Darren Parry

Do Economic Factors Affect Their Respective Sports Team's Performance?

Introduction

Being an avid sports fan, I have always wondered whether there is any correlation between the team's performance in their respective sports leagues and economic factors throughout the city and region each team is in. It is obvious that financial factors within the organization are a huge factor in determining the success of a team. For example, if one team's profit is one of the league's highest and they have the resources to spend money on the best players in the league, there will be a positive connection between winning percentage and the team's economic resources. However, throughout this process, I was wondering whether the economic prosperity had any correlation to the respective team's success. For example, six of the past ten NBA Finals winners were in the west region of the country. Given this information, the research question I wanted to answer was how does team performance in American professional sports leagues relate to various economic factors of the cities or regions in which the teams are based?

Research Steps

The first steps in developing my research question involved gathering data on the four professional sports leagues in the United States and finding each team's winning percentage in the last ten years. The respective leagues I am doing are Major League Baseball (MLB), National Basketball Association (NBA), National Football League (NFL), and the National Hockey League (NHL). I also gathered the number of championships won by each team in the last ten years; combined with the winning percentage in the last ten years, this is a good metric to determine each team's success in the last ten years. Regarding the economic factors, I chose to include the total population of the city, the average household income, the average household size, the median age of the residents, and the total retail sales per capita in the city. These factors are a pretty good indication of the well-being of the city, and it could give surprising results as to how these factors affect certain team's performances. Most of the teams in these leagues have a city next to their team's name, like the San Francisco Giants, however if one of the team names had a state, like the Arizona Cardinals, I chose the city that the stadium was located in to make all of the data points similar sizes. After using statmuse.com to find the respective sports data, the US census to find the economic data, and a region map to determine which teams were in what region of the United States I manually put it all in one datasheet, so it was ready to be regressed in two different models in Stata.

OLS Regression

Using the OLS regression model, there are two regression equations that I want to interpret to determine if these economic factors are statistically significant in the sense of each team's performance. These equations are $win\_percentage = \beta 0 + \beta 1 total\_retail\_sales\_per\_capita + \beta 2 population + \beta 3 average\_household\_income + \beta 4 median\_age + \beta 5 average\_household\_size + u$ and $championships\_won = \beta 0 + \beta 1 total\_retail\_sales\_per\_capita + \beta 2 population + \beta 3 average\_household\_income + \beta 4 median\_age + \beta 5 average\_household\_size + u$. To regress these performance metrics on certain economic variables, we first need to check if these economic metrics are statistically significant on their own according to the different data. After putting everything in STATA and regressing each team's performance statistic on their specific economic metric and using a t-value of 2.26 with a 95% confidence interval, the only statistically significant variable was regressing *championships_won* on *total_retail_sales_per_capita.* Although this isn't a great start in figuring out statistically significant variables regarding team performance, there is still hope in figuring out if certain economic factors affect team performance.

Source	SS	df	MS	Number of obs	=	124
				F(1, 122)	=	6.82
Model	2.80938044	1	2.80938044	Prob > F	=	0.0102
Residual	50.2873938	122	.412191752	R-squared	=	0.0529
				Adj R-squared	=	0.0451
Total	53.0967742	123	.431681091	Root MSE	=	.64202

championships_won	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
average_household_income	7.83e-06	3.00e-06	2.61	0.010	1.89e-06	.0000138
_cons	-.2011753	.2087399	-0.96	0.337	-.6143969	.2120462

Figure 1. Regression analysis on championships_won and average_household_income

Now regressing *win_percentage* on *total_retail_sales_per_capita, population*, *average_household_income*, *median_age*, and *average_household_size*, I found that all of the variables combined are not statistically significant in figuring out win percentage because of a low F-value. Using an F-statistic of roughly 2.7 with a 95% confidence interval and the degrees of freedom for the numerator being 5 and the denominator being 107, there seems to be no statistical significance between win percentage and the other economic variables combined. On the other hand, when regressing *championships_won* on *total_retail_sales_per_capita, population*, *average_household_income*, *median_age*, and *average_household_size*, there seems to be a correlation between the variables because the p-value of the F-statistic is less than 0.05, being at .0057. Looking at the coefficients of each variable and keeping every other variable constant for every one unit increase in *total_retail_sales_per_capita, championships_won* decreases by .00000484, for every one unit increase in *population, championships_won* decreases by .0000000414, for every one unit increase in *average_household_income, championships_won* increases by .00001, for every one unit increase in *median_age, championships_won* increases by .07934, for every one unit increase in *average_household_size, championships_won* decreases by .327.

