Resolving Simultaneity Bias: Using Features to Estimate Causal Effects in Competitive Games

Anders Harboell Christiansen
Independent Researcher
Lyngby, Denmark
andershc1993@gmail.com

Emil Gensby
Independent Researcher
Lyngby, Denmark
emil@gbcnet.dk

Bryan S. Weber
Department of Economics
College of Staten Island: CUNY
Staten Island, New York
bryan.weber@csi.cuny.edu

***Abstract*—In this paper, we walk through an application of the instrumental variables (IV) and control function approach (CF) estimators to identify the causal effect of in-game features on the probability of winning. Without use of IV or CF, typical methods of identifying the effect of in-game features (like player performance) will be strongly biased. Practical uses of unbiased estimates include predictive game balance when introducing new heroes and features, enhanced estimation of player rankings, and better interpretation of individual player statistics.**

***Index Terms*—Instrumental Variables, Control Function, Causal Effects, Dota, Matchmaking, Winning Probability**

I. Introduction

Many players and game developers are particularly interested in observing the causal effect of player performance within games on their probability of winning. For example, a game creator may be interested in releasing a new item to the game, increasing damage. Game developers are naturally concerned with estimating the game balance consequences. In addition, matchmaking services for games are often keyed to player's performance within the game beyond simply winning, such as in Overwatch [1], [2]. However, the causal effect of these features are likely estimated by a variety of naive approaches, and as a result the matchmaking system will be biased. In fact, Overwatch matchmaking eventually removed this system for advanced players [3].

Continuing this example, the Overwatch League frequently displays head-to-head statistics of players on opposing teams after a match. However, those statistics are often misread as the cause of victory rather than a symptom of victory. For example, even a weak player on a winning team is likely to have died very little and generated many kills relative to their counterparts on the opposing team. As a result, it is difficult to display these statistics in a meaningful way and they often feel random or misleading [4].

As we demonstrate in this paper, naive estimations will produce inaccurate results that are often biased strongly in the wrong direction. We review the theoretical background of this problem, and explore an application to identify the causal effects of changing critical features in a popular game, Defense of the Ancients 2 (Dota 2). This discussion is important in the context of computer games, where computer scientists naturally tend towards machine learning (ML) approaches. ML approaches emphasize the quality of the prediction, but the estimated causal effect of changing features (such as improved individual player performance, or game designers altering the mechanics) can be strongly biased [7].

978-1-7281-1884-0/19/$31.00 ©2019 IEEE

II. Causal Effects

In this case, we are interested in *causal inference*, estimating how a change in a particular feature will lead to a change in outcomes. However, both causation and correlation can be useful for making accurate predictions, causing the two to be frequently confused.

Consider the case of a coaching method being applied to a team in a gambling-heavy sport. Coaches would be interested in the causal effect of the coaching method. A coach, therefore, would prefer a classic randomized experiment with a treatment and control group to determine if a coaching method is effective. On the other hand, if the coaching method is publicly used, gamblers would be interested in predicting the outcome of any game, and can earn money with a strong and consistent correlation, perhaps one identified with standard ML techniques.

Like coaches in the above analogy, game developer's interests are primarily on the causal effect of the game features. They hope to make inferences about game balance and match quality using the actual effect of the features on the outcomes. Furthermore, the outcomes of any matches, and any non-causal, purely correlative associations between the features and the match outcome are largely uninteresting.

For this specific purpose, measures of fit like R^2, which measure the predictive power of the features, are uninformative. We are not particularly interested in what portion of variation in the outcome variable is being predicted, rather we are interested in what portion of that variation is caused by the features we have selected. However, the average partial effect of the features are very important [5], such as those measured by the slope coefficient of ordinary least squares regression (OLS). For our paper, we are going to provide an example where the actual causal effect will not properly be measured by standard techniques (they are biased). This problem will be exemplified by OLS because of its broad universality, but persists in random forests [6] and other ML techniques that are considered standard [7]. In this paper, we highlight both the *instrumental variables* and *control function approach* from

the econometric literature as ways to resolve this problem. To accomplish this goal, we will need a complete data set which exposes how common the problem is, and how to resolve it. For this purpose, we have collected a data set of Dota 2 games using OpenDota API [8].

A. Dota 2

The 2013 video game, Defense of the Ancients 2 (Dota 2), is a five versus five game featuring a wide range of playable heroes. During the game, the players kill units, obtaining "last hits" (LH) for gold and experience. They also kill the enemy team's heroes and towers for resources. The primary goal is to destroy the enemy's main base, which is hidden behind at least three towers. Once a team succeeds in this task, they win and the match concludes.

In a typical game, after the matchmaking process has been completed, each player selects a hero from a list with a wide variety of abilities (with no duplicates). Each hero has a primary attribute - strength (str), agility (agi) and intelligence (int) - which denotes the general hero role in the game, with some exceptions. Typically, str heroes generally do less damage and tend to have a large number of hit points. Agi heroes do the most damage over time and are strongest during late game. Int heroes provide powerful abilities and are an important source of instant damage. For example, a team of two str, two agi, and one int hero is a balanced team. All Random (AR) games in Dota 2, however, bypass the hero selection process and randomly assign a hero to each player, almost like an experiment. AR games create suboptimal teams - for example, a team of four int heroes and one str hero would have trouble accomplishing several important tasks in the game.

