Causal Inference and Major League Baseball

Jamie Thornton

Department of Mathematical Sciences
Montana State University

May 4, 2012

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Jamie Thornton

This writing project has been read by the writing project advisor and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

Date

Steve Cherry
Writing Project Advisor

Date

Mark C. Greenwood
Writing Project Coordinator

1. Introduction

Statistics has been a integral part of baseball for many years. Over a decade ago, baseball managers started relying heavily on statistics to make on field decisions. At the beginning, many of these statistics were simple stats calculated by hand. How many runs did each batter hit in (RBI's), or how many times did each batter strike out? These easy statistics laid the ground work for some of the more complicated statistics that are calculated in major league baseball today. The simple stats still have plenty of value in the eye of the manager and team owner, but there have also been new additions and improvements over the years. For example, a relatively new statistic to the game, slugging percentage, looks at the total number of bases earned per at bat. (Sports Reference LLC, 2009)

While the goal of the players may be to win a World Series, the goal of the owner is to make a profit. The reality is that major league baseball is a business: a multi-billion dollar business. With the cost of operating a franchise continuing to rise, owners must be looking for a way to decrease costs and increase revenue. The most costly aspect of owning a major league baseball team is usually the salaries that the players earn. During the 2011 season the average salary for MLB players was just under $3.1 million, with the minimum salary at $414,000 (Associated Press, 2011). With 40-man rosters, it's clear that paying players can quickly cost a team over $100 million a year.

With salary costs soaring, teams are constantly searching for a way to balance the costs of the team with potential revenue. It makes sense that winning teams tend to sell more tickets, and thus make more revenue. Obviously, fans are much more likely to jump on the bandwagon and pay for tickets if they are going to watch a winning team. So the question becomes; how much is winning games worth? And more importantly, if owners do invest in more expensive players, will their investments be rewarded with more games won?

There are several examples of teams in major league baseball that spare no expense when it comes to having the highest paid players. The New York Yankees are a prime example of a team that might pay for their wins. In 2011, for the 13th season in a row, the Yankees sported the highest league payroll. But the Yankees also sported 97 wins, which was the best record in the American League (ESPN, 2011). But, there are also examples of teams with small payrolls winning a lot of games. For example, the Tampa Bay Rays made it to the World Series in 2008 with a payroll of only $43.8 million, which was second-to-lowest in the league that year. So the real questions is, does a higher payroll lead to more wins?

Through-out this paper, I will be investigating work done by Derek Stimel (Stimel, 2011) as he looked at the dependence relationships between payroll and winning percentage in major league baseball. Stimel not only investigated how payroll affected the winning percentage of a team, but also how the winning percentage affected payroll. For example, after the Rays were in the World Series in 2008, the 2009 payroll for the team jumped $20 million dollars from the previous year to $63.3 million (Kendrick, 2011). Was this jump a direct result of the success the previous year? Should the owners expect this increase in payroll lead to continued success? These are all questions that we will investigate.

2. Possible Relationships

In the beginning of the paper by Stimel (2011), it is noted that there are three potential relationships that could exist between a team's payroll and their winning percentage. First, the payroll of a team could affect the winning percentage. We would imagine that this is a positive relationship, meaning that an increase in payroll should lead to an increase in winning percentage. This makes sense in the context of major league baseball where a player is supposedly paid according to their skill and abilities. It would stand to reason that if you are spending more on your payroll, then you should be paying for more skilled players; and more skilled players should lead to a higher winning percentage.

The second relationship that could exist would be where an increase in winning percentage leads to an increase in payroll. More success is likely to bring not only more fans to the ball park, but also increase revenue through TV deals, merchandise, advertising, etc. A team that is making more money is likely to spend more of that revenue on players, thus increasing the team payroll. The team would spend more money on what they hope are more skilled players, which would hopefully continue to keep the winning percentage high.

The last relationship that is possible between payroll and winning percentage is that they are both influenced by on field performance. A team that performs better on the field should win more games, leading to a higher winning percentage. But if a team is performing better on the field, then some individual players are likely performing better, and with better performance comes higher salaries, which would increase payroll. It is common practice for contracts to be renegotiated, especially after a player does particularly well. If a team suddenly does better on the field than it has previously, it would make sense that the players on the team that helped produce the increase in wins would want to be compensated for their production. This could easily lead to an increase in payroll.

