *win\_percentage* be negatively correlated.
  + **Therefore, I believe that this variable should stay in my regression.**
* *win\_percentage* is highly correlated with *player\_salary* (.7413).
  + This correlation **is a data-based issue.** It would be expected that the team’s win percentage would be highly and positively correlated with how much the team pays their players, because of the assumption that money can afford better talent.
  + **For this reason, I believe it is necessary to leave *win\_percentage* in my regression.**

Part 2: My Regression

Pre-regression Notes:

It is important to note that my data measures teams over time, thus it is panel data.

* I need to manipulate my data so that STATA is aware I have panel data.
  + The command I used is below:



Before running my regression I used STATA commands to summarize and describe my data; ensuring no mistakes or discrepancies were made during data entry or importation.

* I ran both the describe command and summarize command and found no discrepancies.

Confidence level and Null Hypothesis:

I am testing my results at a 95% confidence level with a sample size of n=221. Therefore there are 2 standards that my variables will have to meet.

* Z-score must be more than 1.984.
* P-value must be below 2.5%.
* 0 cannot be located within the confidence interval.

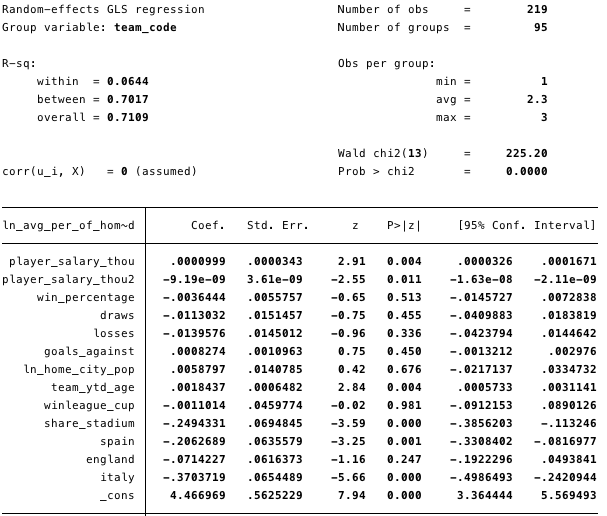
The null hypothesis for all variables is that they are statistically insignificant (H=0) until they have met all standards.

All interpretations are the average marginal effect, ceteris paribus.

(Results on next page)

Regression Results: Log-Linear Model

****

****

Part 3: Interpretation

Overall R2:The overall R2 for this regression is .7109, therefore, 71.09% of the variation in *ln\_avg\_per\_of\_home\_seat* around it’s mean is explained by the independent variables in my regression.

*player\_salary\_thou & player\_salary\_thou2:* **This variable is statistically significant at a 95% confidence level, therefore, the null hypothesis of H=0 is rejected.** I know that this variable is significant based on three reasons:

* *player\_salary\_thou* has a z-score of 2.91, which is above the threshold of 1.984.
* *player\_salary\_thou* has a p-value of .004.
  + .4% probability we would make a type 1 error by rejecting the null.
* 0 is not located within the confidence interval.

**Since *player\_salary\_thou2* is also statistically significant, I have determined with 95% confidence that there is a nonlinear relationship between *player\_salary\_thou* and *ln\_avg\_per\_of\_home\_seat*.** I know that this variable is significant based on three reasons:

* *ln\_avg\_per\_of\_home\_seat* has a z-score of 2.55, which is above the threshold of 1.984
* *ln\_avg\_per\_of\_home\_seat* has a p-value of .011.
  + 1.1% probability we would be making a type 1 error by rejecting the null.
* 0 is not located within the confidence interval.

The equation for marginal effect of *player\_salary* on *ln\_avg\_per\_of\_home\_seat* is:

*(player\_salary\_thou)* = .00999 - .000001838(*player\_salary\_thou*)

* **As shown by the equation, the effect of *player\_salary\_thou* depends on the valueof *player\_salary\_thou*.**
  + In other words, every amount of *player\_salary\_thou* has a different marginal effect on *ln\_avg\_per\_of\_home\_seat*.