In our estimation process, we will be exploiting the random assignment in AR games to make inferences about how game features cause winning in Dota 2 in general. These game features will typically be measures of player performance (last hits, kills, etc.). We do not have interest in balance within AR games – the games themselves are clearly designed to be unbalanced, a property which we intend to exploit as if it were a mock experiment. We focus on AR games because random assignment of heroes will dramatically alter a team's in-game features, but we can be confident this variation will not be caused by behavioral changes over the course of winning a game. Instead, the variation in features will be created by a (presumably) random assignment of heroes. We then estimate the causal impact of changes in game features deriving exclusively from the randomized change in hero composition, and not from any other source.

III. Data

Data was acquired by issuing requests to the OpenDota API [8] via Python. The procedure for gathering data is as follows:

1) Request a list of match IDs and their game mode.
2) Save the match IDs if the game mode tag is AR.
3) Request the match data for all the AR matches by using the match IDs.
4) Lastly, pull out the specified statistics from each match and insert them into a CSV file that can be used for further experimentation.

Each match has team averages per minute for a wide range of gameplay related statistics, which we will be using going forward. We collect the following statistics: average kills per minute, average last hits per minute, average tower damage per minute, the largest premade party size, and the number of heroes belonging to each of the three primary attributes. We have also collected the matchmaking ranking (MMR) difference between the teams. We note that the MMR we have found, when players choose to make it public, is for competitive games where heroes are *chosen* by the player, and not AR games. To obtain this MMR, we took each player rank (when public) and converted it to MMR via [9]. We averaged the MMR of all public profiles on a team, and if a team had no public profiles the match was discarded. For this paper we have gathered a total of 3342 matches with complete information, for a total of 6684 observed wins and losses.

TABLE I
Data Summary: Player Averages for Both Teams

Statistic	Mean	St. Dev.	Min	Max
won	0.500	0.500	0	1
$kills/min$	0.184	0.070	0.000	0.511
lh/min	3.517	0.940	0.663	7.811
$towerdmg/min$	63.220	51.021	0	259
$largestpartysize$	2.571	1.087	1	5
str	1.617	1.020	0	5
agi	1.602	1.026	0	5
int	1.781	1.043	0	5
mmr_diff	0.000	736.878	−3,360	3,360
N	6,684			

Our total table provides 6,684 observed wins and losses. Since we observe both sides of every match, exactly half of the teams in our dataset win and the other half have lost. The statistics have been averaged per player (5 players per team) over the length of the game (typically about 40 minutes).

On average, each player has an average of 0.184 kills/min. Some teams failed to kill any of the opposing team, resulting in a minimum of 0 kills/minute. Each player also average about 3.5 last hits/minute as a primary source of income, out of a maximum theoretically possible 8 creeps/minute.[1] The tower damage hovers around 63 tower damage/minute. Over the course of the game, the maximum possible tower damage is 23,400 – but players can win by taking a direct path to the enemy for as little as 7,800 total tower damage.

Random mode is typically a casual mode filled with relatively large parties (average size of about 2.5). There is no "all random" Dota 2 publicly available MMR, but the players with public profiles suggest the teams are typically of diverse skill sets for standard competitive Dota 2. Lastly, the distribution of the three hero attributes (str, agi, int) is almost perfectly even

[1]The number of creeps spawned increases as the game goes on, but 8 is the number of initially spawned creeps per lane.

in our sample, suggesting that the mode does in fact randomly assign these heroes.

IV. Methodology

A. Bias in Simultaneous Equations

Suppose we are attempting to provide an equation to estimate the causal effect of changing some input on team's chance of winning. For our example, we are using Dota 2, and an important feature within the game is the rate at which a team generates *last hits* (LH). Strategy games like Starcraft may instead use economic features [10], while first person shooters like Halo may be interested in accuracy or kill/death ratios [11], [12]. Through LH the team generates income and experience points to power up their heroes. A naive model estimating this causal effect might be linear – however, it can be one of several functions, including random forests [6] or nonparametric functions [13], [14]:

$$win_i = \alpha_0 + \beta_1 LH_i + \epsilon_{1i} \tag{1}$$

Where we are interested in the causal effect of last hits on winning, estimated by the average partial effect: $\frac{dE(win_i|LH_i)}{dLH_i} = \beta_1$. We expect that $\beta_1 > 0$, players with many last hits should eventually win. However, this naive approach fails to consider the fact that the causality for this equation also runs in reverse. Winning players find opportunities to make last hits, losing players miss or are not given such opportunities. Indeed, Dota 2 permits damaging one's own units in order to prevent last hits. As such, equation 2 is also reasonable, where again this function can be of several types (we will show the linear case):

$$LH_i = \alpha_1 + \beta_2 win_i + \epsilon_{2i} \tag{2}$$

Where we believe $\beta_2 > 0$, winning conveys numerous positional advantages which assist in getting last hits. At the same time, players may have their last hits denied if they are losing. A quick substitution exercise will show the following:

$$\begin{aligned} LH_i &= \alpha_1 + \beta_2(\alpha_0 + \beta_1 LH_i + \epsilon_{1i}) + \epsilon_{2i} \\ LH_i &= \alpha_1 + \beta_2\alpha_0 + \beta_2\beta_1 LH_i + \beta_2\epsilon_{1i} + \epsilon_{2i} \\ (1-\beta_2\beta_1)LH_i &= \alpha_1 + \beta_2\alpha_0 + \beta_2\epsilon_{1i} + \epsilon_{2i} \\ LH_i &= \frac{\alpha_1 + \beta_2\alpha_0 + \beta_2\epsilon_{1i} + \epsilon_{2i}}{1-\beta_2\beta_1} \\ Cov(LH_i, \epsilon_{1i}) &= \frac{\beta_2\sigma_{1i}^2}{1-\beta_2\beta_1} \neq 0 \end{aligned} \tag{3}$$