3. Data

To investigate which of these situations is more likely, Stimel used the PC algorithm and graph theory to look at the directional relationships between payroll, winning percentage, and on field performance. Before we delve into what the PC algorithm is actually doing, we will investigate the choice of variables. Stimel used 29 variables in an effort to describe the on field performance of each team. All the variables were taken at the team level and recorded for 21 seasons between 1985 and 2009. The variables were separated in to 5 basic categories; overall variables, batting variables, fielding variables, pitching variables and base running variables.

The overall variables included winning percentage and payroll. Winning percentage is defined as the number of games won during a season divided by the total number of games played. Each team plays 162 regular season games with up to 19 additional games for the playoffs. Payroll is defined in millions of dollars, so a payroll value of 75.4 would correspond to a team having a payroll of $75.4 million that season. All the payroll numbers were inflated to represent 2008 dollar value (Stimel, 2011).

The batting variables included at bats, batting average, walks, batter park factor, singles, doubles, triples, home runs, total bases, runs on base percentage, slugging percentage and strikeouts. Again, all of these variables are at the team level; for example at bats represents the total number of at bats that a team had during a season. Batting average is the total number of hits divided by total at bats, i.e. the average number of hits per at bat. Walks, singles, doubles, triples, home runs, and strikeouts are simple counts of how many of each type of outcome resulted from a plate appearance. Total bases counts how many total bases a team achieved in a season. For example, a double that was moved over on a sacrifice fly would get 3 total bases; 2 for the double and 1 for the sacrifice fly. Slugging percentage, as explained before, represents the average number of bases per hit. The last variable, batter park factor, tells us if the team plays in a hitter-friendly park or not. If a team plays in a hitter-friendly park it may not only affect the number of hits or home runs the team achieves, but also the type of player they are willing to pay for. A team with a hitter-friendly park may be more willing to pay large amounts of money for a proven hitter, than a team that is not in a hitter-friendly park (Stimel, 2011).

The only fielding variable included was fielding percentage. Fielding percentage is calculated by dividing the total number of outs completed by a team by the total number of opportunities they had to get an out. For example, when a team commits an error and does not throw a runner out at first base, that is considered a missed opportunity for an out. The fielding percentage represents the average number of outs a team gets per opportunity (Stimel, 2011).

Pitching variables include walks against, complete games, earned runs, earned run average, hits against, home runs against, innings pitched, pitching park factor, runs against, strikeouts against, and saves. Walks against, complete games, hits against, home runs against, runs against, and strikeouts against are simple counts; for example, hits against represents the total number of hits opposing teams got against a team. Earned runs are runs scored by non-defensive errors. This means that the runner had to get on base by a hit and be scored by a hit. Earned run average is the average number of earned runs per nine innings, i.e. earned runs divided by total innings played times nine. Innings pitched represents the total number of innings the pitching staff pitched in a season. Saves are the total number of games relief pitchers finished where the team had the lead when the relief pitcher came in and went on to win the game. Finally, pitching park factor is a lot like batter park factor, where we have a variable to represent how friendly a teams home field is to pitchers. Again, if a team plays in a pitchers park they might be more willing to pay money for a good pitcher (Stimel, 2011).

The last category, base running variables, contains 2 variables; caught stealing and stolen bases. Caught stealing is a count of how many times a team had a runner thrown out as they were attempting to steal a base. Stolen bases is a count of how many times a team was successful in stealing a base (Stimel, 2011).

All variables, except those that are already on a percentage scale, were transformed by taking the natural log. The variables that were not transformed include: winning percentage, batting average, on base percentage, slugging percentage, and fielding percentage. I believe this transformation was used to decrease the wide range that would have been present in some of the variables with out the transformation (Stimel, 2011).

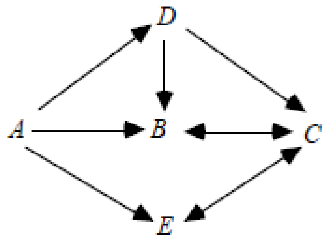
4. PC Algorithm

Stimel's goal in his paper was to establish cause and effect relationships between the aforementioned variables. However, it is well known that causal inferences in observational studies are problematic in statistics. In the book *Causation, Prediction, and Search* the authors Peter Spirtes, Clark Glymour and Richard Scheines note that while it is true that most statistical methods cannot be used to draw cause and effect inferences, that is exactly what these methods are often used to do. They note that while it is stated that linear regression can only be used as a means of fitting a line to data and predicting new values, it is often used "to predict values of a variable when the regressors are manipulated, that is, when action or policy forces some novel distribution on the regressors," (Glymour et al., 2001, pg. 1).