**From the regression, I was able to compute the maximum amount of *player\_salary\_thou* a team should pay their players.**

* This point is the inflection point (IP) and is computed by the equation below:
  + IP = B1/2(B2)
  + IP = .00999/2(.000000919)
    - IP = $5,435.255713 thousands of dollars
      * IP = $5,435,255.71
* **Therefore, the marginal effect of *player\_salary\_thou* on *ln\_avg\_per\_of\_home\_seat* is positive and getting smaller before the inflection point of $5,435.26 thousands of dollars and after the inflection point the marginal effect is negative and getting larger.**

*win\_percentage:* **This variable is statistically insignificant at a 95% confidence level, therefore I failed to reject the null of H=0.** I know that this variable is insignificant based on three reasons:

* *win\_percentage* has a z-score of .65, which does not meet the threshold of 1.984.
* *win\_percentage* has a p-value of .513.
  + 51.3% probability we would be making a type one error by rejecting the null.
* 0 is located within the confidence interval.

*draws:* **This variable is statistically insignificant at a 95% confidence level, therefore I failed to reject the null of H=0.** I know that this variable is insignificant based on three reasons:

* *draws* has a z-score of .75, which does not meet the threshold of 1.984.
* *draws* has a p-value of .455.
  + 45.5% probability we would be making a type one error by rejecting the null.
* 0 is located within the confidence interval.

*losses:* **This variable is statistically insignificant at a 95% confidence level, therefore I failed to reject the null of H=0**. I know that this variable is insignificant based on three reasons:

* *losses* has a z-score of .96, which does not meet the threshold of 1.984.
* *losses* has a p-value of .336.
  + 33.6% probability we would be making a type one error by rejecting the null.
* 0 is located within the confidence interval.

*goals\_against:* **This variable is statistically insignificant at a 95% confidence level, therefore, I failed to reject the null of H=0.** I know that this variable is insignificant based on three reasons:

* *goals\_against* has a z-score of .75, which does not meet the threshold of 1.984.
* *goals\_against* has a p-value of .450.
  + 45% probability we would be making a type one error by rejecting the null.
* 0 is located within the confidence interval.

*ln\_home\_city\_pop:* **This variable is statistically insignificant at a 95% confidence level, therefore, I failed to reject the null of H=0.** I know that this variable is insignificant based on three reasons:

* *ln\_home\_city\_pop* has a z-score of .42, which does not meet the threshold of 1.984.
* *ln\_home\_city\_pop* has a p-value of .676.
  + 67.6% probability we would be making a type one error by rejecting the null.
* 0 is located within the confidence interval.

*team\_ytd\_age:* **This variable is significant at a 95% confidence level, therefore, the null hypothesis of H=0 is rejected.** I know that this variable is significant based on three reasons:

* *team\_ytd\_age* has a z-score of 2.84, which is above the threshold of 1.984.
* *team\_ytd\_age* has a p-value of .004.
  + .4% probability we will be making a type one error by rejecting the null.
* 0 is not located within the confidence interval.

**The marginal effect of *team\_ytd\_age* on *ln\_avg\_per\_of\_home\_seat* is:**

* For every additional year of *team\_ytd\_age*,percentage of home seats sold increases by .18437%.

*winleague\_cup:* **This variable is statistically insignificant at a 95% confidence level, therefore, I failed to reject the null of H=0.** I know that this variable is insignificant based on three reasons:

* *winleague\_cup* has a z-score of .02, which does not meet the threshold of 1.984.
* *winleague\_cup* has a p-value of .981.
  + 98.1% probability we will be making a type one error by rejecting the null.
* 0 is located within the confidence interval.