This violates the Gauss-Markov conditions in Equation 1: no independent variable may be correlated with the error term, and so a linear estimate of β_1 in equation 1 will be biased in the direction of $\frac{\beta_2\sigma_{1i}^2}{1-\beta_2\beta_1}$. Given the assumptions that $\beta_1 > 0, \beta_2 > 0$, our estimates of β_1 will be too small if $Cov(LH_i, \epsilon_{1i}) < 0$, and too large if $Cov(LH_i, \epsilon_{1i}) > 0$ [15]. This problem happens frequently in econometrics literature. The term "endogenous" is used to describe any input variable which is influenced by either the outcome, or another omitted variable through the error term ϵ. Independent variables which have no such confounding problems are called "exogenous".

In order to resolve this simultaneity bias for a vector of potentially endogenous variables y_2, and accurately measure the causal effect of an increase in y_2 we take advantage of a family of procedures called Instrumental Variables (IV) and Control Function approaches (CF).

Formally, our goal is to predict the average partial effect of an input on the predicted value of y_1 using a set of potentially endogenous regressors y_2 and exogenous-well behaved regressors x. Critically, we also need additional exogenous variables (instruments), denoted (z), which must satisfy several properties. IV requires that instruments must be uncorrelated with the error terms, $Cov(z, \epsilon) = 0$ (instrumental exogeneity), and must be a relevant predictor of y_2, $Cov(z, y_2) \neq 0$ (instrumental relevance) [16]. In short, z must *cause* x, but *not cause* wins except through the intermediaries provided.

The IV approach is a direct two step procedure which works in a linear case (only):

$$\begin{aligned} &a)\ y_2^{Stage1} = \beta_z^{Stage1} z + \beta_x^{Stage1} x + \epsilon_1^{Stage1} \\ &b)\ y_1 = \beta_x^{IV} x + \beta_y^{IV} \hat{y_2}^{Stage1} + \epsilon_2^{IV} \end{aligned} \tag{4}$$

We estimate the first equation, and use the estimates of y_2 in the second. The initial standard errors from this second stage will be misleading, because they do not reflect the proper variability of $\hat{y_2}$. Proper standard errors from this second stage can be found via bootstrapping, as well as other methods outlined in [16] for estimation and evaluation.

CF requirements are similar but also require a distributional exclusion restriction, where $(\sim)$ denotes equality of conditional distributions.

$$\begin{aligned} \epsilon_2|x, z &\sim \epsilon_2|x, \epsilon_1 \\ &\sim \epsilon_2|\epsilon_1 \end{aligned} \tag{5}$$

See [13], [14] for more details. The CF approach is a similar multi-step procedure:

$$\begin{aligned} &a)\ y_2^{Stage1} = \beta_z^{Stage1} z + \beta_x^{Stage1} x + \epsilon_1^{Stage1} \\ &b)\ y_2 - \hat{y_2}^{Stage1} = \hat{\epsilon_1}^{Stage1} \\ &c)\ y_1 = f(\beta_x^{CF} x + \beta_y^{CF} y_2 + \beta_\epsilon^{CF} \hat{\epsilon_1}^{Stage1} + \epsilon_2^{CF}) \end{aligned} \tag{6}$$

Again, we estimate the first equation, then the second, and use the residuals calculated in parts *6a* and *6b* to estimate the final values. Here the IV estimates are consistent and the control function approach will return the same estimates as the IV approach if f is linear.[2] In order to estimate the value of $\hat{y_1}$, one may use an average structural function, calculated:

$$E[f(\beta_x^{CF} x + \beta_y^{CF} y_2)] = N^{-1} \sum_{i=1}^{N} (\beta_x^{CF} x + \beta_y^{CF} y_2 + \beta_\epsilon^{CF} \hat{\epsilon_{1i}}^{Stage1}) \tag{7}$$

which averages out the endogenous residuals and provides an unbiased estimate of $E(y_1|x, y_2)$, where y_1 is the value from

[2]While tempting, estimating a nonlinear IV approach such as $y_1 = f(\beta_{x2} x + \beta_{y2}\hat{y_2} + \epsilon_2)$ does not obtain consistent estimators, and when y_2 is binary the mistake is sometimes called the "forbidden regression" [15].

6c. Again, standard errors can be found by bootstrapping. We note that because winning is binary, we prefer the CF approach and choose f as Probit (denoted Φ), which is bounded to produce results between 0 and 1. We note Athey and Tibshirani [6] have done a great deal of work expanding the range of possible f functions to include random forests, and Blundell and Powell [13], [14] have done work on the process of estimating nonparametric f.