This inherent battle between what statistics can tell us and what researchers actually want to know lead Spirtes, Glymour and Scheines to develop several new algorithms that help show where cause and effect relationships actually are present. The PC algorithm that Stimel used in his paper is one such algorithm. Stimel was hoping to use the PC algorithm to investigate whether payroll depended on winning percentage, if winning percentage depended on payroll or if they were both caused by on field performance.

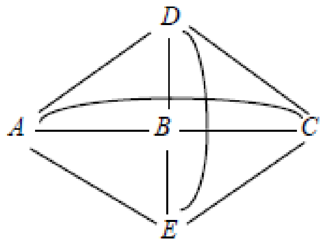
The PC algorithm is based on the theory behind independence relationships and directional acyclic graphs, otherwise known as DAGs. There are 6 situations in which a relationship, or "statistical dependency," between variable X and Y can be observed from a sample of the variables. This relationship can be detected when: X causes Y; Y causes X; X and Y cause each other; a third variable causes both X and Y; the sample is not representative; or when the values of X and Y form time series. Experimentation is generally accepted as the only option to separate the most common and interesting of cases, when X causes Y, Y causes X, they cause each other, or some third variable causes them both (Glymour et al., 2001 pg. 22). However, when other variables are measured we may be able to infer causal relationships if we can assume that all variables that affect the system are included in our data set (Glymour et al., 2001).

Graph theory can be used in many statistical situations to help illustrate the relationships between the variables. The graphs consist of vertices, **V**, and edges, **E**. The vertices represent the variables that are in play in a given problem. The edges connect the vertices or variables. There are several types of edges that can connect 2 vertices. Directed, bi-directed, undirected and partially directed edges are possible. In this paper we will not discuss partially directed edges. The plot below shows an example of several different types of edges (Glymour et al., 2001).



In the graph above, you can see directional edges between verticies A & D, A & B, A & E, D & B, and D & C. Bidirectional edges are found between B & C and E & C. There are no undirected edges in this graph, but those are simply edges with no arrow heads. A directed edge from A to B (arrow pointing toward B) is said to be out of A and into B. It is also considered that A is the parent of B and B is the child of A. In the graph above, B would be a collider because there are two arrows that meet head on at B. E would also be a collider for the same reason (Glymour et al., 2001).

Also, a graph is considered complete if every vertex is connected to every other vertex. The graph below is a complete undirected graph: complete because each vertex is directly linked to every other vertex, and undirected because all the edges are undirected (Glymour et al., 2001).

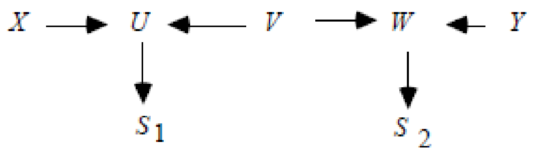


These graphs can prove very useful when it comes to visualizing the relationships between variables. A graph could have a vertex for each variable in the problem, then appropriate edges could be used to indicate the relationship between variables. If it was know that A caused B then we would expect to see a directed edge out of A into B. If A and B caused each other, then a bi-directional edge could be used between the two vertices. If other variables caused both A and B, then there could be directed edges out of those other variables into A and B, with no edge directly linking A and B. Thus, these graphs can be a useful tool for visualizing the relationships between variables.

Before we can apply the idea of directional acyclic graphs to our questions about the effects of winning percentage and payroll in major league baseball, we need to understand how edges in the graphs become directed. The graphs use standard definitions of independence between any 2 variables. If the joint distribution of the variables can be expressed as a product of the individual variables, then they are considered independent. For example, variables X and Y would be considered independent if the joint density of X and Y was equal to the product of the density of X and the density of Y. Furthermore, we can say that X and Y are independent conditional on Z when the density of X and Y given Z equals the product of the density of X given Z and the density of Y given Z (Glymour et al., 2001). Obviously, in most situations the true probability distributions for the variables are unknown, and so data must be used to estimate these distributions.