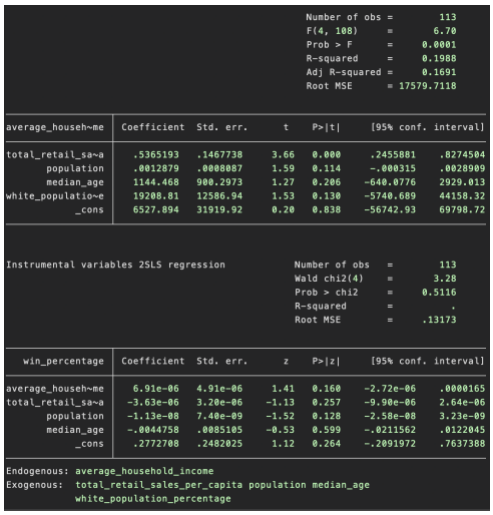
Source	SS	df	MS	Number of obs	=	113
				F(5, 107)	=	3.50
Model	7.19674299	5	1.4393486	Prob > F	=	0.0057
Residual	44.0244959	107	.411443887	R-squared	=	0.1405
				Adj R-squared	=	0.1003
Total	51.2212389	112	.457332491	Root MSE	=	.64144

championships_won	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
total_retail_sales_per_capita	-4.84e-06	5.50e-06	-0.88	0.381	-.0000157	6.06e-06
population	-4.14e-08	2.97e-08	-1.39	0.166	-1.00e-07	1.75e-08
average_household_income	.00001	3.48e-06	2.89	0.005	3.15e-06	.0000169
median_age	.07934	.0335502	2.36	0.020	.0128308	.1458493
average_household_size	-.3270768	.2638697	-1.24	0.218	-.8501677	.196014
_cons	-2.173332	1.211713	-1.79	0.076	-4.57541	.2287474

Figure 2. Regression analysis on championships_won and the economic factors

Instrumental Variables Regression

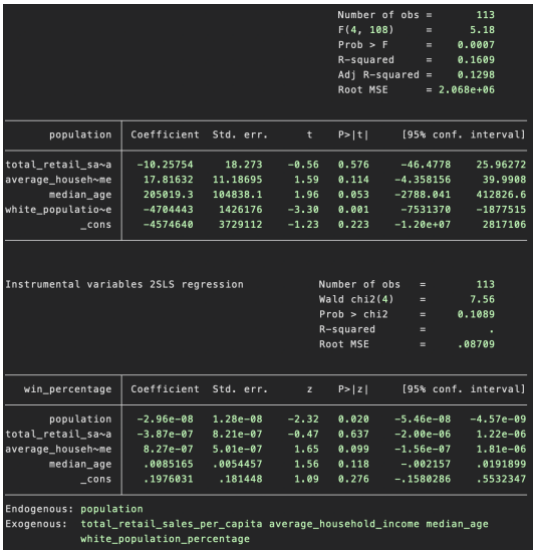
The introduction of instrumental variables in this type of experiment is important because they control and take into account the errors in this experiment. These errors could include the presence of heteroscedasticity, omitted variables, or other factors in regression that would make the data unusable. In the case of my problem of trying to connect a team's performance with its city's economic status, a key instrumental variable that could be used would be *male_percentage_population* in the respective city. I also tried to use *white_population_percentage* in the respective cities but trying to regress that as an instrumental variable for *average_household_income* deemed not statistically significant. I thought these two variables would work well together because of the positive omitted variable bias in the original economic variables of the white population in a city. Since white households hold about 80% of the nation's assets, one would assume that if there was a higher white population, then the *average_household_income* would also increase. However, this was not the case because the absolute value of the z-value was less than the z-statistic at 95%, 1.65.



```
                                          Number of obs =       113
                                          F(4, 108)     =      6.70
                                          Prob > F      =    0.0001
                                          R-squared     =    0.1988
                                          Adj R-squared =    0.1691
                                          Root MSE      = 17579.7118

------------------------------------------------------------------------------
average_househ~me | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
------------------+-----------------------------------------------------------
total_retail_sa~a |   .5365193   .1467738     3.66   0.000     .2455881    .8274504
       population |   .0012879   .0008087     1.59   0.114    -.000315    .0028909
       median_age |   1144.468   900.2973     1.27   0.206    -640.0776    2929.013
white_populatio~e |   19208.81   12586.94     1.53   0.130    -5740.689    44158.32
            _cons |   6527.894   31919.92     0.20   0.838    -56742.93    69798.72
------------------------------------------------------------------------------

Instrumental variables 2SLS regression          Number of obs   =       113
                                                Wald chi2(4)    =      3.28
                                                Prob > chi2     =    0.5116
                                                R-squared       =         .
                                                Root MSE        =    .13173

------------------------------------------------------------------------------
   win_percentage | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
------------------+-----------------------------------------------------------
average_househ~me |   6.91e-06   4.91e-06     1.41   0.160    -2.72e-06    .0000165
total_retail_sa~a |  -3.63e-06   3.20e-06    -1.13   0.257    -9.90e-06    2.64e-06
       population |  -1.13e-08   7.40e-09    -1.52   0.128    -2.58e-08    3.23e-09
       median_age |  -.0044758   .0085105    -0.53   0.599    -.0211562    .0122045
            _cons |   .2772708   .2482025     1.12   0.264    -.2091972    .7637388
------------------------------------------------------------------------------
Endogenous: average_household_income
Exogenous:  total_retail_sales_per_capita population median_age
            white_population_percentage
```