B. Application & Practice

In practice, it is difficult to find such instruments. The instruments can be weak [17], or potentially still endogenous. In our case we have some strong instruments, and some potentially weak instruments, which carries a risk of potential bias in small sample sizes, particularly if the instrument is not perfectly exogenous from all omitted variables [15]. In practice, game developers can easily introduce instruments into their games. Some examples of perfect instruments, each which directly affect only a single aspect of the game, include:

- A randomly given "holiday gift box" which provides +10 damage for the duration of the current game.
- A randomly appearing "demonic sprite" which provides -10 to all healing effects for the duration of the current game.
- A random "demolition box", appearing in one side's spawn. It can be used once to do 200 damage against structures but no damage to players.

We quickly reiterate our instrument of choice. In Dota 2, players have an AR mode where all players are assigned a random hero possessing one of the three primary attributes: $\{str, int, agi\}$. When given a choice of heroes, teams strongly prefer a balanced composition. Compositions such as five strength heroes or five intelligence heroes are considered very suboptimal. Team composition is critical, though there is much heterogeneity among classes. OpenAI lost its only game against professional players when it was assigned a questionable composition at game start. OpenAI assessed its probability of winning at the start of the game roughly 2.9% [18].[3] This assessment is not because of a fundamental prohibition on winning with any particular combinations of heroes, but rather because it becomes impossible to accomplish the basic tasks of the game, such as last hitting, killing enemy heroes, or destroying towers. This suggests that certain combinations of hero attributes can influence in-game performance up or down. Therefore, these randomly assigned hero attributes serve as experimental variation (instruments) for predicting those otherwise endogenous game features, and the combination of $\{str, int, agi\}$ seems to be a good proxy for that property.

We instrument using indicators for the number of heroes for each attribute. Since these are randomly assigned before the match, they are fully exogenous. If there are other omitted variables through which hero composition influences victory, our results still suffer from the remaining bias, though if these channels are small and our instruments strongly predictive, our bias is small [15]. As such, choice in instrument is extremely important, as is the exhaustive collection of controls. Conveniently, in this digital environment, the collection of features can be fairly comprehensive. If the instrument choice is still a concern, we note that a key point of this paper is that game developers can insert perfect instruments. These perfect instruments can directly affect a single aspect of the game, such as the examples we provided above. From there, this family of IV/CF procedures allows for inference of the causal effect of those game features.

C. Using Instrumental Variables

We have collected the following endogenous variables (y_2) for both friendly teams A and enemy teams B (marked in subscripts as a, b), and found the per player average over the game duration. This comes to a total of six endogenous variables.

1) lh_a, lh_b: The last hits, where a player kills an enemy non-hero unit. This is the primary method of collecting income.
2) $kills_a, kills_b$: The kills, where a player kills an enemy hero unit. This is the secondary method of collecting income.
3) $towerdmg_a, towerdmg_b$: The amount of tower damage dealt every minute. This is a tertiary method of collecting income.

We have the following four exogenous variables (x), which are determined prior to the start of the game and are not changed during the game.

1) $mmrdiff_a$: The amount that team A is behind in MMR.
2) $largestpartysize_a, largestpartysize_b$: The number of players grouped together on the team.
3) A vector of ones: To create an intercept term, included here to save on notation.

We have the following instrumental variables (z), which are determined prior to the start of the game and are expected to directly influence only the in-game statistics y_2. We need at least one instrument for each of the six variables in y_2. Here we use $C(attribute)_{team}$ to represent the count of heroes from $team$ with $attribute$ as factors.

1) $C(str)_a, C(int)_a$: This is a full set of indicator variables for *str* and *int* heroes that take the value 1 if team A has a particular count of that hero attribute or 0 otherwise. The counts range between 0 and 5, and must total to 5 or less.[4]
2) $C(str)_b, C(int)_b$: This is a second set of indicator variables for team B.

Our goal is ultimately to determine the causal effect of an increase in endogenous features (y_2) on winning (y_1):

1) win_a: An indicator taking the value 1 if team A has won and 0 otherwise.

[3]We note their composition had heroes that do not perform according to the stereotypical roles we listed above.

[4]We note that $C(agi)_a = 5 - C(str)_a - C(int)_a$, so it is redundant to include $C(agi)_a$.

We estimate this causal effect using each of the previous Equations 4 and 6. We contrast these IV/CF results with naive estimators, shown below.

The naive approaches of these methods are the corresponding naive linear probability model:

$$win_{ai} = \beta_x^{IV-N} x_i + \beta_y^{IV-N} y_{2i} + \epsilon_{2i}^{IV-N} \tag{8}$$

And a naive version of the corresponding Probit estimation.

$$win_{ai} = \Phi(\beta_x^{CF-N} x_i + \beta_y^{CF-N} y_{2i} + \epsilon_{2i}^{CF-N}) \tag{9}$$

We then calculate the average structural function evaluated at the mean of $\bar{x}$ and along the range of several selected endogenous variables of interest from y_2: last hits/minute, kills/minute, and tower damage/minute. We show the biased naive estimators are broadly different from the IV/CF family of estimators.

