FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES ESCUELA DE ECONOMÍA

Factores que influyen en la probabilidad de ganar un partido de fútbol: Análisis de la Liga 1 peruana en el periodo 2015 – 2019

Descripción breve

El intercambio, la negociación, la estratégia y el juego son temas de interés central en el análisis económico y el futbol en tal sentido, aporta un gran escenario, con resultados empíricos que incrementan este interés analítico.

AUTOR
ARRUNATEGUI ZETA MARCELO ALEJANDRO
PROFESOR MENTOR
CHANG ROJAS VICTOR ALEJANDRO
COEDICIÓN Y COORDINACIÓN
Victor Echevarría Alvarado
Edith Rosas López
REVISIÓN COMITÉ DE CALIDAD
Jorge W. Rosas Santillana
Mónica Vega Arana
Víctor Chang Rojas
Luis Peralta Letiche

Lima, enero 2021

RESUMEN

El presente trabajo de investigación tiene como objetivo determinar las variables que influyen de manera significativa en la probabilidad de ganar un partido de fútbol. Específicamente, para el caso de la Liga 1 peruana para el periodo 2015-2019. Para ello se tomarán variables estadísticas, deportivas y climatológicas y, respecto a la metodología, se empleará tanto el modelo Logit Ordenado como Logit Ordenado Generalizado para realizar el análisis econométrico. Los resultados corroboran los efectos negativos que generan las tarjetas rojas y el tener jugadores con mayor peso corporal en la probabilidad de empatar y ganar un partido. A su vez, se muestra que el tener un mayor porcentaje de posesión durante el partido y tener jugadores más altos incrementa la probabilidad de obtener un resultado de empate o victoria. Los datos muestran también que tener jugadores, con una edad levemente mayor al promedio, incrementa estas probabilidades. Finalmente, se ratifica los efectos positivos que genera el anotar una mayor cantidad de goles y jugar en la modalidad de local.

Palabras Clave: Modelo Logit, economía del deporte, metodología econométrica

ABSTRACT

The present research work aims to determine the variables that significantly influence the probability of winning a soccer match. Specifically, for the case of the Peruvian League 1 for the period 2015-2019. For this, statistical, sports and weather variables will be taken and, regarding the methodology, both the Ordered Logit model and the Generalized Ordered Logit will be used to perform the econometric analysis. The results corroborate the negative effects of red cards and having players with higher body weight on the probability of drawing and winning a match. In turn, it is shown that having a higher percentage of possession during the match and having taller players increases the probability of obtaining a draw or victory result. The data also show that having players, slightly older than the average, increases these probabilities. Finally, the positive effects generated by scoring a greater number of goals and playing at home are ratified.

Key Words: Logit model, sport economics, econometric methodology

INTRODUCCIÓN

El término deporte ha tomado varias interpretaciones con el paso del tiempo debido a que es muy amplio y está abierto al análisis personal. Una definición general sería decir que es la realización de cualquier tipo de actividad que fomente o promueva las capacidades tanto físicas como mentales. Sin embargo, la que se utilizará para fines de esta investigación es la siguiente: "El deporte es un fenómeno complejo, abierto que expresa una idea en constante evolución acorde a los tiempos y que constituye un componente significativo de la experiencia vital del ser humano como individuo y del colectivo social" (Paredes, 2002). A su vez, existe una amplia gama de disciplinas (futbol, baloncesto, atletismo, boxeo, etc.) donde cada una posee características tanto generales como específicas que las hacen atractivas de practicar, según laspreferencias de cada individuo.

Tabla 1 - Ligas deportivas mayor remuneradas a nivel mundial

Liga	País	Deporte	Ingreso ($)
NFL	EEUU	Fútbol American o	$ 13 billones
MLB	EEUU	Béisbol	$ 9,5 billones
Premier League	Inglaterra	Fútbol	$ 5,3 billones
NBA	EEUU	Baloncesto	$ 4,8 billones
NHL	EEUU/Canadá	Hockey	$ 3,7 billones
Bundesliga	Alemania	Fútbol	$ 2,8

			billones
La Liga	España	Fútbol	$ 2,2 billones
Serie A	Italia	Fútbol	$ 1,9 billones
Ligue 1	Francia/Mónaco	Fútbol	$ 1,5 billones
B. Profesional	Japón	Béisbol	$ 1,1 billones

Fuente: Howmuch.net/Which professional sports leagues make the most

El deporte, con el paso del tiempo, ha pasado de ser una razón de agrupamiento social, destinada a satisfacer una necesidad de entretenimiento, a considerarse como un producto intangible, compuesto por una parte de servicios y otra productiva, cuya inversión, consume infraestructura, producción y administración responde a aspectos de índole económico. Es decir, paso a ser un generador de desarrollo y es aquí donde toma participación lo que se denomina como economía del deporte.

El análisis deportivo desde el punto de vista económico ha ido adquiriendo importancia en el ámbito académico, especialmente porque se le reconoce como un sector con las mismas características de los sectores tradicionales (Mesa y Arboleda, 2007). Esto debido a que, detrás de cualquier practica o evento deportivo, se encuentran una gran cantidad de actividades que guardan relación con lo económico. Por ejemplo, detrás de un partido profesional de fútbol televisado se necesitan elementos de otros sectores como indumentaria, alimentos, transporte, infraestructura, servicios médicos, medios de comunicación, entre otros. Por lo tanto, el sector deportivo tiene carácter interdependiente con otros sectores de la

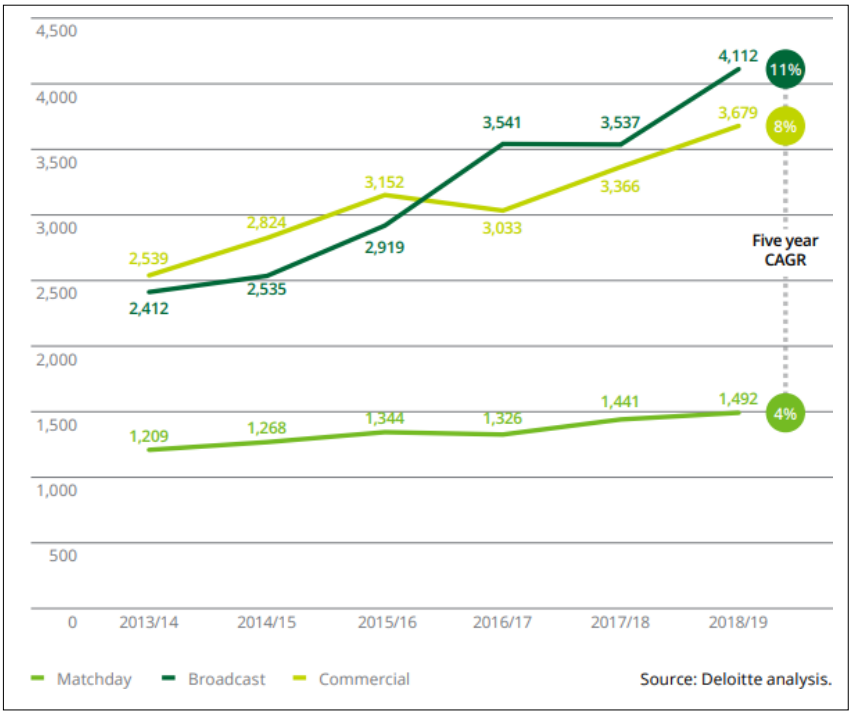
economía y va generando cada vez mayor interés ya que influye en la demanda y oferta agregada de los países. Hoy en día se puede apreciar una variada cantidad de literatura respecto a este sector que respalda lo antes mencionado, destacando principalmente la de países desarrollados donde las actividades y eventos deportivos representan un considerable porcentaje de su producto bruto interno. Sin embargo, varia para el caso de América Latina en donde, si bien existen investigaciones económicas para la mayoría de países que lo conforman, para el caso peruano es escaso encontrar literatura de esta naturaleza.

Se hace énfasis en el análisis econométrico, metodología que se empleará en esta investigación, debido a la variedad de aplicaciones que posee en las distintas áreas académicas y del conocimiento. Con esta herramienta se pueden analizar interrogantes en el sector deportivo como: que variables influyen en el rendimiento de un equipo, en la cantidad de puntos anotados, en el traspaso de un jugador, en la asistencia a eventos deportivos, en el mercado de apuestas y su pronosticación, entre muchos más. Las variables a emplear (de desempeño, estadísticas, demográficas, climatológicas, etc.) varían según el deporte y de lo que se quiere examinar. Pero, para el caso peruano, investigaciones econométricas dentro del sector deportivo son prácticamente inexistentes en donde la falta de una base de datos completa en el país se aproxima a ser la principal causa. A pesar de que existe información en fuentes oficiales, a nivel general, sobre las distintas disciplinas en el Perú, no se cuenta con una base de datos detallada y de diferentes años lo cual es una enorme limitante al hacer análisis económico utilizando la metodología mencionada. Sin embargo, existen sitios (en este caso páginas webs) dedicados a la recolección estadística en donde algunos de estos cuentan con información de distintos años. Por lo cual se analizará, para fines de esta investigación, el caso del denominado “deporte rey” ya que en el país este es el que más predomina a nivel sociocultural.

El futbol es uno de los deportes, por no decir el principal, que más predomina a nivel mundial el cual practicarlo o presenciarlo agrupa masas a nivel de género, raza, condición socioeconómica, religión, etc. Se manifiesta a través de las modalidades amateur y profesional siendo esta última la que más destaca, en especial, la liga profesional de futbol europeo. Según la consultora Deloitte en su

reporte “Football Money League 2020”, los 20 clubes de futbol (todos pertenecientes a la eurozona) con mayor facturación en el mundo han acumulado 9.283 millones en ingresos durante la última temporada logrando un incremento interanual del 12%, obteniendo un nuevo récord histórico. Como se mencionó anteriormente, el sector deportivo posee una relación de interdependencia con otras actividades económicas y sociales en donde, de no existir el futbol, muchas de estas desaparecerían lo cual disminuiría el bienestar de la población (Coremberg, Sanguinetti, Wierny, 2016). La realización de eventos de gran proporción como las copas mundiales y partidos entre equipos favoritos producen impactos económicos considerables sobre el país donde se realiza, llegando a modificar incluso su estructura económica por la necesidad de adecuar los servicios públicos y generar nueva infraestructura necesaria para la realización de este tipo de eventos de gran magnitud.

Figura 1 - Crecimiento de ingresos de los 20 mejores clubes europeos



Fuente: Deloitte Sports Business Group January 2020

¿Y qué es lo que busca todo equipo profesional? La respuesta cae por si sola, buscan lograr la mayor cantidad de victorias posibles en una determinada temporada. El ganar partidos consecutivos, tener una buena posición en las estadísticas de los mejores equipos o incluso ganar un torneo trae efectos positivos como: patrocinio de diversas marcas, mayor cantidad de asistencia del público a los partidos, mejor publicidad y captación de los medios, entre otros. Y a su vez, respecto a los jugadores, implicaría un incremento en su nivel de ingreso entre otros beneficios debido a la relación directa que poseen estos con el club respectivo (mayor rentabilidad de una empresa conlleva efectos positivos para sus trabajadores). Entre algunas de las responsabilidades del plantel técnico está la de analizar, en la mayoría de lo posible, los factores que puedan llegar a influir en que el equipo gane u obtenga otro resultado en el próximo partido a disputar. Sin embargo, no es solo el plantel, dicho análisis es de interés común tanto por el público, programas deportivos televisivos o radiales, páginas webs de fanáticos del deporte, etc. Y es que una de las cualidades que más atrae de este deporte es su incertidumbre. No se puede predecir un resultado destinado al azar, pero si tratar de pronosticarlo examinando factores que se cree que tengan influencia alguna en la oportunidadde salir victorioso.