The PC algorithm was developed to efficiently detect the needed edges in a graph and find proper directions for these edges. The PC algorithm starts with a complete undirected graph by connecting all the variables in the problem. Next, the algorithm iteratively attempts to remove unnecessary edges in the graph by testing pairs of adjacent vertices for d-separation (see below for more on d-separation). Adjacent vertices are defined as vertices that are directly connected by an edge. So in the beginning, when we have a complete undirected graph, each variable will be adjacent to every other variable (Glymour et al., 2001).

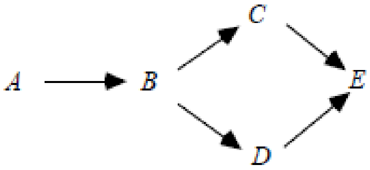
D-separation is a way of looking for independence and conditional independence relationships. Variables X and Y are said to be d-separated by Z if there is no path from X to Y without traversing a collider or traversing a member of Z. The exception to these rules are that if a collider has a child in Z, then the path from X to Y can traverse the collider without being d-separated. It's important to note that when we consider paths from X to Y for the sake of determining d-separation we don't need to consider the direction of the edge (Pearl, 2009).



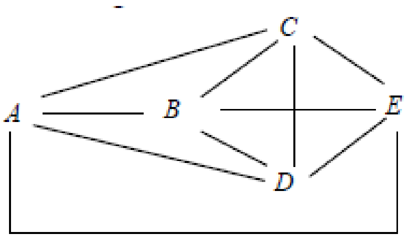
In this graph from Glymour et al. (2001), X and V are d-separated because there is no path from X to V without traversing U, which is a collider. Because we don't need any variables in the set Z to condition on to declare this d-separation, we say that X and V are d-separated given the empty set. X & S1, X & W, X & Y, and X & S2 are all d-separated given the empty set because any path from X to another variable would need to traverse the collider at U, which can't be done. If we considered Z to include S1, then X and V are no longer d-separated; we say they are d-connected. Because S1 (a child of the collider U) is in Z, we can now traverse U in our path from X to V. While X and V are now d-connected, X and Y are still d-separated because the collider W is still in the path from X to Y; we would say X and Y are d-separated given S1. If the set Z included S1, S2, and V, then X and Y would be d-separated given the set $\{S1, S2, V\}$. Even though we can now traverse the colliders U and W because they each have a child in the set Z, we cannot find a path from X to Y that does not traverse a member of Z, namely V. There is no way to move from X to Y without passing through V.

When n=0, the algorithm attempts to determine any pairs of variables that are d-separated given the empty set. This will only occur when there are colliders present in the true graph. When n=1, the algorithm looks for pairs of variables that are d-separated given 1 other variable. When n=2, the algorithm looks for pairs of variables that are d-separated given 2 variables. This continues until all unnecessary edges have been removed from the complete undirected graph.

Let's consider an example. Suppose we were considering a problem with 5 variables: A, B, C, D and E. Suppose the true directional acyclic graph (Glymour et al., 2001) was:

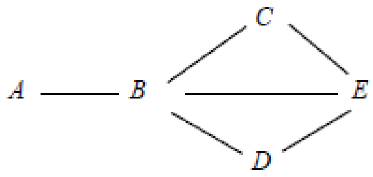


Using the PC algorithm, we would start with the complete undirected graph, shown below.

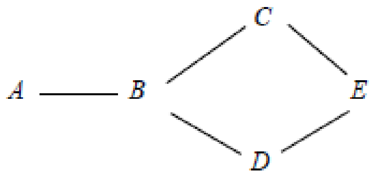


When n=0, the PC algorithm would not remove any edges because there are no pairs of variables that are d-separated given the empty set. There is one collider in the true graph, vertex E, but there are no pairs of variables that are d-separated given the empty set because there is another path between all the variables that does not involve traversing E. The only likely variables that could be d-separated given the empty set are C and D, but there is a path from C to D through B so that you are not required to traverse the collider at E. So when n=0, all variables are still d-connected.