Figure 3. Failed instrumental variable regression analysis

After trying all of the economic variables with the two possible instrumental variables I had data on, I finally found a statistically significant instrumental variable. While regressing *win_percentage* on the different economic variables and setting *white_population_percentage* as an instrumental variable for *population,* I found that it is statistically significant and that the coefficient for population, -.0000000296, is estimating the returns to *population* for the percentage of white people in that population, so, in the end, it is influencing the race in the population, therefore saying that the higher white population in the city, the higher the population, which therefore means the lower the win percentage of the respective team.



```
                                          Number of obs =       113
                                          F(4, 108)     =      5.18
                                          Prob > F      =    0.0007
                                          R-squared     =    0.1609
                                          Adj R-squared =    0.1298
                                          Root MSE      = 2.068e+06

------------------------------------------------------------------------------
      population | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-----------------+-----------------------------------------------------------
total_retail_sa~a |  -10.25754     18.273    -0.56   0.576    -46.4778    25.96272
average_househ~me |   17.81632   11.18695     1.59   0.114   -4.358156    39.9908
       median_age |   205019.3   104838.1     1.96   0.053   -2788.041    412826.6
white_populatio~e |  -4704443    1426176    -3.30   0.001    -7531370   -1877515
            _cons |  -4574640    3729112    -1.23   0.223   -1.20e+07    2817106
------------------------------------------------------------------------------

Instrumental variables 2SLS regression          Number of obs   =       113
                                                Wald chi2(4)    =      7.56
                                                Prob > chi2     =    0.1089
                                                R-squared       =         .
                                                Root MSE        =    .08709

------------------------------------------------------------------------------
   win_percentage | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
------------------+-----------------------------------------------------------
       population |  -2.96e-08   1.28e-08    -2.32   0.020    -5.46e-08   -4.57e-09
total_retail_sa~a |  -3.87e-07   8.21e-07    -0.47   0.637    -2.00e-06    1.22e-06
average_househ~me |   8.27e-07   5.01e-07     1.65   0.099    -1.56e-07    1.81e-06
       median_age |   .0085165   .0054457     1.56   0.118    -.002157    .0191899
            _cons |   .1976031    .181448     1.09   0.276    -.1580286    .5532347
------------------------------------------------------------------------------
Endogenous: population
Exogenous:  total_retail_sales_per_capita average_household_income median_age
            white_population_percentage
```

Figure 4. Successful instrumental variable regression analysis

Results

Resulting of the previous two regression models, we have found that even though the economics variables are not necessarily statistically significant with a team's performance metric on their own with a t-statistic, there is some correlation between *championships_won* and all of the different economic variables because of the high F-statistic. These variables and the amounts changed are described in the OLS section of this report. Looking at other statistical values in this regression, there is a low R-squared, but that is not necessarily a bad thing. R-squared measures the goodness of fit in the model and you want to have as of an R-squared as possible, but in this case of our R-squared value being .1405, it is not

necessarily something that would ruin our experiment because our variables are statically significant regarding *championships_won.*

The instrumental variable regression gives us more information regarding *win_percentage* instead of that of *championships_won.* However, these are both successful measures of how well a team is performing, so it is okay that we had both present regarding both of these models. As one can see from Figure 4, we reduced the omitted variable bias by adding an instrumental variable of *white_population_percentage* for *population.* This forced us to have better metrics of what determines *win_percentage* with different economic factors. Looking at the coefficients of each variable and keeping every other variable constant for every one unit increase in *total_retail_sales_per_capita, win_percentage* decreases by .000000387, for every one unit increase in *population* with the instrument variable present*, win_percentage* decreases by .0000000296, for every one unit increase in *average_household_income, win_percentage* increases by .000000827, and for every one unit increase in *median_age, win_percentage* increases by .0085165. I have concluded from both models that there is a negative correlation between a team's performance and population, total retail sales per capita, and average household size and there is a positive correlation between a team's performance and average household income and median age.

Conclusion

To answer the question of whether economic factors in a city affect their team's performance on the field or court is pretty much up in the air as of now. I feel like I have answered some questions regarding this through my models. One of the things I found is that the higher the population one city has, the fewer championships it will have, and the higher the median age for one city the more championships that city will win. Of course, this experiment comes with a lot of limitations, including that each economic variable isn't statistically significant with the team's performance on its own. These variables didn't really seem to be correlated to a team's performance and we had to assume a lot of things in these experiments. To perform the OLS regression model, I had to assume that the model is linear in the coefficients and the error term, the error term has a population mean of zero, all of the economic variables are uncorrelated with the error term, observations of the error term are uncorrelated with each other, and there is no heteroscedasticity present. There was also a pretty small sample size given that I had to input all of the data by myself, whereas if I had an already imputed data set, this would have made it a lot easier to get thousands of pieces of data. There was the presence of some teams in Canada, whereas the majority of this experiment was supposed to happen in the United States. Given all of these assumptions and the possibility of omitted variable biases, this experiment was far from perfect. However, I still believe I found out some useful information regarding a team's success and how it is connected to the associated city's economic and demographic factors. Next time I perform this experiment I would change a couple of things, including adding more variables that are more relevant to the team's performance. I would even consider changing my research question entirely and having it more sports focused. I would try and regress each team's performance with player salary, fan engagement, team experience, injury rates, coaching changes, media exposure, team history, and other player or team metrics.