Y es que, volviendo al análisis literario, existen escasos artículos que tratan esta problemática. Contreras y Muñoz (2015) utilizan variables deportivas como posesión, si jugó de local o visitante, faltas, goles a favor y en contra, entre otras más para examinar el rendimiento de los equipos, en dicho caso, de la Premier League que permita la facilitación en la toma de decisiones tácticas. Sus resultados muestran que si existen variables generales que expliquen en cierta parte el comportamiento de un partido y que jugar de local tiene efectos significativos. Santiago Celis (2013) analiza si la transferencia de talento extranjero tiene un impacto sobre el desempeño de un equipo. Para ello tomo variables tanto deportivas (que comparten cierta similitud con las de los autores antes mencionados) como geográficas que expliquen de manera puntual el comportamiento de este fenómeno en la liga Postobón en Colombia. A pesar de que no se obtuvieron conclusiones especificas respecto a los jugadores extranjeros en el desempeño, el resto de variables si mostraron una clara influencia tanto positiva como negativa. Chumacero (2009) analiza los determinantes detrás de los

resultados de partidos de clasificación en la zona sudamericana y toma en consideración la altitud del estadio como una estas. El autor concluye que dicha variable climatológica no se considera relevante, pero si otros como la temperatura o humedad y, también, se muestra evidencia que la ventaja de jugar de local es extremadamente importante.

Se analiza, también, los factores que influyen en el salario de un jugador, en la afluencia del público a los distintos eventos, los efectos que trae un cambio en el plantel técnico, etc. Lo que se trata de decir con estas referencias mencionadas es que investigaciones en torno a lo que puede influir en la probabilidad de ganar un partido son bastante limitadas y, para el caso peruano, se podría argumentar que es prácticamente nulo. Esto último es la principal motivación detrás de este trabajo. Una propuesta de investigación de índole econométrica en este sector sería un gran aporte ya que, como se mostró y se mostrará más delante de forma más detallada, el deporte y la economía poseen una fuerte relación.

Es aquí donde se centrará esta investigación ya que el futbol peruano, en términos de resultados, mantiene un patrón de escasas victorias y usuales derrotas. La última participación de la selección peruana en la copa del mundo, con excepción de la celebrada en Rusia en 2018, fue en 1982. Analizando otros eventos, nunca se ha ganado la Copa Libertadores (competencia más importante de la región) donde la mayoría de equipos peruanos no pasaron de las fases iniciales y, también, solo se ha ganado en dos ocasiones la Copa América, siendo la última hace 45 años.

Entonces, ¿Por qué los logros deportivos del futbol peruano son escasos o excepcionales?

¿Qué impide que nuestro futbol sea competitivo a nivel internacional? Si alguna vez tuvimos una “época dorada” de logros, ¿Cuáles han sido las razones por las que no hemos podido seguiradelante con ese legado? (Panfichi, Vila, Chávez, Saravia, 2018).

Objetivo del estudio

Por lo tanto, el presente trabajo tiene como pregunta de investigación la siguiente:

¿Qué factores influyen en la probabilidad de ganar de los equipos de futbol de la Liga1 peruana durante el periodo 2015 – 2019?

El objetivo principal está centrado en determinar los factores que influyan de manera significativa en la probabilidad de ganar un partido de futbol, de manera más específica, los de primera división de la liga peruana en la conocida Copa Movistar. Se examina a nivel regional debido a que, como se mencionó anteriormente, no se cuenta con una fuente de datos de diferentes años lo suficientemente amplia que muestre estadísticas de los equipos peruanos que hayan tenido participación en torneos de carácter internacional. Cabe resaltar, que la unidad de análisis de la investigación son los partidos. Por ello, toda la información a recolectar en torno a las variables estará en términos de los partidos. Como ya se mencionó, el periodo de tiempoque se analizará es desde el año 2015 hasta 2019.

Respecto a las variables, como dependiente, se tiene una variable categórica que contempla si el equipo perdió, empato o gano. Por otro lado, como independientes se tomarán los años de antigüedad del equipo, cantidad de goles anotados, cantidad de tarjetas amarillas, cantidad de tarjetas rojas, edad promedio de los jugadores, si el equipo jugo de local o visitante, cantidad de jugadores extranjeros, altitud del estadio, altura promedio de los jugadores, peso promedio de los jugadores, porcentaje de posesión y cantidad de substituciones. Es decir, se tomarán variables estadísticas, deportivas y climatológicas. En cuanto a la metodología, se empleará tanto el modelo Logit Ordenado como Logit Ordenado Generalizado para realizar el análisis econométrico. Las fuentes de donde se obtendrán los datos son las paginas CeroAcero.es y Resultados-futbol.com, dedicadas a mostrar todo tipo de información y estadísticas de los equipos de fútbol de distintas ligas del mundo.

En este sentido se plantea la siguiente hipótesis.

El presente trabajo nos permitirá determinar los factores que influyan de manera significativa en la probabilidad de ganar un partido de fútbol para los equipos de primera división de la liga peruana en el periodo 2015 - 2019.

Así mismo se plantean las siguientes hipótesis específicas

- Los años de antigüedad que posee un equipo no influye de manera significativa en laprobabilidad de ganar un partido de fútbol.
- La cantidad de tarjetas amarillas obtenidas influye de manera significativa en laprobabilidad de ganar un partido de fútbol.
- La cantidad de tarjetas rojas obtenidas influye de manera significativa en la probabilidadde ganar un partido de fútbol.
- La cantidad de goles anotados influye de manera significativa en la probabilidad deganar un partido de fútbol.
- La edad de los jugadores influye de manera significativa en la probabilidad de ganar unpartido de fútbol.
- El jugar en la modalidad de local influye de manera significativa en la probabilidad deganar un partido de fútbol.
- El tener una mayor cantidad de jugadores extranjeros influye de manera significativa enla probabilidad de ganar un partido de fútbol.
- La altitud del estadio donde se disputa el encuentro no influye de manera significativaen la probabilidad de ganar un partido de fútbol.
- La altura de los jugadores influye de manera significativa en la probabilidad de ganarun partido de fútbol.
- El peso de los jugadores influye de manera significativa en la probabilidad de ganar unpartido de fútbol.
- El porcentaje de posesión de balón influye de manera significativa en la probabilidad deganar un partido de fútbol.
- La cantidad de substituciones no influye de manera significativa en la probabilidad deganar un partido de fútbol.

En este capítulo se hizo una breve introducción del sector deportivo, su relación con la economía y la problemática a investigar. En el segundo capítulo se realizará el marco teórico dentro del cual estarán los antecedentes, idénticos o similares, al tema a investigar y los principales conceptos asociados al futbol. El tercer capítulo se enfocará en explicar lasvariables y metodología a utilizar para obtener los resultados. En el cuarto y quinto se muestran los resultados obtenidos y conclusiones, respectivamente.

METODO

Antecedentes

Revisión Literaria del Tema de Investigación

Autor y año	Objetivo	Unidad de análisis y datos	Variables		Metodología	Principales resultados y Conclusiones
			Dependientes	Independientes		
Magel y Melnykov (2014)	Desarrollar modelos que puedan utilizarse para predecir los resultados de partidos europeos de futbol	Las primeras 33 rondas de futbol disputadas en la liga española, inglesa y italiana durante la temporada 2011 - 2012	gh-ga: Diferencia entre goles anotados por los equipos locales y visitantes 1 = local gano o empato y 0 = local perdio	X1: Σ diferencias entre la cantidad de goles recibidos por un equipo local y la cantidad de goles recibidos por su contricante en las ultimas k rondas X2: Σ diferencias entre la cantidad de goles recibidos por el equipo visitante y la cantidad de goles recibidos por su contricante en las ultimas k rondas X3: Σ diferencias entre la cantidad de tarjetas recibidas por un equipo local y la cantidad de tarjetas recibidas por su contricante en las ultimas k rondas X4: Σ diferencias entre la cantidad de tarjetas recibidas por el equipo visitante y la cantidad de tarjetas recibidas por su contricante en las ultimas k rondas X5 y X6: indican los paises (España, Inglaterra y Italia)	Modelo de regresion MCO Modelo de regresion Logit	Los modelos predijeron corectamente entre el 73% - 79% de los partidos usando los valores recopilados de X1, X2, X3 y X4 sobre los ultimos k partidos disputados Usar las variables en base a las ultimas 4 rondas de partidos disputados entre dos equipos dieron, aproximandamente, los mismos resultados que cuando se utilizaron las ultimas 6, 8, 10 y 12 rondas
Liu et al. (2015)	Determinar las relaciones entre 24 estadisticas deportivas y el resultado de un partido de futbol (victoria, derrota y empate)	48 partidos de la fase grupos del Mundial Brasil 2014 38 partidos cerrados de la fase de grupos del Mundial Brasil 2014	Resultado del partido (victoria, derrota y empate)	Asociadas a los goles: Shot, Shot on Target, Shot Blocked, Shot from Open Play, Shot from Set Piece, Shot from Counter Attack, Shot from Inside Area, Shot from Outside Area Asociadas a pases y organización: Ball Possesion (%), Pass, Pass Accurracy (%), Long Pass, Short Pass, Through Ball, Average Pass Streak, Cross, Dribble, Offside, Corner, Aerial Advantage (%) Asociadas a defensa: Tackle, Foul, Yellow Card, Red Card	Modelo Lineal Generalizado / Regresion Logistica Acumulativa	Shot from Counter Attack, Ball Possesion, Short Pass y Average Pass Streak tienen un efecto significativo en la probabilidad de ganar (efectos fueron mas fuertes en partidos cerrados) Equipos deben plantear estrategias proactivas en base a las variables que tuvieron efectos positivos en la probabilidad de ganar
Yue et al. (2014)	Identificar los factores mas importantes para ganar un partido de futbol y, de esa forma, determinar las tacticas mas adecuadas	126 partidos disputados en las 14 jornadas de la Bundesliga durante la temporada 2011	Resultaldo del partido	E: La eficiencia de goles (numero de goles/numero de tiros) S: Numero de tiros del equipo en el partido P: Numero de pases del equipo en el partido C: Numero de contactos con el balon del equipo en el partido D: Distancia promedio del equipo en el partido R: Numero de piques del equipo en el partido OO: Numero de "uno y uno" del equipo en el partido	Analisis Estadistico	La eficiencia de goles es por mucho el parametro de equipo mas importante para el resultado de un partido (calidad de goles es mas importante que la cantidad de disparos para ganar un partido) Los resultados favorecen a "Juego Directo" en lugar de "Juego de Posesion" debido a que presenta mayor eficiencia de goles
Oberstone (2009)	Identificar los factores de campo mas importantes que expliquen el éxito de un club futbolistico	20 equipos de la Liga Premier Inglesa durante la temporada 2007 - 2008	Éxito del equipo (puntos ganados durante la temporada)	Promedio de goles por partido, Numero de tiros a puerta, Tiros en el blanco (%), Goles a disparar (%), Goles marcados fuera del area (%), Numero total de pases, Proporcion de pases cortos/largos, Finalizacion general del pase (%), Numero total de cruces, Finalizacion de cruces (%), Promedio de goles concedidos por partido, Numero de placajes, Placajes ganados (%), Numero de bloqueos, espacios libres e intercepciones, Numero de faltas, Numero de tarjetas amarillas, Numero de tarjetas rojas	Modelo de Regresión Multiple	Goles marcados fuera del area (%), Goles a disparar (%), Numero de tarjetas amarillas, Proporcion de pases cortos/largos, Numero total de cruces y Promedio de goles concedidos por partido mostraron ser significativos para el éxito de un club de la Liga Premier Inglesa
Lopez et al. (s.f.)	Examinar los factores que intervienen en la probabilidad de ganar un partido de futbol americano	256 partidos disputados en la NFL durante la temporada 2015 - 2016	PROBGAN: Probabilidad de ganar un partido (1 = victoria y 0 = derrota)	YDSAIRE: Yardas por aires obtenidas por el mariscal de campo YDSTIERRA: Yardas por tierra obtenidas por los jugadores que tuvieron acarreos RATING: Eficiencia del pase TERCERA: Conversiones efectivas en tercera oportunidad (%) DIF: Diferencia entre balones recuperados y entregados sea por balon suelto o intercepcion CAPTURAS: Capturas obtenidas por jugadores defensivos sobre el mariscal YDSDES: Yardas logradas en devolucion de despejes	Modelo Probit Multivariado	YDSTIERRA, RATING y CAPTURAS tienen un claro impacto en la probabilidad de ganar un partido de fútbol americano Se considera valioso a un mariscal de campo que mantenga un RATING y TERCERA altos. A su vez, una defensiva que recupere balones y presione al mariscal opositor (CAPTURAS) es igual de valiosa.