V. Results

Preemptively, one might expect cross validations, where the sample is split into groups. A model is then fit on a training set composed of one such group, and applied to a test set. Some measure of fit, such as R^2 is then retained and the model is discarded. After multiple models are compared, the model with the best fit measure is then used to fit the entire data set, and the discussion continues from there. This would not be useful in this context because a better fit is not the focus of the process. In fact, we expect the unbiased process to obtain a weaker fit by most or all measures. This is specifically because these unbiased approaches sacrifice the quality of fit in order to obtain more accurate measures of the causal effect of the inputs. Our emphasis is instead on how the measured effect of the inputs is misleading when using an biased method. This difference is visible through the direct comparison of the biased OLS and unbiased IV coefficient estimates in Table III, and the dramatically different estimated effect of the inputs on the probability of winning between the CF and simple Probit approaches.

A. First Stage

The battery of tests for the first stage are common to both IV and CF estimations. Since the hero attributes are randomly assigned before the game, we are not concerned with the hero attributes being caused by winning, but we can use these randomly assigned attributes to provide an estimate for each the six endogenous variables, $lh_{a,b}, kills_{a,b}, towerdmg_{a,b}$. These first stage estimates of game features will be derived entirely from the randomly assigned hero attributes prior to the game, and thus, will not have any simultaneity problems as the game veers towards victory or defeat. We discuss the predictive power of the commonly shared first stage in Table II.

We note that while the initial tests indicate we do not have weak instruments [19], the rule of thumb is that one ought to have an F-statistic over 10 [17], [20], which some (but not all) of our instruments pass. Since some of our instruments are weak by that standard, our estimates may have higher variance and may be biased in small samples as the errors compound between stages. (This is a particularly common problem, so when choosing or creating instruments, make sure they are strong enough to have an easily measurable effect on the game.) Game developers may also benefit from significantly larger data sets than the one we have collected. We suspect the main reason only a few of our instruments have strength over 10 is due to the fact that there are many heroes who do not behave according to their attribute, such as agility heroes that serve as tanks, ex: Lone Druid or Razor [21]. This means attributes are not a perfect predictor of in-game performance but does not nullify the hero attributes influence on the in-game performance. We also have no evidence of over-identification with a Sargan test [22]. The Wu-Hausman test checks the hypothesis that the original OLS is the same as linear IV, which we firmly reject [23] [24] [25]. The findings of the Wu-Hausman test in Table II are reiterated by the fact that we see substantial differences in the point estimates of β below.

TABLE II
Stage 1 Tests

	df1	df2	statistic	p-value
Weak instruments (lh_a)	20.00	6660.00	16.18	0.00***
Weak instruments ($kills_a$)	20.00	6660.00	4.97	0.00***
Weak instruments ($towerdmg_a$)	20.00	6660.00	3.50	0.00***
Weak instruments (lh_b)	20.00	6660.00	16.18	0.00***
Weak instruments ($kills_b$)	20.00	6660.00	4.97	0.00***
Weak instruments ($towerdmg_b$)	20.00	6660.00	3.50	0.00***
Sargan	14.00		18.42	0.19
Wu-Hausman	6.00	6668.00	5.45	0.00***

Note: *p<0.1; **p<0.05; ***p<0.01

B. Linear Instrumental Variables

In Table III we compare the results of the Naive OLS to the linear IV approach which takes into account the endogeneity of these in-game features. We are interested in the average partial effects listed below: $\frac{dE(y_1|\bar{x},y_2)}{dy_2}$. Again, for linear models, this is simply the coefficient β, and IV and OLS are both linear models estimating the same β. We have put these coefficients and their standard deviations, along with measures of fit in Table III.

Typically, the IV estimates have much larger standard errors than OLS because of the use of the first stage. Furthermore, these linear probability models do not recognize the boundary on win_a between 0 and 1. As a result the standard errors, even after correcting for IV, are not ideal [16].[5] We also point out that the IV process has a slightly weaker fit by all simple measures, smaller R^2, smaller adjusted R^2, slightly larger residual standard error. The point, then, is emphatically not that IV leads to better predictive power, but IV estimates the causal effect for each of the input features, rather than a predictive effect that depends only on correlation.

When we are reviewing these results, we will be looking for the following specific cases. Firstly, where winning leads to an increase in the input variable. In these cases, the naive

[5]The CF approach, used later, is adaptable to such nonlinearities.

TABLE III
NAIVE OLS AND IV

	Dependent variable:	
	won_a	
	Naive OLS	*IV*
	(1)	(2)
Team A - Endogenous		
lh_a	0.020***	0.081***
	(0.004)	(0.031)
$kills_a$	0.463***	0.890
	(0.059)	(0.779)
$towerdmg_a$	0.004***	0.003*
	(0.0001)	(0.002)
Team B - Endogenous		
lh_b	−0.020***	−0.081***
	(0.004)	(0.031)
$kills_b$	−0.463***	−0.890
	(0.059)	(0.779)
$towerdmg_b$	−0.004***	−0.003*
	(0.0001)	(0.002)
Exogenous		
mmr_diff_a	0.00000	0.00000
	(0.00000)	(0.00000)
$largestpartysize_a$	0.0005	−0.004
	(0.003)	(0.005)
$largestpartysize_b$	−0.0005	0.004
	(0.003)	(0.005)
Constant	0.500***	0.500
	(0.026)	(0.406)
Observations	6,684	6,684
R^2	0.786	0.768
Adjusted R^2	0.786	0.768
Residual Std. Error (df = 6674)	0.231	0.241
Note:	*p<0.1; **p<0.05; ***p<0.01	

approach will conflate the two, and the average partial effect of the input will be overestimated. Alternatively, when winning leads to a decrease in the input variable, the naive approach will then underestimate the average partial effect of the input. We see both types of errors in the naive process.