When n=1, the PC algorithm would find several pairs of variables that are d-separated given one other variable. A and C are d-separated given B. There is no path from A to C that does not pass through B. Because these variables are d-separated the PC algorithm would remove this edge of the graph. The algorithm would also find that A & E, A & D, and C & D are all d-separated given B. These represent all the d-separated relationships in the true graph that are conditional on only one variable. After this step, the PC algorithm would remove the edges between the variables that are determined to be d-separated. The graph would then look like the following (Glymour et al., 2001):



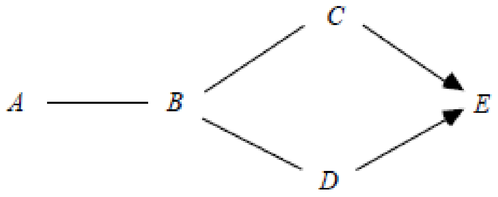
When n=2, the PC algorithm is looking for pairs of variables that are d-separated given sets of 2 other variables. The only relationship in this example that meets this requirement is B and E, which are d-separated given the set $\{C, D\}$. There is no path from B to E that does not pass through either C or D. The PC algorithm would also remove this edge, resulting in an undirected graph that has all the correct edges (Glymour et al., 2001):



At this point the iteration process would stop because there are no more remaining pairs of adjacent variables where there are more than 2 other adjacent variables, so the n=3 step is not needed.

The final step of the PC algorithm is to attempt to correctly orient the edges in the simplified undirected graph so that the edges match the true relationships. We can follow several rules to orient the edges. For each group of three vertices, X, Y, and Z, such that X & Y and Y & Z are adjacent in the reduced graph but X & Z are not adjacent, orient the edges $X \to Y \leftarrow Z$ if and only if X & Z are not d-separated given Y (Glymour et al., 2001). This means that if there are three variables so that X and Y are connected and Y and Z are connected, but X and Z are not connected, the edges will both point to Y as long as X and Z are not d-separated given Y. This is because we don't want a situation where (1) X and Z are d-separated given Y, and then (2) we make Y a collider, as that would eliminate the need for Y to separate X and Z. If Y was a collider, then X and Z would be d-separated given the empty set, not Y. Following a similar thought process, we can continue to orient edges such that if $A \to B$ and B & C are adjacent but A & C are not adjacent, then the edge between B and C should be oriented such that $B \to C$.

In our example, there are 6 groups of 3 variables that satisfy the conditions above: A-B-C, A-B-D, C-B-D, B-C-E, B-D-E, and C-E-D. One such group, C-E-D, allows us to orient the edges $C \rightarrow E \leftarrow D$ because C and D are not adjacent but they are not d-separated given E. The other triples, such as A-B-C, do not allow us to orient their edges because A and C are d-separate given B. At this point we have no way of knowing which way the edges between A & B and B & C should be oriented. The final graph the algorithm would produce is (Glymour et al., 2001):



This graph is close to the true DAG, but it's not exactly right. There are several big assumptions that are made when the PC algorithm is applied to a data set. The first assumption is that all variables that have an effect on the system of interest have been measured. The second assumption is that every unit in the population follows the same causal relationships. In the baseball study, this means that each team follows the same relationship between payroll and winning percentage. The final assumption is that the algorithm correctly finds the graph, no needed edges were removed and no non-needed edges were left in the graph (Glymour et al., 2001).

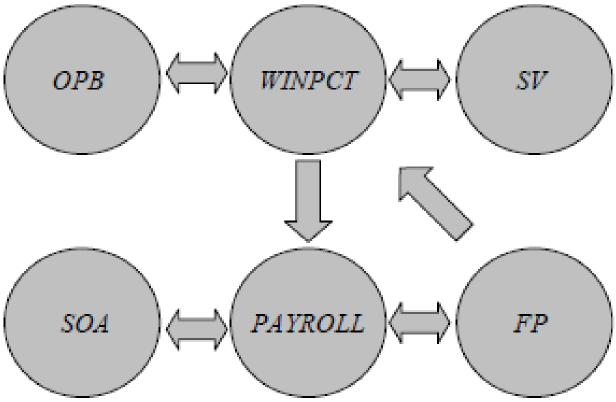
Clearly the PC algorithm is not a perfect system. It can be very difficult to assume that every variable that has an effect on the system has been measured. In addition to the assumptions, the algorithm has trouble directing all the edges, even if it does detect the correct graph. Also, if you have data samples that do not adequately represent the true underlying probability distribution for each variable, then the PC algorithm could seriously misjudge the true relationships that are present. If a needed edge is removed or a non-needed edge is left in the graph it can seriously alter the relationships represented by the graph.