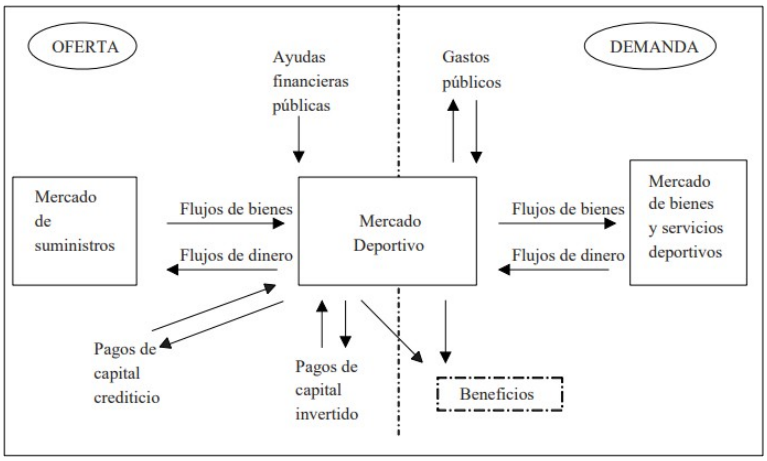
Alves et al. (2010)	Predecir las posibilidades de que un equipo clasifique a la Copa Libertadores y, también, su clasificación al final de este	20 equipos participantes del Campeonato brasileño del año 2008	Wj: Indicador del evento del j-esimo equipo ganador ; j varia de 1 a 20 Lj: Indicador del evento del j-esimo equipo perdedor; j varia de 1 a 20	Xi: Xi = 1 para i = j y el i-esimo equipo que juega de local, Xi = -1 para i = j y el i-esimo equipo que juega de visitante y Xi = 0 en caso contrario ; i varia de 1 a 20 Zi: Zi esta dado por el % puntos ganados por el i-esimo equipo en los ultimos 4 partidos jugados de local si i = j y el j-esimo equipo juega de local, Zi esta dado por el % puntos ganados por el i-esimo equipo en los ultimos 4 partidos jugados de visitante si i = j y el j-esimo equipo juega de visitante y Zi = 0 en caso contrario ; i varia de 1 a 20	Modelo Logit Ordinal	Modelo 1 presenta predicciones más estables, pero es moroso para tener en cuenta los cambios en el comportamiento de los equipos (considerar el desempeño reciente del equipo como variable explicativa mejora el modelo en dicho aspecto) Se pueden emplear otras variables como la fuerza de los equipos (en términos de puntos ganados) para tener en cuenta otros aspectos y detalles
Gómez (2013)	Creación de modelos de predicción para los equipos ingresantes a la etapa final del FPC	270 partidos comprendidos entre las fechas 1 - 15 de la FPC durante la temporada 2013	Resultado del equipo (1 = gana, 0.5 = empate, 0 = derrota) Goles anotados en partido anterior Goles recibidos en partido anterior	Goles anotados en partido anterior, Goles recibidos en partido anterior, Puntos acumulados, Resultado del equipo (1 = gana, 0.5 = empate y 0 = derrota), Goles acumulados a favor del contricante, Posición del equipo en la fecha anterior, Local (1 = equipo es local y 0 = es visitante), Equipo en torneo internacional (1 = si esta en torneo internacional y 0 = lo contario) y Suma de resultados en las últimas 5 fechas	Modelo Logit Ordenado	Los 3 modelos contribuyen en pronosticar los equipos que jugarán en las 3 fechas restantes del FPC Limitantes como la falta de variables y datos necesarios afectaron a que los resultados no tengan el máximo porcentaje de acierto posible
Harrop y Nevil (2014)	Identificar factores de desempeño que discriminen entre partidos que un equipo ganó, empató y perdió Identificar aquellas variables que mejor pronostiquen el éxito del equipo	46 partidos jugados por un equipo perteneciente a la Liga 1 Inglesa durante la temporada 2012 - 2013	Resultaldo (1 = equipo gana y 0 = empata o pierde)	Asociadas a la ofensiva: Total de tiros, Tiros en el blanco, Tiros dentro del área de penalti, Pases, Pases Exitosos (%), Pases en el campo contrario, Faltas recibidas, Regates, Cruzados y Saques de esquina y fuera de juego (cometidos) Asociadas a la defensa: Faltas cometidas, Cruzados en contra, Saques de esquina en contra, Tarjetas amarillas y Tarjetas Rojas Asociadas al contexto: Ubicación del partido (1 = equipo juega de local y 0 = como visitante)	Analisis Estadistico (Prueba Kruskal Wallis) Regresión Logística Binaria	Los resultados apoyan investigaciones previas, las cuales sugieren que Pases / Posesión y Tiros a objetivo son variables que influyen en las probabilidades de éxito de un equipo La habilidad del equipo analizado es insuficiente para mantener una posesión clave del balón. Se debe reducir la cantidad de regates / pases intentados y incrementar el % de pases exitosos / tiros a objetivo en la medida de lo posible. La ubicación puede afectar el rendimiento del equipo y sus chances de éxito
Fuqua (2014)	Predecir los cuatro finalistas en un torneo de baloncesto	510 equipos participantes del Campeonato de Primera División de Baloncesto Masculino (NCAA) entre las temporadas 2006 - 2013	1 = equipo llego a semifinales y 0 = lo contrario	PtsPer100Poss: puntos por cada 100 posesiones (medida de la ofensiva) PtsPer100PossAllow: puntos por cada 100 posesiones permitidas (medida de la defensa) RBSRate: tasa de rebotes (medida del control del balón) SOS: fuerza del horario (para normalizar las estadisticas basadas en el nivel de competencia) REGSTR: fuerza regional (para tener en cuenta las variaciones en la fuerza de las regiones)	Modelo Logit Binomial	El modelo funciona, mejor que cualquier otro sistema de clasificación actual, para predecir los equipos entrantes a la etapa de semifinales del NCAA
Goddard y Asimakopoulos (2004)	Pronosticar los resultados de partidos de fútbol de la Liga Inglesa	19,744 partidos entre las temporadas 1990 – 1999 de los 20 mejores equipos de la Liga Premier y los 72 equipos de las Ligas 1, 2 y 3 (24 por cada liga)	Yij: Resultado de un partido entre los equipos i y j (1= victoria del local, 0.5 = empate y 0 = victoria del visitante)	$P^{d}_{i,y,s}$: puntuación total del equipo local i $R^{H}_{i,m}$: resultado del m-esimo partido más reciente jugado en casa por el equipo local i (m = 1,…,M) $R^{A}_{i,n}$: resultado del n-esimo partido más reciente jugado fuera de casa por el equipo local i (n = 1,…,N) $SIGH_{i,j}$: 1= partido tiene importancia de campeonato, ascenso o descenso para equipo local i pero no para el visitante j, 0 = lo contrario $SIGA_{i,j}$: 1= partido tiene importancia para el equipo visitante j pero no para el local i, 0 = lo contrario CUP_i: 1= equipo local es eliminado de la Copa FA, 0 = lo contrario $DIST_{i,j}$: log. natural de la distancia geograica entre los terrenos de los equipo i y j $P^{d}_{f,y,s}$ $R^{H}_{f,m}$ $R^{A}_{f,n}$ y CUP_j: caso analago para el equipo visitante j	Modelo Probit Ordenado	Los partidos con importancia de campeonato, ascenso o descenso, la participación y desenvolvimiento de los equipos en Copas (FA) y la distancia geografica entre las ciudades de origen de los equipos rivales resultan significativas y contribuyen al modelo de pronostico de resultados

Fuente: Elaboración Propia

Relación economía y deporte

La economía del deporte es un área que ha ido ganando terreno con el transcurso del tiempo. Ha pasado de ser una simple practica con fines recreativos a considerarse, como se mencionó anteriormente, un bien de carácter intangible que guarda relación con aspectos económicos. Es decir, ha pasado a considerarse un sector económico más ya que proporciona empleo, genera riqueza y produce bienes y servicios que se utilizan con mucha frecuencia hoy en día. Antes era poco común encontrar investigaciones econométricas referente al sector deportivo, pero ahora se puede apreciar una mayor cantidad de estos que tratan de explicar diversos fenómenos como el mercado de apuestas deportivo, la medición del salario de un jugador en base a una cierta cantidad de factores, lo que puede influir en el desempeño de un equipo, entre muchos más. La microeconomía, la macroeconomía y la econometría son herramientas fundamentales para explicar las relaciones económicas y sociales que se producen en el sector deportivo (Fiallo, 2017). El análisis del deporte desde el punto de vista microeconómico gira en torno al mercado de bienes y servicios que los distintos agentes ofertan y demandan. El consumo de estos dependerá de la manera en que son ofertados en el mercado, así como de las preferenciasy necesidades del individuo en dicho momento.