The first change is visible in the average partial effect on last hits/minute (for both teams A and B), the IV estimate is more than quadruple the naive estimated association between last hits and game victory.[6] The difference between the biased and unbiased model can be interpreted as follows: teams that are winning get fewer last hits (e.g., once a player is winning they instead collect kills or destroy towers), and teams that begin losing tend to focus on getting more last hits (e.g., it is the players' only safe source of income). For example, this behavior may happen if the winning team has already acquired critical items while the losing team is still struggling to obtain basic equipment. When this pattern of behavior persists, naive estimates will have coefficients that are biased downward. The IV approach, when done correctly, removes this bias.

In the next variable of interest, kills/minute (for both teams), the estimated average partial effect of another kill per minute remains large, but loses its significance. Partially, this large magnitude is due to the scale of kills. An additional kill/minute would put any team well above the maximum observed value of 0.51.[7] After rescaling the biased estimate, an additional kill over a 45 minute game would be associated with a significant 1% increase in win rate. The comparative unbiased estimate is insignificantly different than zero. One possible explanation for the significance change is that kills are prioritized when winning, while losing teams prioritize collecting last hits. If this is the case, the naive approach will conflate winning and kills, and bias the estimated average partial effect of kills upwards and towards significance, while the IV approach does not have that bias.

Lastly, the IV approach estimates a 25% smaller and less significant average partial effect of tower damage on winning for both teams.[8] Again, the direction of this reduction can be explained by winning teams tending to seek out opportunities to destroy towers, and losing teams tending to avoid such risky plays. The naive approach will conflate the two effects and overestimate the relevance of tower damage, while the IV approach does not have such biases.

Between the two estimations, the exogenous variables of MMR and party size (for each team) are small and insignificant. These criteria are matched between the two teams – the matchmaking system always pairs large parties with each other and teams of similar MMR. We note that the coefficient on party size is reduced by 80%, even though it is not exogenous. It is possible that bias in one feature can contaminate estimates of other coefficients in the estimation, so even the average partial effects of exogenous variables can be better estimated after the IV process [16].

C. Nonlinear Control Function Approach

Below we compare the results of the naive Probit approach to the improved CF approach. For the naive approach, we estimate the average partial effect. We then evaluate the average structural function (Equation 7) with those same inputs to generate an unbiased estimate of the probability of winning. Both the estimates are bootstrapped 50 times (both stages) to estimate the standard errors and display ± 1 standard deviation. We note that the standard deviations for the unbiased process are particularly large (in general), a problem exacerbated by the strength of the given instruments. Thus far, the estimated effects of team A and team B's inputs have been inverse of each other, so we only look at team A's inputs for brevity since this pattern continues.

First, we examine the effect of team A's last hits on team A's probability of winning in Figure 1. We begin by noting the biased naive estimates (green) are very confident and particularly narrow. However, these naive estimates are biased and cannot be relied upon. On the other hand, we can see that the unbiased control function approach (black) provides large standard errors. We note that the SD can be reduced if the size of the data set increases, or the precision of the instruments can be further increased – say by constructing better instruments.

[6]There are on average about 3.5 last hits per minute.

[7]There are on average about 0.18 hero kills per minute

[8]There is on average about 63 tower damage dealt per minute.

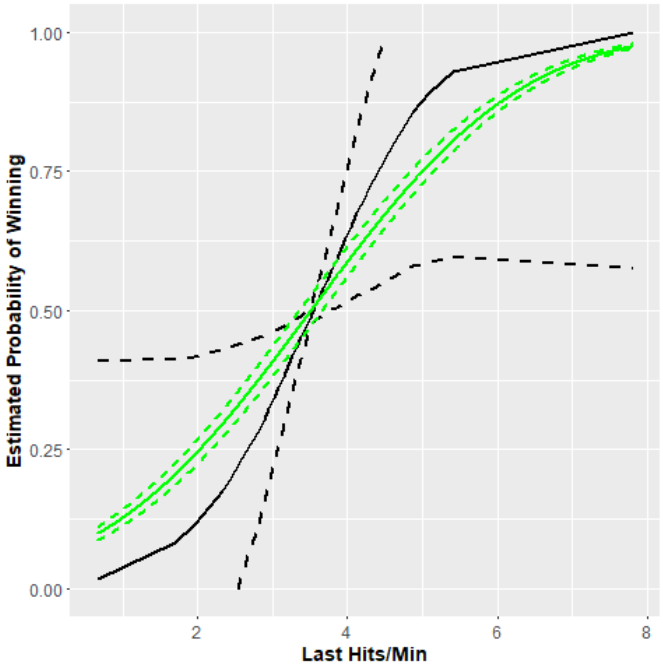

Fig. 1. Green - Naive, Black - Control Function, Dotted Lines - 1 SD
The control function approach puts greater emphasis on last hits than naive methods, suggesting increasing last hits will cause an increased probability of winning, all else being held equal.