5. Results

Before Stimel applied the PC algorithm to his data set, he realized the need to remove the time dependence in the data. All of the variables were at the team level and most likely depended heavily on time as a covariate. For example, a team's batting average is most likely dependent on their batting average from the previous year. Teams are likely to have many of the same players that they did the previous year and it's likely that these players will perform similarly to how they performed the year before. Using two information criteria, Akaike Information Criterion (AIC) and Schwarz Criterion (SC), Stimel found that a one year lag was the best option to remove any time effect before applying the PC algorithm. Stimel was also worried about single team effects, such as the Yankees with their unusually high payroll, unfairly influencing the results. To adjust for individual team effects, Stimel used a panel vector autoregression with fixed effects to filter the variables before the PC algorithm was applied (Stimel, 2011).

When using the PC algorithm, Stimel chose to use a ”less restrictive” alpha level of 10% for all of the independence (or d-separation) tests the algorithm used. Stimel noted that this choice of alpha could lead to a slightly different structure than an alpha level of 1% or even 5%, but hopefully not significantly different. A significance level of 10% allows more edges to stay in the graph than would remain at a 5% or 1% significance level, because you are less likely to fail to reject the null hypothesis that variables X and Y are d-separated at the 10% level. Stimel also chose to orient any undirected edges left by the PC algorithm as bi-direction, meaning that he assumed that if the algorithm was unable to orient the link then the 2 variables depended on each other, or more likely on a third unmeasured variable (Stimel, 2011).

After applying the algorithm, Stimel found a number of interesting dependency relationships. Because the main goal is to discover the relationship between payroll and winning percentage, only that portion of the graph the PC algorithm produced is shown below (Stimel, 2011).



From the graph above, we can see that payroll is dependent on winning percentage but winning percentage is not directly dependent on payroll. Winning percentage is only dependent on payroll through it’s effect on fielding percentage (FP). Winning percentage is dependent on fielding percentage, and has bi-directional relationships with on base percentage (OBP) and saves (SV), meaning that they depend on each other. Payroll is dependent on winning percentage and has bi-directional relationships with fielding percentage and strikeouts against (SOA), again meaning that they depend on each other (Stimel, 2011).

Now that the dependence relationships have been established, we would really like to know how strong these relationship are. How does payroll respond to a change in winning percentage? These are the real questions that get at the heart of the matter.

Using the causal relationships established by the DAG, Stimel fit the regression model

$$WINPCT = \beta_0 + \beta_1 FPCT + \beta_2 OBP + \beta_3 SV + \epsilon$$

Stimel fit several different versions of this model. One version included the one year time lags, while another version included only fixed effects. A third version included both fixed effects and lags, and the final version had neither lags nor fixed effects. In the table below we can see the results of the 4 models (Stimel, 2011).