Figura 2 - Funcionamiento del mercado deportivo



Fuente: Heinemann

Los bienes y servicios que se ofertan y demandan son diversos (ropa e implementos deportivos, alquiler de canchas recreativas, literatura deportiva, etc.). Y siendo el fútbol uno de los deportes más importantes del mundo, la proporción de dichos bienes y servicios es mucho mayor. Por otro lado, la macroeconomía del deporte analiza la influencia que tiene este en las variables agregadas de un país como inversión, empleo, consumo, producto bruto interno (PBI), inflación, entre otros. Por lo tanto, uno de los enfoques relevantes para estudiar el deporte desde el punto de vista macroeconómico se da a partir del modelo de oferta y demanda agregadas (Mesa y Arboleda, 2007). La condición de equilibrio macroeconómico expresa que tanto la demanda como oferta agregadas deben ser equivalentes, por lo que un incremento en el gasto deportivo generará mayores ingresos en dicho sector, lo cual impulsa a que exista más inversiones que causa al mismo tiempo un aumento en la demanda de consumo (a esto se le conoce como efecto multiplicador).

Teoría de la producción económica

Se pueden analizar aspectos como el desempeño de un equipo, su nivel de ingresos, porcentaje de victorias, entre otros más a través de la teoría de la producción. De manera más específica, a través de funciones de producción donde la variable a medir se conoce como “output” y las que se usarán para realizar tal medición se denominan “inputs” o variables de salida. Diversos autores han utilizado funciones productivas en estudios económicos del deporte, siendo el primero de estos Scully (1974) que realizó una comparación entre los salarios y el producto de ingresos marginales de jugadores de la Liga Mayor de Baseball (MLB). Modeló una función donde utiliza el porcentaje de victorias del equipo por temporada, entre 1968 y 1969, como output y variables deportivas asociadas al baseball como inputs. Más adelante se analizará esta modelación a detalle.

En general, una función de producción se denota de la siguiente forma:

$$Q = f(X_1, X_2, \ldots, X_n)$$

Donde Q es el máximo nivel de producción que se puede llegar a tener en base a la combinación especifica de factores X_i que se está utilizando (Pindyck y Rubinfield, 2009). En el caso donde solo se esté utilizando dos factores, la función de producción seexpresa de la siguiente forma:

$$Q = f(L, K)$$

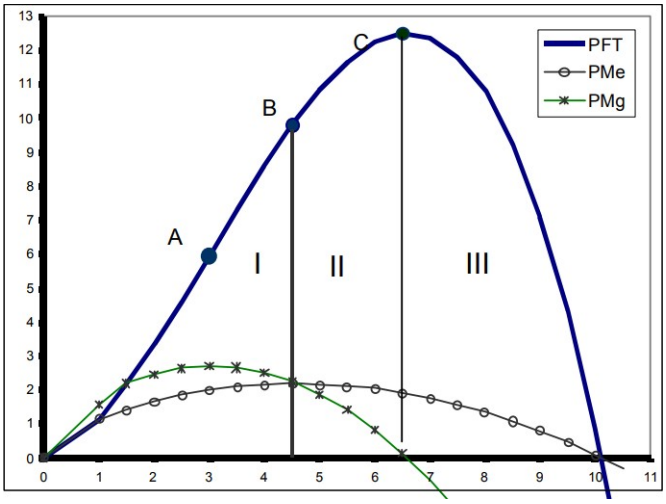
Donde L y K son trabajo y capital, respectivamente. La producción está fuertemente relacionada con la tecnología por lo que, dada una cierta cantidad de factores, la cantidad de Q que se obtendrá dependerá también del estado de conocimientos tanto técnicos como científicos utilizados en ese momento. La eficiencia técnica es otro concepto que brota cuando se entra a la teoría productiva, el cual dicta que las funciones de producción explican lo que es técnicamente viable cuando se produce de forma eficiente. No siempre se puede suponer lo último, pero es racional pensar que las empresas quieran alcanzar el máximo beneficio sin desperdiciar recursos.

El comportamiento de las funciones de producción varía según el horizonte de tiempo. En el corto plazo no es posible cambiar las cantidades de uno o más factores productivos. Es decir, existirá por lo menos un factor que no pueda alterarse y permanecerá fijo (usualmente ese factor es K) por lo que, para aumentar la producción, se debe de incrementar la cantidad de factores variables que se está utilizando. Por tanto, una función de producción a corto plazo se expresa de la siguiente forma:

$$Q = f(L, K^*)$$

Donde K^*es la cantidad fija de capital

Figura 3 - Función de producción a corto plazo



Fuente: Arzubi (2003)

Las funciones productivas a corto plazo poseen tres etapas en donde toman participación las curvas de producto medio (PMe) y marginal (PMg). En la primera etapa, el producto marginal es mayor al medio en un principio, pero a medida que avanza la curva el producto marginal decrece y el medio alcanza su máximo punto igualando al marginal. En la segunda ambos decaen (mas no se vuelven negativos) hasta el punto donde el producto marginal se torna nulo. Finalmente, en la última etapa el producto medio se mantiene mayor (pero decreciente) al marginal mientras que este último se tornanegativo.

Figura 4 - Etapas de la función de producción a c/p

Etapa I:	El producto medio crece. PMg > PMe.
Etapa II:	El producto medio decrece mientras el producto marginal es positivo. PMg < PMe.
Etapa III	El producto marginal es negativo y el Producto medio continua decreciendo. PMg < PMe.

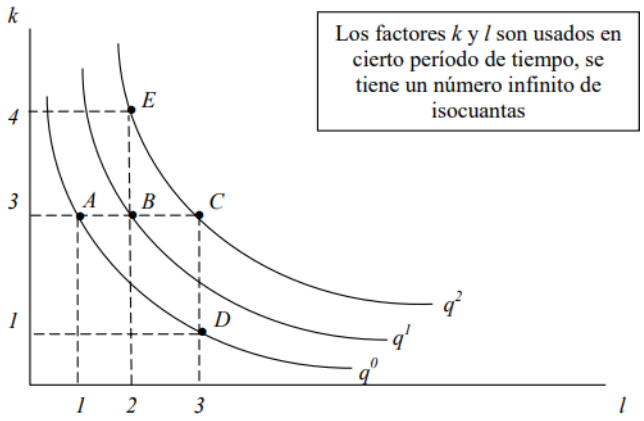
Fuente: Arzubi (2003)

En el largo plazo, tanto el trabajo como capital son variables debido a que se adecuan a las condiciones en las que se encuentra el mercado. Se utiliza una gran cantidad de trabajo y muy poco capital, una gran cantidad de capital y muy poco trabajo, o cantidades equilibradas de ambos factores (Rengifo, 2019). Cuando se estudia la función de producción a largo plazo, aparece el concepto de las isocuantas o isoproductos. Estas son curvas que muestran todas las combinaciones posibles de factores que generan el mismo nivel de producción (Pindyck y Rubinfield, 2009), y se llega a representar en la función de la siguiente manera:

$$q_i = f\,(L,\ K)$$

Donde q_i es el nivel de producto. Las curvas isocuantas tienden a ser convexas al origen, decrecientes y no se interceptan entre sí ya que, de ser el caso, se estaría argumentando que se pueden obtener dos niveles diferentes de producción con una mismacombinación de factores lo cual es imposible.

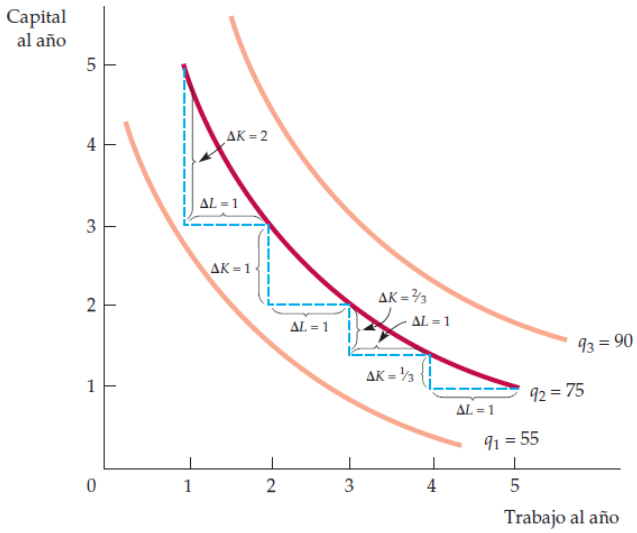
Figura 5 - Curvas isocuantas



Fuente: Mendieta (2005)

La pendiente de una curva isocuanta se le conoce como Relación Marginal de Sustitución Técnica (RMST). La RMST, similar a la Relación Marginal de Sustitución (RMS) de la teoría del consumidor, nos indica como se puede sustituir un factor productivo por otro manteniendo el nivel de producción constante. Es decreciente a medida que se desplaza en sentido descendente a lo largo de una isocuanta (Rengifo, 2019) lo cual refuerza la propiedad convexa de estas curvas.

Figura 6 - Relación Marginal de Sustitución



Fuente: Pindyck y Rubinfiled

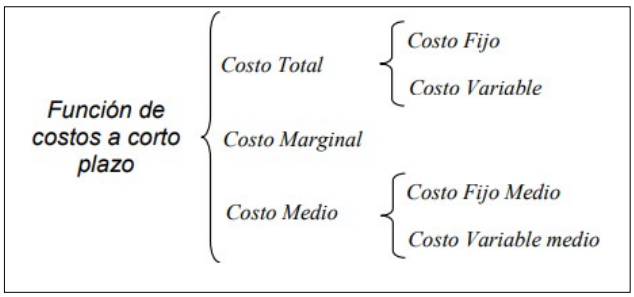
Por otro lado, se debe de analizar también los costes que implica el realizar un proceso productivo. Esto se hace mediante la función de costos de producción la cual muestra el valor mínimo para obtener un determinado nivel productivo, considerando constantes los precios de los factores. Existen distintos tipos de costos, siendo los más principales (Pindyck y Rubinfield, 2009):

- Costo de oportunidad: es el costo correspondiente a las oportunidades que se pierden cuando no se utilizan los recursos para el fin que posee un valor más alto.

- Costo económico: es el costo que tiene para una empresa la utilización de recursoseconómicos en la producción, incluido el costo de oportunidad.
- Costo hundido: es aquel gasto que no se pueden volver a recuperar bajo ningún motivouna vez que se hayan realizado.
- Costo contable: son los gastos reales más los de depreciación del equipo de capital.