*share\_stadium:* **This variable is statistically significant at a 95% confidence level, therefore, the null hypothesis of H=0 is rejected.** I know that this variable is significant based on creasons:

* *share\_stadium* has a z-score of 3.59, which is above the threshold of 1.984
* *share\_stadium* has a p-value of .000.
  + 0% probability we will be making a type one error by rejecting the null.
* 0 is not located within the confidence interval.

**The marginal effect of *share\_stadium* on *ln\_avg\_per\_of\_home\_seat* is interpreted through the following equation:**

* (*share\_stadium*)= 100\*(eB1-1)
  + (*share\_stadium*)= 100\*(e-.2494331-1)
  + (*share\_stadium*)= -22.07%
* Therefore, if the team shares a stadium with another team in its league, the percentage of home seats sold will be 22.06% lower, than when compared to teams that do not share a stadium.

*spain:* **This variable is statistically significant at a 95% confidence, therefore, the null hypothesis of H=0 is rejected.** I know that this variable is significant based on three reasons:

* *spain* has a z-score of 3.25, which is above the threshold of 1.984.
* *spain* has a p-value of .001.
  + .01% probability we will be making a type one error by rejecting the null.
* 0 is not located within the confidence interval.

**The marginal effect of *spain* on *ln\_avg\_per\_of\_home\_seat* is interpreted through the following equation:**

* (*spain*)= 100\*(eB1-1)
  + (*spain)*= 100\*(e-.2062689-1)
  + (*spain)*= -18.64%
* Therefore, if the team is located in Spain, the percentage of home seats sold will be 18.64% lower, than when compared to a team located in Germany.

*england:* **This variable is statistically insignificant at a 95% confidence level, I fail to reject the null of H=0.** I know that this variable is insignificant based on three reasons:

* *england* has a z-score of 1.16, which does not meet the threshold 1.984.
* *england* has a p-value of .247.
  + 24.7% probability I will be making a type one error by rejecting the null.
* 0 is located within the confidence interval.

*italy:* **This variable is statistically significant at a 95% confidence level, therefore, the null hypothesis of H=0 is rejected.** I know that this variable is significant based on three reasons:

* *italy* has a z-score of 5.66, which meets the threshold of 1.984.
* *italy* has a p-value of .000.
  + 0% probability I will be making a type one error by rejecting the null.
* 0 is not located within the confidence interval.

**The marginal effect of *italy* on *ln\_avg\_per\_of\_home\_seat* is interpreted through the following equation:**

* (*italy*)= 100\*(eB1-1)
  + (*italy)*= 100\*(e-.3703719-1)
  + (*italy)*= -30.95
* Therefore, if the team is located in Italy, the percentage of home seats sold will be 30.95% lower, than when compared to a team located in Germany.

Part 4: Conclusion

Bottom Line:

Based off of the data available to me, and the model that I ran; I determined with 95% confidence that 5 independent variables in my model are statistically significant and I was able to compute their marginal effects*.* The 5 significant independent variables again are:

* *player\_salary\_thou*
  + I was able to determine that there is a nonlinear relationship between *player\_salary\_thou* and *ln\_avg\_per\_of\_home\_seat.* 
    - The amount a team should pay its players if it’s goal is to maximize attendance, cet par, is: $5,435,255.71
* *team\_ytd\_age*
* *share\_stadium*
* *spain*
* *italy*

I was also able to determine at a 95% confidence level (**based off of the data available to me and the model I ran**) that performance **does not** have an effect on *ln\_avg\_per\_of\_home\_seat*.

* I came to this conclusion because all of the performance factors I included in my model were **not** significant at a 95% confidence level.
  + This could be explained by fan loyalty; perhaps fans are loyal enough to still attend games even if their team is not doing well.

In hindsight, to increase the accuracy of my model, I probably would have found more independent variables such has advertising dollars spent per season, the number of games broadcasted per season, and marketing dollars per season to see if these factors have an influence on team attendance.