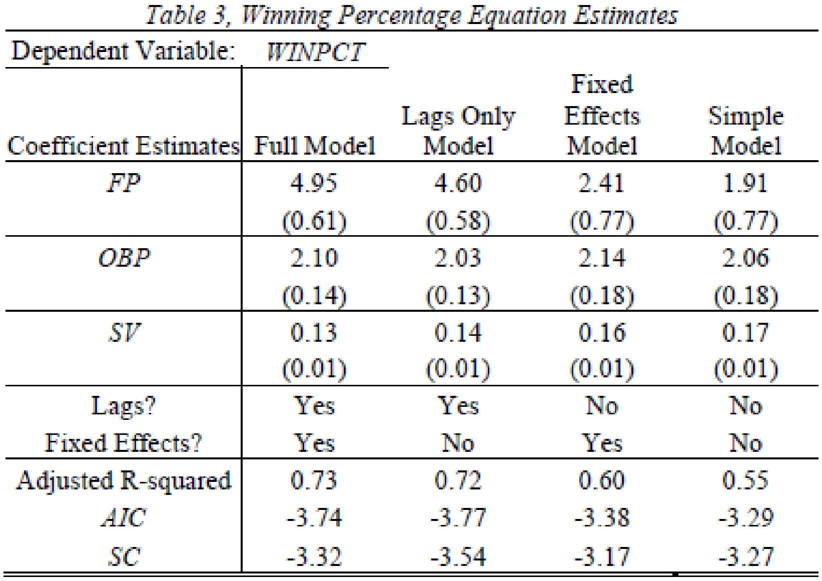


Table 3, Winning Percentage Equation Estimates

Dependent Variable: *WINPCT*				
Coefficient Estimates	Full Model	Lags Only Model	Fixed Effects Model	Simple Model
FP	4.95	4.60	2.41	1.91
	(0.61)	(0.58)	(0.77)	(0.77)
OBP	2.10	2.03	2.14	2.06
	(0.14)	(0.13)	(0.18)	(0.18)
SV	0.13	0.14	0.16	0.17
	(0.01)	(0.01)	(0.01)	(0.01)
Lags?	Yes	Yes	No	No
Fixed Effects?	Yes	No	Yes	No
Adjusted R-squared	0.73	0.72	0.60	0.55
AIC	-3.74	-3.77	-3.38	-3.29
SC	-3.32	-3.54	-3.17	-3.27

It is interesting to note that while the AIC values for all four models are very close (all within 0.5 of each other), the coefficient estimates on some of the variables vary quite a bit. While the coefficient estimates for on base percentage and and saves are relatively stable, the coefficient estimates for fielding percentage range from 4.95 from the full model to 1.91 from the simple model. This variation in coefficient estimates despite similar AIC values could imply a lack of robustness in the estimation process. This is not an issue that Stimel discusses in his paper.

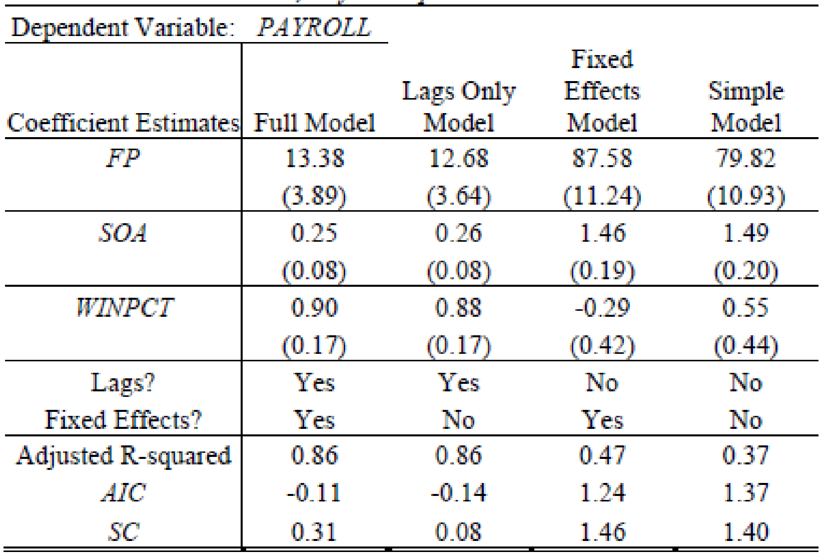
Stimel focused on an interpretation of the coefficients from the lags only model because this model had the lowest AIC value. Using this model, we can interpret the coefficients as follows: the coefficient estimate for fielding percentage of 4.6 means that a 1 point increase in fielding percentage would lead to a 4.6 point increase in winning percentage. For example, if a team had a .500 winning percentage and a .95 fielding percentage, then increases their fielding percentage to .96, we would expect the winning percentage to increase to .546. The coefficient estimate of 2.03 for on base percentage means that a 1 point increase in on base percentage leads to a 2.03 point increase in winning percentage. For example, a team with a .450 on base percentage and a .500 winning percentage could expect their winning percentage to increase to .520 if they increased their on base percentage by 1 point to .451. Finally, the coefficient estimate of 0.14 for saves means that a 10 percent increase in saves in a season would lead to a 1.4 point increase in winning percentage. For example, a team that increased their saves from 30 to 33 could expect their winning percentage to raise from .500 to .514 (Stimel, 2011).

A similar method was used to fit a model for the payroll variable. From the causal graph, we found that the dependent relationship was

$$PAYROLL = \beta_0 + \beta_1 FP + \beta_2 SOA + \beta_3 WINPCT + \epsilon$$

Again, fitting this model using four different methods (both lags and fixed effects, lags only, fixed effects only, and a simple model with no fixed effects or lags), we see the following coefficient estimates (Stimel, 2011).