Al igual que en la función de producción, la de costes se puede analizar según el tamaño de su horizonte de tiempo. Como ya se mencionó, a corto plazo siempre habrá un factor que permanece invariante mientras los demás si lo hacen cuando se altera le nivel productivo. Aquí toman participación tres conceptos importantes de costos, los cuales son:

Figura 7 - Función de costos a corto plazo



Fuente: Rengifo (2019)

El costo total (CT) es la sumatoria de todos los factores utilizados en el proceso productivo. Se divide en costos fijos y variables y esta denotado de la siguiente forma:

$$CT = CF + CV$$

Respecto al costo marginal (CMg), es básicamente la variación incremental que experimenta el costo total cuando se produce una unidad adicional de producción.
Siendo

específicos, la variación que experimenta el costo variable con dicha unidad productiva adicional (costo fijo invariante). Su formulación es:

$$CMg = \Delta CT/\Delta Q \quad o \quad CMg = \Delta CV/\Delta Q$$

El coste medio (CMe) es el costo total entre cada unidad que se ha producido. Es decir, es el costo total de la empresa dividido por su nivel de producción (Rengifo, 2019). Este tipo de costo, como en el caso del costo total, se encuentra compuesto por el costo fijo medio (CFMe) y costo variable medio (CVMe) los cuales comparten la misma definición que el CMe (son los costos fijos y variables entre el nivel de producción).

$$CT/Q = CF/Q + CV/Q$$

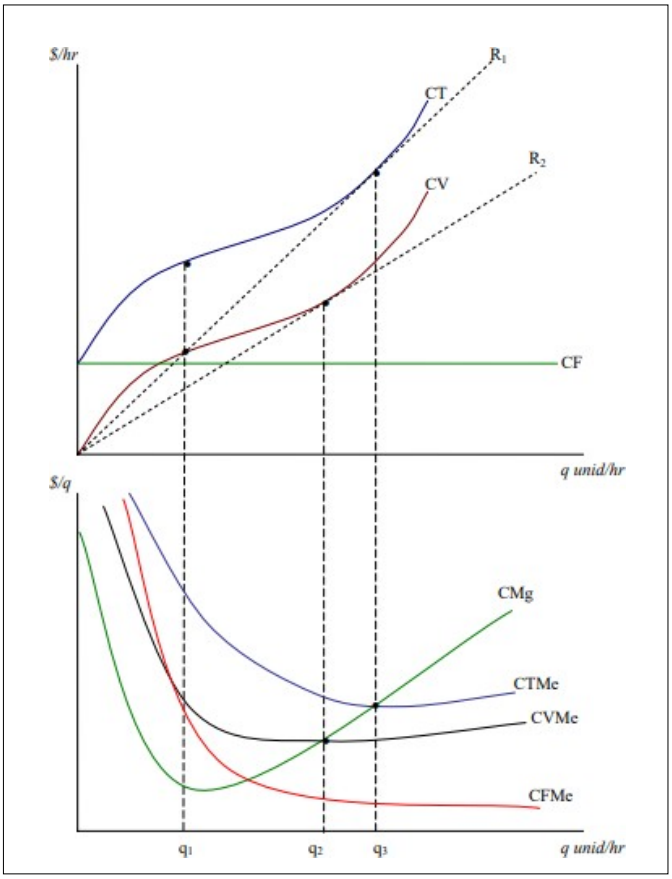
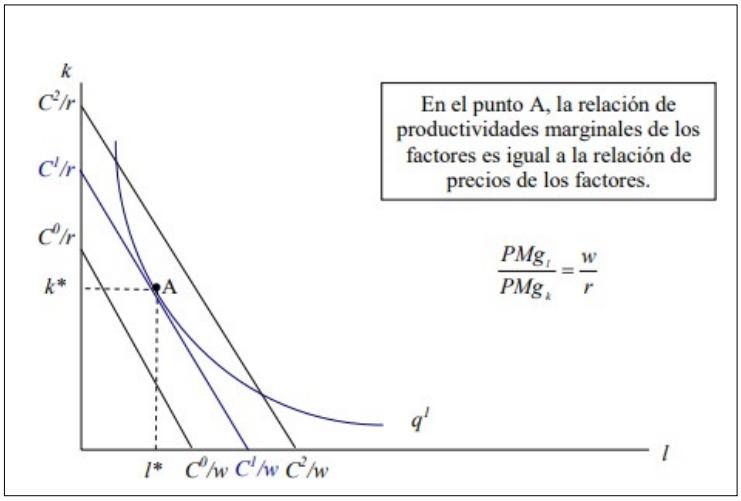


Figura 8 - Curva de costos a corto plazo

Fuente: Mendieta (2005)

En el largo plazo, ya no se cuenta con ningún factor invariante. Por tanto, la elección de los factores depende tanto de los costes relativos de los factores de producción como del grado en que la empresa puede sustituir unos factores por otros en su proceso productivo (Pindyck y Rubinfield, 2009). Aquí entra el concepto de recta isocoste, la cual muestra la cantidad de combinaciones de factores de producción donde cada una de estas poseen el mismo costo. Cuando la recta isocoste se encuentre tangente a la isocuanta, mostrará las cantidades optimas de factores que maximizarán el nivel productivo dados unos costos. Es decir, representa el gasto mínimo respecto a un nivel de producción dado.

Figura 9 - Curva isocoste



Fuente: Mendieta (2005)

El deporte en la teoría de la producción

Volviendo a las funciones de producción deportivas, el output debe ser producido a través de otro productor (en este caso, serían los equipos). La estimación de estas funciones, para una variedad de deportes, se han realizado generalmente con el objetivo de probar cierta teoría económica o describir algún aspecto de la operación de los mercados laborales deportivos. En este tipo de mercados, definir y medir la producción es más complicado a comparación de otros

(Leeds y Von Allmen, 2016). Como se mencionó anteriormente, el estudio de Scully (1974) fue uno de los primeros en lo que

respecta a funciones de producción deportivas. A continuación, se explicará el modelamiento de la función productiva que propuso este autor para hacer la comparación entre el producto de ingresos marginales y salarios de un jugador profesional de beisbol siendo la variable de salida el porcentaje de victorias del equipo y una serie de variables deportivas las de entrada. A pesar de que el objetivo de esta investigación es determinar lo que influye significativamente en la probabilidad de ganar un partido y no los salarios, esta modelación servirá como apoyo al planteamiento econométrico que se está proponiendo en este trabajo.

En dicho estudio, se plantea que la calidad de los partidos de beisbol se mide a través del porcentaje de victorias *W*, el cual se encuentra relacionado con dos categorías generales de inputs: un vector de habilidades del jugador A_i y un vector de inputs de no jugadores (entrenadores, gerentes, etc.) I_j. Por lo tanto:

$$W = W\,(A_1, A_2, \ldots, A_n; I_1, I_2, \ldots, I_m)$$

A su vez, los equipos obtienen ingresos, principalmente, a través de las entradas vendidas el día del partido y los derechos de transmisión televisiva y radial. Se afirma que lo antes mencionado se relaciona de forma directa con el porcentaje de victorias de un equipo y el área poblacional, y de forma indirecta con el desempeño del jugador. Entonces, el estudio expresa lo siguiente:

$$R = p.\ T\,[W\,(A_i, I_j), P_g] + B\,[W\,(A_i, I_j), P_b]\ ;\ \ i = 1, \ldots, n\ y\ j = 1, \ldots, m$$

Donde *R* es el ingreso del equipo, *p* el precio de entrada al estadio, *T* el número de entradas vendidas, *W* el desempeño del equipo, P_g la población potencial atraída al beisbol, P_b potenciales transmisora televisas o radiales y *B* los ingresos por transmisión. Por otro lado, el costo *C* del equipo se encontrará determinado por el nivel de habilidad de los jugadores y el nivel de inputs de los no jugadores. Como el mercado de trabajo delos equipos deportivos posee características monosopnicas, los costos de jugador estarán relacionados de manera endógena a su nivel de habilidad.

Entonces, se define los costos del equipo como:

$$C = \Sigma_i A_i S_i (A_i) + \Sigma_j r_j I_j \quad ; \quad i = 1, \ldots, n \text{ y } j = 1, \ldots, m$$

Donde S_i (A_i) son funciones de suministro de los jugadores y r_j factores de remuneración de los no jugadores. Las ganancias del equipo se encuentran definidas como:

$$\Pi = R - C$$

Las condiciones de primer orden para un máximo son obtenidas diferenciando con respecto a A_i y I_j:

$$\frac{\partial \pi}{\partial Ai} = p \frac{\partial T}{\partial W} \frac{\partial W}{\partial Ai} + \frac{\partial B}{\partial W} \frac{\partial W}{\partial Ai} - Ai \frac{\partial Si}{\partial Ai} - Si \quad ; \quad i = 1, \ldots \ldots, n$$

$$\frac{\partial \pi}{\partial Ij} = p \frac{\partial T}{\partial W} \frac{\partial W}{\partial Ij} + \frac{\partial B}{\partial W} \frac{\partial W}{\partial Ij} - rj \quad ; \quad j = 1, \ldots \ldots, n$$

Los resultados de las condiciones de primer orden revelan que los equipos maximizan sus ganancias seleccionando un nivel de habilidades de jugador y inputs de no jugadores tal que estos reciban un salario equivalente a sus productos de ingresos marginales, restándole las rentas que se consideran monopsónicas.

Los otros factores son retribuidos igual a sus productos de ingreso marginal. La manifestación de beneficios monopólicos se presenta en el modelo, donde puede llegar a tomarse en cuenta si se observa que:

$$\frac{\partial \pi}{\partial Pg} = p \frac{\partial T}{\partial Pg} > 0$$

$$\frac{\partial \pi}{\partial Pb} = \frac{\partial B}{\partial Pb} > 0$$

Utilizando todo esto, Scully (1997) plantea el siguiente modelo:

$$PCTWIN_{it} = \alpha + \beta TSA_{it} + XTSW_{it} + \delta NL_{it} + \Phi CONT_{it} + \Phi OUT_{it} + \xi_{it}$$

Donde *TSA* es el promedio de *SLG* del equipo, *TSW* indica la proporción de ponches por base, *NL* y *CONT* son variables categóricas para la Liga Nacional y los ganadores de división la temporada pasada, respectivamente; y *OUT* es también una categórica para los equipos que al final de la temporada se encontraban 20 o más partidos fuera de lugar. Este estudio sirvió como puente a que se realizarán mayores investigaciones económicas en el ámbito deportivo utilizando funciones de producción. Bairam, Howells y Turner (2006) utilizan estas funciones para el caso del cricket australiano y neozelandés.

Otros Carmichael y Thomas (1995) lo aplican al rugby, donde analizan la producción y eficiencia de los equipos deportivos. A su vez, Borland (2006) planteó la forma general de una función de producción para un equipo, expresándose de la siguiente forma:

$$Pit = y\ (Q_{1it},\ \ldots\ldots,\ Q_{Jit}\ M_{Jit},\ X_{it},\ T_{1it},\ \ldots\ldots,\ TN_{it})\ \ ;\ i = 1,\ \ldots,I\ ;\ t = 1,\ \ldots,T$$

Donde *P* es el desempeño o la salida. La mayoría de estudios que estiman funciones de producción, en base a los equipos, han utilizado la temporada como el periodo de tiempo para analizar el rendimiento y el porcentaje de victorias como el output. El modelo tiene al rendimiento del equipo *i* en el momento *t* dependiente de la calidad *Q* de los jugadores *J* en el equipo, la calidad del entrenador *M,* otros factores *X* y la calidad de los otros *N* equipos de la competencia. Respecto a la elección de la forma funcional, determinar la apropiada requiere de un análisis de la manera en que los inputs se combinen para generar el output del equipo y de cómo le afectarán variaciones incrementales. Los principales enfoques en estudios de funciones de producción, en base a equipos deportivos, han tomado modelos que especifican una relación lineal o logarítmica entre el output y los inputs. Los lineales suponen una separación implícita aditiva de los inputsmientras que los logarítmicos suponen una interacción multiplicadora entre inputs individuales que determinan el output del equipo.