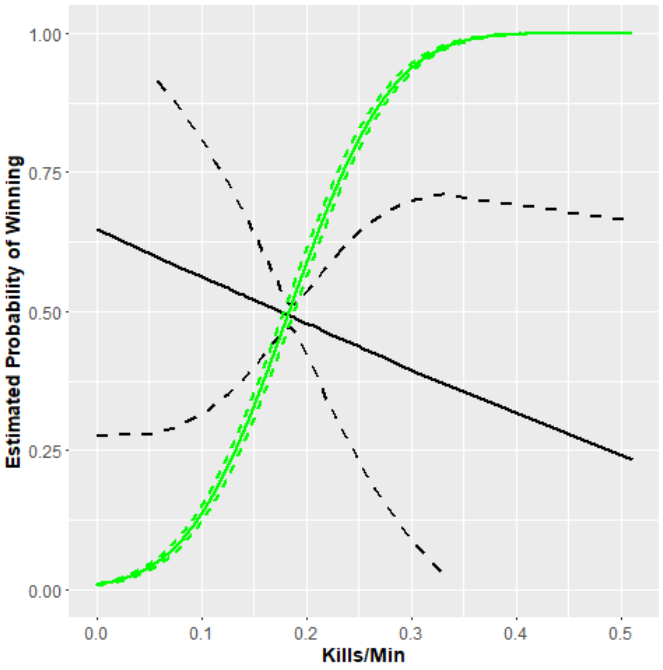

Fig. 2. Green - Naive, Black - Control Function, Dotted Lines - 1 SD
The control function approach removes a great deal of bias towards kills. This suggests that kills alone have little effect on the probability of winning, but rather teams that are winning pursue kills.

As it stands, the CF estimates suggest a team that has 5 last hits/minute is very likely to win the game (89%), while the naive estimates suggest they are only about 75% likely to do so. As a reminder, both estimates keep all other inputs constant at the mean. Again, this result would occur if teams that are winning do not emphasize acquiring last hits compared to teams that are losing. The CF approach removes this bias.

In Figure 2 we show the effect of team A's kills on team A's probability of winning. The biased naive estimates (green) are very confident and particularly narrow. The estimates suggest a team that has 0.1 kills/minute, about 1 standard deviation below the mean, are only about 13% likely to win which seems low. In contrast, the CF estimates (black) suggest that a team with 0.10 kills/minute continue to be about evenly matched with their opponents (56%). We again emphasize this estimate keeps all other values constant at the mean. As with the IV case, this substantial difference could be explained by kills/minute having little effect on the game independently, but teams that are already winning tend to seek them out while teams that are losing tend to avoid these risky encounters.

For this third and final Figure 3, we show the effect of team A's tower damage/minute on team A's probability of winning. Again, the biased naive estimates (green) are again very confident and particularly narrow. The naive estimates (black) suggest a team that has 100 tower damage/minute, about 1 standard deviation above the average, are almost certainly winning (94%). This estimate seems unreasonably high. Conversely, the CF estimates suggest a team with the same tower damage remains about equally matched with

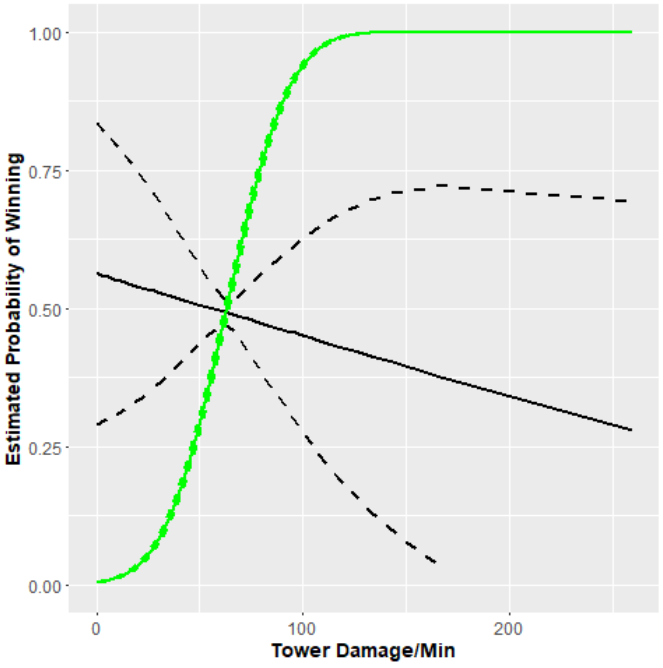

Fig. 3. Green - Naive, Black - Control Function, Dotted Lines - 1 SD
The control function approach removes a great deal of bias towards tower damage, suggesting tower damage alone has only a small impact on the probability of winning, but rather teams that are winning pursue tower damage.

their opponents (45%). This substantial difference between the biased and unbiased estimates can be explained by the fact that winning teams tend to attack towers to end the game, and teams that are losing tend to be unwilling or unable to attack enemy towers. However, all else being equal, some additional tower damage will have almost no measurable impact on the outcome of the game.