Table 4, Payroll Equation Estimates



Dependent Variable: *PAYROLL*

Coefficient Estimates	Full Model	Lags Only Model	Fixed Effects Model	Simple Model
FP	13.38	12.68	87.58	79.82
	(3.89)	(3.64)	(11.24)	(10.93)
SOA	0.25	0.26	1.46	1.49
	(0.08)	(0.08)	(0.19)	(0.20)
WINPCT	0.90	0.88	-0.29	0.55
	(0.17)	(0.17)	(0.42)	(0.44)
Lags?	Yes	Yes	No	No
Fixed Effects?	Yes	No	Yes	No
Adjusted R-squared	0.86	0.86	0.47	0.37
AIC	-0.11	-0.14	1.24	1.37
SC	0.31	0.08	1.46	1.40

Again, Stimel focused on an interpretation of the coefficients from the lags only model because this model had the lowest AIC value. Using this model, the coefficient estimate for fielding percentage of 12.68 means that a 1 point increase in fielding percentage would lead to a 12.68 percent increase in payroll. For example, a team with a $75 million payroll would expect their payroll to jump to $84.51 million if the fielding percentage jumped 1 point. The coefficient estimate on strikeouts against of 0.26 means that a 10% increase in strikeouts against would lead to a 2.6 percent increase in payroll. Finally, the coefficient estimate of 0.88 on win percentage means that a 10 percent increase in winning percentage would lead to an 8.8 percent increase in payroll (Stimel, 2011).

It is important to note again that these coefficient estimates vary greatly among the different models even though the AIC values are fairly similar. The AIC values range from -0.11 for the full model to 1.37 for the simple model. The coefficient estimate for fielding percentage is especially variable, with estimates ranging from 87.58 from the fixed effects model with no lags to 12.68 from the lags only model. An estimate of 87.58 for the fielding percentage coefficient would mean that a 1 point increase in fielding percentage would lead to an 87.58 percent increase in payroll, which is clearly too high. It is most likely that in this situation, there are other variables affecting the causal relationship that are not present in this model (Stimel, 2011).

6. Conclusions

This study introduced me to several new and interesting aspects of statistical analysis of causal relationships. Causal graphs and the PC algorithm were both new to me. While there are several big assumptions that go along with the algorithm, the ability to draw causal inference is an appealing advantage. A carefully planned study could have the ability to meet the assumptions of the algorithm, i.e. that all units in the population follow the same relationships and that all needed variables are measured.

I think the application of the PC algorithm and causal graphs to the relationship between payroll and winning percentage is something that could be applied to many areas in sports. For example, it would be interesting to investigate the effect of number of years of college on the length of careers for NFL and NBA players; or the effect of concussions on the lifespan of NFL and NHL players. Possibly the most interesting area of application I can think of would be the relationship between on field success and academic success for NCAA sport programs. Do those programs that neglect the student side of student-athletes have more success? In addition to these few areas of interest, I'm sure there are ample opportunities to expand this research to other sport-related questions, as well as other areas of statistical investigation.

References

Associated Press. (Dec. 6, 2011) Average baseball salary rises to $3.1M. http://espn.go.com/mlb/story/_/id/7319810/major-league-baseball-average-salary-increases-31-million.

Buhlmann, P. and M. Kalisch. (Feb. 2, 2008) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. http://arxiv.org/pdf/math/0510436.pdf.

Donnachie, E. (July 14, 2006) Graphical Models and the PC Algorithm. http://uni-leipzig.de/ strimmer/lab/courses/ss06/seminar/slides/ewan-2x4.pdf.

ESPN. (2011) MLB Standings - 2011 [Data File]. http://espn.go.com/mlb/standings/_/year/2011/seasontype/2.

Glymour, C., R. Scheines, and P. Spirtes. (2001) *Causation, Prediction, and Search. Cambridge, MA: MIT Press.*

Kendrick, S. (Dec. 31, 2011) 2009 Baseball Team Payrolls: Total Salaries for Major League Teams. http://baseball.about.com/od/newsrumors/a/09teamsalaries.htm.

Pearl, J. (2009) d-Seperation Without Tears. http://bayes.cs.ucla.edu/BOOK-2K/d-sep.html.

Sports Reference LLC. (Aug. 16, 2009) Slugging Percentage. Accessed March 14, 2012, http://www.baseball-reference.com/bullpen/Slugging_percentage.

Stimel, Derek S. (2011) Dependence Relationships between On Field Performance, Wins, and Payroll in Major League Baseball. Journal of Quantitative Analysis in Sports: Vol. 7: Iss. 2, Article 6.