Proceso metodológico

En esta sección se va a explicar a detalle todo lo referente a los datos y la modelación econométrica que se utilizará para obtener las estimaciones y resultados principales. Como se mencionó anteriormente, el objetivo de esta investigación es determinar los factores que influyan de manera significativa en la probabilidad de ganar un partido de fútbol para los equipos de primera división de la liga peruana en la Copa Movistar en el periodo 2015 - 2019. Por lo tanto, la modelación que más se adecua para obtener los resultados es la de un modelo de respuesta categórica de datos de panel o, dicho de otra forma, se planteará tanto una modelación Logit Ordenado como Logit Ordenado Generalizado debido al

carácter categórico que posee la variable dependiente planteada (toma el valor de 0 si el equipo perdió, 1 si empato y 2 si ganó).

Tipo y diseño de investigación

Claro está, que la investigación a realizar es del tipo cuantitativa ya que se está recopilando y examinando datos sobre las variables que consideramos de interés. La diferencia fundamental entre la investigación cuantitativa y cualitativa es que la primera estudia la asociación o relación entre variables cuantificadas y la segunda lo hace en contextos estructurales y situacionales (Strauss, 1987). A su vez, es del tipo no experimental debido a que no se está alterando o manipulando de ninguna forma los datos de las variables de interés. Es decir, se está realizando el análisis en la forma natural de la información a recopilar. En lo que respecta al diseño de la investigación, como ya se mencionó, comprende básicamente la modelación de un panel de datos ya que se cuenta con datos a nivel individual para distintos periodos de tiempo. Asociando lo último a la problemática de investigación, se tienen datos para la unidad de análisis de interés que son los partidos para los años comprendidos desde el 2015 al 2019 que es el horizonte detiempo planteado.

Participantes

La unidad de análisis son los partidos disputados en cada temporada de nuestro horizonte de tiempo. Se planteó de manera inicial que fueran los equipos, pero surgieron inconvenientes en lo que respecta a la recopilación de datos. Por ejemplo, se trató de encontrar una variable que cuantifique la cantidad de partidos ganados respecto al total o algún porcentaje de ese tipo. Sin embargo, tratándose de una investigación econométrica deportiva peruana, la carencia de datos jugó en contra de considerar los equipos como unidad de análisis para la investigación. Es por eso que se recopilará datos respecto a cada partido jugado de los equipos de primera división de la Copa Movistar. Siendo específicos, los disputados en las modalidades de verano (grupos A y B), apertura, clausura, semifinales y finales. Cabe mencionar, que la estructura de la Copa Movistar ha variado en algunos años. En las temporadas 2015, 2016 y 2019 no hubo torneo de verano, solamente se

disputaron partidos de apertura, clausura, semifinales y final. Sin embargo, en las temporadas 2017 y 2018 si se realizó este torneo. Se puede generar cierta incertidumbre de si esto puede llegar a afectar la investigación al momento de hacer las estimaciones, pero, como lo que nos interesa son la cantidad de partidos disputados (unidad de análisis), no generará inconvenientes. Respecto a la muestra o cantidad de observaciones, se disputaron un total de 1,662 partidos en las temporadas 2015 (281), 2016 (359), 2017 (354), 2018 (358) y 2019 (310). Para cada partido se recopilarán datos de las variables independientes descritas anteriormente.

Tabla 2 - Variables del modelo

$Outcome_{it}$	Variable categórica: 0 = perdió, 1 = empatoy 2 = gano
$Antiquity_{it}$	Años de antigüedad del equipo por partido
$Yellow\_card_{it}$	Cantidad de tarjetas amarillas por partido
Red_card_{it}	Cantidad de tarjetas rojas por partido
$Goals_{it}$	Cantidad de goles anotados por partido
Age_{it}	Edad promedio de los jugadores por partido
$Local_{it}$	Variable categórica: 1 = local y 0 = visitante
$Foreign_{it}$	Cantidad de jugadores extranjeros por partido
$Altitude_{it}$	Altitud del estadio por partido
$Height_{it}$	Altura promedio de los jugadores por partido
$Weight_{it}$	Peso promedio de los jugadores por partido
$Possesion_{it}$	Posesión del balón por partido (%)
$Substitution_{it}$	Cantidad de substituciones por

	partido

Instrumentos

La fuente de donde se obtendrán los datos, como se mencionó en la parte del planteamiento del problema, son de las páginas webs "CeroAcero.es" y "Resultados- fútbol.com" que muestran información de diferente tipo respecto al fútbol (noticias, fichajes, clasificaciones, etc.). Lo que se destaca es que ambas poseen una sección con estadísticas de varios equipos y ligas del mundo, lo cual servirá de gran herramienta para el objetivo de esta investigación. Los datos de la mayoría de las variables se obtendrán de estas fuentes, con excepción de la variable que hace referencia a la altitud del estadio por partido (*Altitude*) la cual se obtendrá a través del aplicativo Google Earth.

Procedimiento

A continuación, se pasará a realizar un ahondamiento sobre la metodología de datos de panel y las ventajas y desventajas de su aplicación. Luego, se explicará en que consiste un modelo de respuesta binaria (Logit) de datos de panel y sus principales características.

A su vez, se ahondará en los modelos Logit Ordenado y Logit Ordenado Generalizado, los cuales servirán como 31ltimo31als31s primordial para obtener los resultados 31ltimo31als del presente trabajo de investigación y, por 31ltimo, se realizará la especificación del modelo.

Los paneles de datos

La naturaleza de los datos influye de manera significativa en el análisis econométrico a utilizar. Cuando se quiere examinar a uno o más individuos en un mismo instante en el tiempo, la aplicación de corte transversal es la más adecuada. Si lo que se desea es analizar una cierta cantidad de variables en distintos periodos temporales, una metodología de series de tiempo es la mejor herramienta para lograr ese cometido y si, por último, se quiere dar seguimiento (o se intenta) a los mismos individuos, empresas, ciudades, países o cualquier otra cosa a lo largo del tiempo una modelación de series de tiempo es lo más recomendable (Wooldridge,

2015). Es decir, combina de cierta manera ambos tipos de modelaciones (corte transversal y series de tiempo) y es un instrumento de mucha utilidad cuando se quiere realizar un análisis econométrico de índole microeconómico. Y siendo el propósito de esta investigación observar los factores que influyen en la probabilidad de ganar un partido de fútbol durante un periodo determinado, esta metodología es la que se debe utilizar. La modelación de datos de panel ha ido ganado fuerza con el paso del tiempo debido, principalmente, a la mayor necesidad de examinar fenómenos de carácter individual en el tiempo y, también, a la optimización tecnológica en lo que respecta a recopilación de bases de datos más detalladas. El principal objetivo de aplicar y estudiar los datos en panel, es la de capturar la heterogeneidad inobservable (Mayorga y Muñoz, 2000) que se presenta de manera constante en el tiempo de lo que se quiere analizar (países, estados, departamentos, etc.). La heterogeneidad inobservable es el error generado al no disponer de ciertas variables debido a su carácter no observable, pero que si poseen correlación con las que sí y es aquí donde entran los modelos de efectos fijos y aleatorios ya que ambos son métodos que permiten realizar estimaciones aceptadas aún en presencia de este error. Se explicarán estos modelos más adelante.

La especificación de una regresión con datos de panel (según Burdisso, 1997) es la siguiente:

$$Y_{it} = \alpha + X_{it}\beta + u_{it} \quad ; \quad i = 1, \ldots, N \quad y \quad t = 1, \ldots, T$$

Donde α es un escalar, β un vector de K parámetros, X_{it} se refiere a la i-ésima observación en el momento t para las K variables independientes, u_{it} el término de error, i = 1, …,N son las observaciones de corte transversal y t = 1, …,T se refiere a las observaciones de series de tiempo. Según el análisis literario, es recomendable que se cuente con un N grande y un t pequeño debido a la existencia actual de bases de datos extensas para horizontes de tiempo pequeños, lo cual le da cierto refuerzo a la creciente frecuencia de modelaciones de datos de panel en estudios de carácter microeconómico. Siguiendo a Burdisso (1997), el error u_{it} puede llegar a desagregarse de la siguiente forma:

$$U_{it} = \mu_i + \delta_t + e_{it}$$

Donde u_i es el error para el individuo i, δ_t los efectos que no logran medirse pero que si llegan a variar en el tiempo y e_{it} el error netamente aleatorio. Tomando como referencia este modelo base se pueden encontrar otros tipos de modelaciones de datos de panel, los cuales se diferencian con el planteado anteriormente en ciertos supuestos y restricciones. Se ahondará en ellos más adelante. Como toda modelación econométrica, trae consigo ciertas ventajas y desventajas al momento de su aplicación y los modelos de datos de panel no son la excepción (no existe algo conocido como el "modelo perfecto"). Por lo tanto, las virtudes y limitantes que posee esta metodología son las siguientes ventajas:

- Se pueden analizar distintas problemáticas de carácter individual que no pueden llegara ser ahondadas por los métodos de series de tiempo y corte transversal.
- Permite realizar estimaciones aceptadas en presencia del error conocido como heterogeneidad inobservable. Es decir, ayuda a controlar los efectos individuales que no se pueden observar directamente (latentes) pero que se llegan a deducir a partir de variables que sí se observan. Esto se puede lograr utilizando distintas herramientas como variables instrumentales, el modelo de efectos fijos, aleatorios y multinivel, entre otros. Las modelaciones de corte transversal y series de tiempo no tratan de corregir este error y usualmente esta denotado como n_i.
- Brinda una mayor cantidad de observaciones (la cual está dada por N x t), menor grado de colinealidad entre variables independientes, mayor precisión en las estimaciones realizadas y mayor rango de variabilidad.

Así mismo se presentan las siguientes desventajas:

- Cuando ya no es posible darle seguimiento. Este limitante se puede observar cuando se realiza recopilación de datos sobre la unidad de análisis de interés en base a encuestas o entrevistas. Ejemplos de este tipo de limitantes son: cobertura de la población a analizar, preguntas no muy claras, distorsión deliberada de respuestas, entre otros (Mayorga y Muñoz, 2000).

- La existencia de paneles incompletos o no equilibrados. Ello causado, principalmente, a la existencia de mayor cantidad de información para ciertos individuos en comparación a otros.

Existen diversos tipos de modelaciones de datos de panel, en base a su planteamiento general, donde se diferencian en ciertas propiedades, supuestos y restricciones ya sean matemáticas, estadísticas, econométricas, etc. En una estim ación con datos de panel se consideran generalmente tres modelos: de datos agrupados (*pooled*), efectos fijos (*fixed effects*) y efectos aleatorios (*random effects*) (Carbajal, Carrillo y De Jesús, 2018). En modelos de respuesta cualitativa, la dependiente a explicar es una variable aleatoria que puede tomar un número finito de resultados (usualmente este número es pequeño). El caso principal surge cuando la dependiente es una respuesta binaria, que toma los valores de 0 y 1 si un cierto evento ocurre o no (Wooldridge, 2001). Por ejemplo, $Y = 1$ si una persona está casada e $Y = 0$ si no lo está, $Y = 1$ si un equipo jugó de local e $Y = 0$ si jugó de visitante. Haciendo un lado la definición de la dependiente, se suele referirse a $Y = 1$ como un éxito y a $Y = 0$ como fracaso.