VI. Conclusion

In this paper, we show that naive estimates of causal effects can be severely biased in linear cases. We also show the appropriate method for linear cases (instrumental variables) and nonlinear cases (control functions). We take a novel data set from OpenDota [8] and use it to estimate causal effects of critical game variables, (last hits/minute, kills/minute, and tower damage/minute) using both naive and the improved unbiased methods. The naive estimates are overconfident and strongly biased, particularly in the case of kills/minute and tower damage/minute. We found that the naive approach overestimates the causal effect of an additional 40 tower damage/minute (about 1 SD) by nearly 49%, holding other factors constant. A similar bias emerges when investigating the causal effect of kills/minute.

The naive methods, however, appear to slightly underestimate the causal effect of last hits, which are the fundamental source of resources in Dota 2. These biases have been created by the fact that winning teams tend to seek out opportunities to obtain kills and tower damage, while losing teams tend to seek out opportunities to acquire last hits. The unbiased methods provide estimates that are far more reasonable. The IV and CF approaches compensate for the fact that winning teams tend to seek out opportunities to find kills and damage towers. The unbiased method more accurately measures the impact of small changes in last hits, kills, and tower damage on winning the game.

As a result, game developers should be cautious when using naive estimates to predict the impact of new features, or the relevance of player attributes towards winning. Ideally, game developers should instead reference unbiased estimates as demonstrated in this paper.

References

[1] S. Mercer (Principal Designer), "Overwatch matchmaking system," Blizzard Official Forum, Oct. 2016. [Online]. Available: https://us.battle.net/forums/en/overwatch/topic/20749737390

[2] A. H. Christiansen, B. F. Nielsen, and E. Gensby, "Multi-parameterised matchmaking: A framework," in *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 2018, pp. 377–380.

[3] S. Mercer (Principal Designer), "Overwatch matchmaking system," Blizzard Official Forum, Dec. 2017. [Online]. Available: https://us.battle.net/forums/en/overwatch/topic/20759648155

[4] "Broadcasters struggling to give context and meaning to game stats - by karahol," Dec 2017. [Online]. Available: https://www.winstonslab.com/news/2017/12/07/owl-overwatch-context-statistics/

[5] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *International conference on machine learning*, 2016, pp. 3020–3029.

[6] S. Athey, J. Tibshirani, S. Wager *et al.*, "Solving heterogeneous estimating equations with gradient forests," *arXiv preprint arXiv:1610.01271*, pp. 168–189, 2016.

[7] J. Pearl, *Causality*. Cambridge university press, 2009.

[8] OpenDota, "Dota 2 statistics," Mar. 2019. [Online]. Available: https://www.opendota.com/

[9] Dota 2 Wiki, "Matchmaking/Seasonal Rankings," Mar. 2019. [Online]. Available: https://dota2.gamepedia.com/Matchmaking/Seasonal_Rankings

[10] B. S. Weber, "Standard economic models in nonstandard settings–starcraft: Brood war," in *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 2018, pp. 417–424.

[11] D. Buckley, K. Chen, and J. Knowles, "Rapid skill capture in a first-person shooter," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 9, no. 1, pp. 63–75, 2017.

[12] K. J. Shim, K.-W. Hsu, S. Damania, C. DeLong, and J. Srivastava, "An exploratory study of player and team performance in multiplayer first-person-shooter games," in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. IEEE, 2011, pp. 617–620.

[13] R. Blundell and J. L. Powell, "Endogeneity in nonparametric and semiparametric regression models," 2003.

[14] R. W. Blundell and J. L. Powell, "Endogeneity in semiparametric binary response models," *The Review of Economic Studies*, vol. 71, no. 3, pp. 655–679, 2004.

[15] J. M. Wooldridge, *Introductory econometrics: A modern approach*. Nelson Education, 2015.

[16] ——, *Econometric analysis of cross section and panel data*. MIT press, 2010.

[17] J. Bound, D. A. Jaeger, and R. M. Baker, "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak," *Journal of the American statistical association*, vol. 90, no. 430, pp. 443–450, 1995.

[18] OpenAI, "Openai five benchmark: Results," Mar. 2019. [Online]. Available: https://www.opendota.com/

[19] C. Kleiber and A. Zeileis, *Applied Econometrics with R*. New York: Springer-Verlag, 2008, ISBN 978-0-387-77316-2. [Online]. Available: https://CRAN.R-project.org/package=AER

[20] D. O. Staiger and J. H. Stock, "Instrumental variables regression with weak instruments," 1994.

[21] Dota 2 Wiki, "Dota 2 wiki," Mar. 2019. [Online]. Available: https://dota2.gamepedia.com/

[22] J. D. Sargan, "The estimation of economic relationships using instrumental variables," *Econometrica: Journal of the Econometric Society*, pp. 393–415, 1958.

[23] J. Durbin, "Errors in variables," *Revue de l'institut International de Statistique*, pp. 23–32, 1954.

[24] W. De-Min, "Alternative tests of independence between stochastic regressors and disturbances," *Econometrica (pre-1986)*, vol. 41, no. 4, p. 733, 1973.

[25] J. A. Hausman, "Specification tests in econometrics," *Econometrica: Journal of the econometric society*, pp. 1251–1271, 1978.