El uso de modelos de respuesta binaria ayuda a superar las limitaciones que posee el modelo de probabilidad lineal (MPL) en lo que respecta a intentar expresar una probabilidad que se encuentre comprendida entre 0 y 1. En este tipo de modelos, el interés yace principalmente en la probabilidad de respuesta (Wooldridge, 2015). Por lo tanto, se debe de expresar la probabilidad de un evento condicionado a un conjunto de variables independientes:

$$E\,(Y_i / \boldsymbol{X}) = Pr\,(Y_i = 1 / \boldsymbol{X}) = Pr\,(Y_i = 1 / X_1, X_2, \ldots, X_k)$$

Donde **X** comprende el conjunto de independientes. Para contrarrestar las limitaciones antes mencionadas respecto al MPL, se debe expresar un modelo de respuesta binaria donde se incorpore una función que asuma de manera estricta valores entre cero y uno. Puede estar denotado de la siguiente forma:

$$Pr\,(Y_i = 1 / \boldsymbol{X}) = G\,(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k) = G\,(\beta_0 + \boldsymbol{X\beta})\ ;\ 0 < G(z) < 1$$

Donde **Xβ =** $\beta_1 X_1 + \ldots + \beta_k X_k$ y G es dicha función que asegura que las probabilidades del modelo de respuesta binaria se encuentren entre cero y uno. En el modelo de probabilidad lineal G es una función de identidad, por lo que las probabilidades de respuesta no pueden estar entre 0 y 1 para todo X y β. Diferentes funciones de carácter no lineal se han planteado para lograr que dichas probabilidades se mantengan dentro de ese rango, pero las dos que más resaltan o que se ven con mayor frecuencia en la práctica son las de los modelos Logit y Probit siendo G, en el primero, una función logística creciente que se ubica entre cero y uno para todo z (todos los números reales). Se expresade la siguiente forma:

$$G(z) = exp(z)/[1 + exp(z)] = \Delta(z)$$

En el caso del modelo Probit, se plantea G como una distribución acumulada normal estándar creciente que también cerciora que z se ubique entre cero y uno. Para este modelo, G se expresa de la siguiente forma:

$$G(z) = \Phi(z) \equiv \int_{-\infty}^{z} \Phi(v$$

En el análisis econométrico estas han sido las que más se han utilizado (por no decir las únicas). Es decir, las ventajas de utilizar estos modelos frente al MPL es que asegura lo antes mencionado y hace que los efectos parciales sean decrecientes, pero la interpretación de sus resultados es más difícil siendo esta su principal desventaja. Por lo tanto, para aplicar de manera satisfactoria los modelos Logit y Probit, es importante saber cómo interpretar los β_j en variables explicativas tanto continuas como discretas (Wooldridge, 2001). Bajo las modelaciones tanto Logit como Probit la función de regresión es no lineal en los β, por lo que realizar estimaciones mediante mínimo cuadrados ordinarios (MCO) queda descartado. Por lo tanto, según la revisión literaria econométrica, se debe estimar utilizando el método de máxima verosimilitud. Este se basa en la distribución de la variable dependiente dada el conjunto de independientes **X**. Al momento de su aplicación, se tiene que obtener lo que se denomina estimador de máximaverosimilitud (EMV) condicionado sobre las variables independientes. Para ello se necesita la densidad de Y_i dada X_i. [Vea Wooldridge (2015, capitulo 17) para un ahondamiento más detallado sobre este método].

Los modelos Logit y Probit también cuentan con pruebas de hipótesis, siendo más específicos, cuando se quiere probar restricciones múltiples de exclusión. Dichas pruebas se le conoce como: multiplicador de Lagrange, prueba de Wald y la de razones de verosimilitudes (RV). No se entrará en detalle para estas pruebas.

Modelo ligit Ordenado

El modelo logit ordenado es un modelo de regresión basado en las probabilidades acumuladas de una variable de respuesta ordinal. De forma particular, se asume que el logit de cada probabilidad acumulada es una función lineal de las covariables con coeficientes de regresión constantes en las respuestas categóricas. La respuesta a sobre qué tan satisfecha está una persona con su calidad de vida es un ejemplo de variable ordinal, la cual puede variar entre 1 a 10 siendo 1 “muy insatisfecho” y 10 “muy satisfecho”. Otro ejemplo es que *Y* sea una calificación crediticia en una escala de 0 a 6 siendo 6 la calificación más alta y 0 la más baja. Resulta tentador analizar los resultados ordinales con el modelo de regresión lineal, asumiendo distancias equivalentes entre categorías. Sin embargo, dicho enfoque posee varios inconvenientes. Por lo tanto, cuando la

variable de interés posee una naturaleza ordinal, es recomendable emplear una modelación especifica como el logit ordenado.

Sea Y_i una variable de respuesta ordinal con C categorías para el i-ésimo sujeto, junto con un vector de covariables X_i. Un modelo de regresión establece una relación entre lascovariables y el conjunto de probabilidades de las categorías:

$p_{ci} = P_r(Y_i = y_c|\boldsymbol{x_i}), c = 1, \ldots, C$De manera general, los modelos de regresión para respuestas ordinales no se encuentran expresados en términos de probabilidades de las categorías. Se refieren, másbien, a convenientes transformaciones individuales, como las probabilidades acumuladas:

$$g_{ci} = P_r(Y_i \preceq y_c|\boldsymbol{x_i}), c = 1, \ldots, C$$

Se debe tomar en cuenta que la última probabilidad acumulada es necesariamente igual a 1, por lo que el modelo especifica solo C-1 probabilidades acumuladas. Un modelo logit ordenado para una respuesta ordinal Y_i con C categorías se define mediante un conjunto de C-1 ecuaciones, donde las probabilidades acumuladas $g_{ci} = P_r(Y_i \preceq y_c|x_i)$ están relacionadas con un predictor lineal $Q' = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots$ a través de la función logit:

$$logit(g_{ci}) = \log\left(\frac{g_{ci}}{1 - g_{ci}}\right) = \alpha_c - Q'_{\boldsymbol{Xi}}, c = 1,2, \ldots, C - 1 \quad (1)$$

Los parámetros α_c, llamados umbrales o puntos de corte, están en orden creciente ($\alpha_1 < \alpha_2 < \cdots < \alpha_{C-1}$). No es posible estimar, de manera simultánea, el intercepto general β_0 y todos los puntos de corte *C*-1. De hecho, agregar una constante de naturaleza arbitraria al intercepto general β_0 puede ser contrarrestado agregando la misma constante a cada punto de corte α_c. Este problema de identificación es resuelto, usualmente, ya sea omitiendo la constante general del predictor lineal ($\beta_0 = 0$) o fijando el primer punto de corte ($\alpha_1 = 0$). El vector de pendientes **β** no se encuentra indexado por el índice categórico *C*, por lo que los efectos de las covariables son constantes en todas las categorías de respuesta. Esto último es denominado como el supuesto de regresión paralela. De hecho, plotear $logit(g_{ci})$ contra una covariable produce *C*-1 líneas paralelas (o curvas paralelas en caso de una especificación no lineal). En (1) el signo negativo antes de **β** implica que aumentar una covariable con pendiente positiva se encuentra asociado con undesplazamiento hacia el extremo derecho de la escala de respuesta. Dicho de otra forma, un aumento en las probabilidades de las categorías de respuesta superiores. Algunos autores escriben el modelo con un signo positivo antes de **β** (en tal caso se invierte la interpretación de los efectos de las covariables).

De la ecuación (1), la probabilidad acumulada para la categoría *C* es la siguiente:

$$g_{ci} = \frac{\exp(\alpha_c - Q'^{\boldsymbol{Xi}})}{1 + \exp(\alpha_c - Q'^{\boldsymbol{Xi}})} = \frac{1}{1 + \exp(-\alpha_c + Q'\boldsymbol{Xi})} \qquad (2)$$

El modelo logit ordenado también es conocido como el modelo de probabilidades proporcionales (Proportional Odds Model), debido a que el supuesto de regresión paralela implica la proporcionalidad de las probabilidades de no exceder la *C*-ésima categoría

$odds_{ci} = g_{ci}/(1 - g_{ci})$. De hecho, la proporción de estas probabilidades para dos unidades, digamos *i* y *j*, es $odds_{ci}/odds_{cj} = \exp[(\beta'(x_j - x_i)]$, que no depende de *C* y, por tanto, es constante en todas las categorías de respuesta. El logit ordenado es un miembro de la amplia clase de modelos ordinales acumulativos, donde el modelo logit es reemplazado por una función de enlace general. Las funciones de enlace más comunes son: Logit, Probit y Log-Log complementario (estos modelos se conocen en psicometríacomo Modelos de Respuesta Gradual).

Supuesto de regression paralela

Uno de los supuestos subyacentes a las regresiones logit y probit ordenadas es que la relación entre cada par de grupos de resultados es el mismo. Es decir, la regresión logit ordenada asume que los coeficientes que describen la relación entre, digamos, las categorías más bajas frente a todas las más altas de la variable de respuesta son los mismos que los que describen la relación entre la siguiente categoría más baja y todas las categorías superiores, etc. A esto se le conoce como supuesto de regresión paralela, deprobabilidades proporcionales o de líneas paralelas. En la práctica, violar este supuesto puede o no alterar las conclusiones del modelo y, de ser el caso, se necesita comprobar mediante pruebas.

Sea un logit ordenado, el supuesto indica que el valor del odds-ratio es el mismo paratodos los valores Y_j puesto que para cualquier categoría:

$$\frac{\Omega_j(x_k + 1)}{\Omega_j(x_k)} = e^{-\beta k}$$

Estos es una consecuencia de que en la especificación de las probabilidades los coeficientes β que afectan a las variables explicativas son los mismos para todas las categorías:

$$P(y_i = y_1|\boldsymbol{x}_i) = F(y_1 - \boldsymbol{x}_i'\beta) \qquad si\ y = y_1$$

$$P(y_i = y_2|\boldsymbol{x}_i) = F(y_2 - \boldsymbol{x}_i'\beta) - F(y_1 - \boldsymbol{x}_i'\beta) \qquad si\ y = y_2$$

$$P(y_i = y_3|\boldsymbol{x}_i) = F(y_3 - \boldsymbol{x}_i'\beta) - F(y_2 - \boldsymbol{x}_i'\beta) \qquad si\ y = y_3$$

$$P(y_i = y_j|\boldsymbol{x}_i) = 1 - F(y_{j-1} - \boldsymbol{x}_i'\beta) \qquad si\ y = y_j$$

El anterior sistema muestra que las probabilidades de respuesta varían solo como consecuencia de que los puntos de corte Y_j son distintos entre categorías. Pero, entre las distintas ecuaciones, los coeficientes son los mismos. Una consecuencia de esto es que los efectos de las exógenas sobre las probabilidades acumuladas son los mismos en todas las categorías:

$$P(y_i \preceq y_1) = F(y_1 - \boldsymbol{x}_i\,'\beta)$$

$$P(y_i \preceq y_2) = F(y_2 - \boldsymbol{x}_i\,'\beta)$$

$$P(y_i \preceq y_j) = F(y_j - \boldsymbol{x}_i\,'\beta)$$

Las pruebas de este supuesto comparan el logit ordenado con un logit ordenado generalizado completo siendo la hipótesis nula (*Ho*) que el supuesto no se viola. El software Stata, por ejemplo, cuenta con cinco pruebas implementadas en el comando "oparallel" para comprobar el supuesto:

- Prueba de Razón de Verosimilitud
- Prueba de Wald
- Prueba de Puntuación (Score Test)
- Prueba de Wolfe-Gould (o de Razón de Verosimilitud Aproximada)
- Prueba de Brant (o Prueba Aproximada de Wald)

Figura 10 - Supuesto de Regresión Paralela

```
. oparallel

Tests of the parallel regression assumption

                    Chi2     df   P>Chi2
-------------------------------------------
     Wolfe Gould   21.33      3   0.000
           Brant   17.45      3   0.001
           score   26.66      3   0.000
likelihood ratio   22.05      3   0.000
            Wald   28.88      3   0.000
```

Fuente: CBN-ITI TRAINING (2016)

Según el cuadro anterior, se incumple el supuesto de regresión paralela. Para tal caso,se pueden plantear las siguientes alternativas:

- Seguir empleando la regresión logit ordenada dado que las implicancias, en la práctica, de violar este supuesto son mínimas.
- Utilizar un modelo logit multinomial. Esto libera a uno del supuesto de proporcionalidad, pero es menos parsimonioso y, a menudo, dudoso por motivos de fondo.
- Dicotomizar el resultado y emplear una regresión logística binaria. Sin embargo, se perdería información y podría alterar las conclusiones principales.
- Emplear un modelo que no asuma la proporcionalidad. Aquí es donde entra el logit ordenado generalizado siendo, esta alternativa, la que se utiliza con mayor frecuencia.

El supuesto de regresión paralela, de modelos acumulativos, puede ser demasiado restrictivo. Tal suposición puede ser apalancada permitiendo que los umbrales dependan de covariables o, alternativamente, que estas tengan pendientes específicas de categoría. A estos modelos se les denomina Proporciones de Probabilidad Parcial (Peterson y Harrel, 1990). Los modelos que violan el supuesto de regresión paralela deben usarse con cuidado ya que aumentan los problemas de identificación e interpretación.

Modelo Ligit ordenado generalizado

Desafortunadamente, la experiencia sugiere que los supuestos del modelo logit ordenado son con frecuencia violados (Long y Freese, 2014). Por lo general, a los investigadores se le ha dejado dos opciones: quedarse con un método el cual se sabe que sus suposiciones son violadas o cambiarse a uno mucho menos parsimonioso y más difícil de interpretar (por ejemplo, un logit multinomial que no emplea la información sobre el orden de categorías). Sin embargo, existe una tercera opción: el modelo logit ordenado

generalizado. Esta modelo relaja de forma selectiva los supuestos del logit ordenado solo cuando es necesario produciendo, potencialmente, resultados que no poseen los problemas de un logit ordenado siendo, a su vez, de fácil interpretación.

Para una variable de respuesta ordinal con *M* categorías, el modelo logit ordenado generalizado es el siguiente:

$$P(Y_i > j) = \frac{\exp(\alpha_j + X_i\beta_j)}{1 + [\exp(\alpha + X\ \beta\ )]j} \quad , \ j = 1,2, \dots , M - 1 \qquad i\ \ j$$

Por ejemplo, si la variable de respuesta tiene cuatro posibles valores el modelo tendrá tres conjuntos de coeficientes (estimando tres ecuaciones de manera simultánea). Un logit ordenado generalizado brinda resultados similares a las regresiones logit binarias/acumulativas. El logit ordenado es caso espacial del logit ordenado generalizado donde los β son iguales para cada *j* (es decir, los subíndices j son innecesarios en la fórmula anterior). Entre estos dos extremos se encuentra el modelo logit ordenado. Con este último, algunos de los coeficientes β son los mismos para todos los valores de *j*, mientras que para otros puede diferir. En el siguiente logit ordenado, por ejemplo, los β para X_1 y X_2 se encuentran limitadas a ser iguales en todos los valores de *j*, pero los β paraX_3 no:

$$P(Y > j) = \frac{\exp(\alpha_j + X1_i\beta1 + X2_i\beta2 + X3_i\beta3_j)}{} \quad , \ j = 1,2, \dots , M - 1$$

$$_{i} \qquad 1 + [\exp(\alpha + X1\ \beta 1 + X2\ \beta 2 + X3\ \beta 3\ )]$$

$$j \qquad i \qquad i \qquad i \quad j$$

Tanto un logit ordenado generalizado sin restricciones como un logit multinomial generarán muchos más parámetros que un logit ordenado puesto que, con estos métodos, todas las variables son libres de la restricción de probabilidades proporcionales (aunque el supuesto puede ser violado por uno o alguno de ellos). Sin embargo, con un modelo deproporciones de probabilidad parcial es posible relajar el supuesto de regresión paralela para aquellas variables en las que no se viola el supuesto.

Logit ordenado generalizado: interpretación

De manera empírica, el logit ordenado generalizado puede funcionar de forma apropiada proporcionando un ajuste a los datos sustancialmente mejor que el logit ordenado y, al mismo tiempo, es más parsimonioso que otras alternativas a emplear. Sin embargo, la interpretación y justificación de este modelo es menos sencilla que el logit ordenado. Desafortunadamente, muchos investigadores llegan a notar solamente el ajuste superior del logit ordenado generalizado y comentan poco sobre lo que podrían significarlos resultados.

Entonces ¿Cómo pueden ser interpretados y justificados los resultados del logit ordenado generalizado? Según Williams (2016), existen al menos cuatro enfoques posibles:

- El modelo se encuentra mal especificado:

Una especificación incorrecta puede causar que el supuesto de regresión paralela aparentemente se viole cuando otro modelo (mejor especificado) podría no mostrar tal violación. Por ejemplo, Williams (2010) brinda un caso en donde la simple adición de un término cuadrático al modelo es, teóricamente, más razonable y conduce a pruebas que muestran que los supuestos de dicho modelo se cumplen. Siendo de mucha utilidad el logit ordenado generalizado, los investigadores deben examinar si un logit ordenado modificado es simple,

válido y más fácil de entender. A su vez, deben considerar si variables importantes se han omitido y, de ser el caso, ver si la inclusión de dichas variables genere que el supuesto de regresión paralela deje de ser violado.

- LOG como modelo de probabilidad no lineal:

Según Long y Freese (2006, p. 187): " El logit ordenado también se puede desarrollar como un modelo de probabilidad no lineal sin apelar a la idea de una variable latente". A modo de extensión, lo más simple puede ser interpretar también el logit ordenado generalizado como un modelo de probabilidad no lineal que permita los determinantes y la probabilidad de que ocurra cada resultado. No es necesario confiar en la idea de un subyacente *Y** que cuenta los valores observado de *Y*.

Para emplear dicho enfoque, Long y Freese (2014) indican que los efectos de las variables pueden ser evaluados a través de predicciones ajustadas, efectos marginales y examinando casos prototipo. Si bien este enfoque es simple y conveniente, podría causar errores como la omisión de percepciones importantes que las otras alternativasa emplear pueden ofrecer.

- Efecto de X/Y es no asimétrico y desigual en cada uno de los logit acumulativos:

El modelo logit ordenado asume que, para cada logit acumulado que se pueda estimar, el efecto de *X* sobre *Y* es el mismo. Sin embargo, tal suposición es a menudo irrazonable y demasiado restrictivo. Fullerton y Dixon (2010) se refieren a lo antes mencionado como "efectos asimétricos". Ellos argumentan (p. 649) que el logit ordenado generalizado posee ventajas sobre otros métodos cuando tales asimetrías aparecen: "Los modelos OLS tradicionales no permiten la posibilidad de que la edad, periodo y cohorte puedan afectar el apoyo a la educación. Similarmente, otros modelos (como el logit binario) no permiten dicha posibilidad e incluso pueden confundirla dado que el colapso de categorías (de una variable de respuesta ordinal) resulta en la perdida de información".

Las relaciones asimétricas a menudo poseen un adecuado sentido teórico y, también, revelan percepciones sustantivas sobre las relaciones subyacentes entre variables. Sin embargo, la mayoría de estudios simplemente reportan los coeficientes del logit ordenado generalizado sin explicar por qué existen relaciones asimétricas y su significado. La regresión OLS con una variable

ordinal colapsada y el propio logit ordenado pueden oscurecer completamente las relaciones asimétricas.

- Algunas explicativas afectan la dirección de las respuestas mientras otros afectan su intensidad:

Supongamos, por ejemplo, que las mujeres tienden a asumir posiciones políticas menos extremas que los hombres. La siguiente tabla puede mostrar tal relación:

Figura 11 - Logit Generalizado: Ejemplo de interpretación

Table 1-3. Hypothetical example of proportional odds violated-II*.

	Attitude				
Gender	SD	D	A	SA	Total
Male	250	250	250	250	1,000
Female	100	400	400	100	1,000
Total	350	650	650	350	2,000
		1 versus 2, 3, 4	1 & 2 versus 3 & 4		1, 2, 3 versus 4
OddsM		750/250 = 3	500/500 = 1		250/750 = 1/3
OddsF		900/100 = 9	500/500 = 1		100/900 = 1/9
OR (OddsF/OddsM)		9/3 = 3	1/1 = 1		(1/9)/(1/3) = 1/3
Betas		1.098612	0		−1.098612
Ologit Beta (OR)		0 (1.00)			
Ologit χ^2 (1 d.f.)		0.00 (p = 1.0000)			
Gologit χ^2 (3 d.f.)		202.69 (p = 0.0000)			
Brant Test (2 d.f.)		179.71 (p = 0.000)			

Fuente: Williams (2016)

Usando un enfoque direccional, el modelo ordinal podría no funcionar adecuadamente, mientras que empleando un enfoque de intensidad si llegaría a hacerlo. Pero suponga que, para cualquier otra variable, el enfoque direccional funciona bien en un modelo ordinal. Como muestra la tabla, la suposición del logit ordenado se violará debido a la única variable problemática. Anteriormente solo se hubieran tenido dos opciones: ejecutar el logit ordenado (siendo el modelo no apropiado para la variable de genero) o ejecutar un logit multinomial (ignorando la parsimonia del modelo). Empleando un logit ordenado generalizado se tiene una tercera opción: restringir las variables donde se cumpla el supuesto de regresión paralela, mientras se liberan otras variables.

Por lo tanto, el modelo que se utilizará para fines de esta investigación es el Logit Ordenado Generalizado ya que, a pesar de que la interpretación de sus resultados es tediosa, es el planteamiento econométrico que más se asemeja a la problemática de este trabajo. Entonces, el modelo econométrico que nos ayudará a determinar los factores que influyen de manera significativa en la probabilidad de ganar un partido de fútbol para el caso de los equipos de la Liga 1 peruana en el horizonte de tiempo 2015 – 2019, se encuentra denotado de la siguiente manera:

$$Outcome_{it} = \beta_0 + \beta_1 Antiquity_{it} + \beta_2 Yellow\_card_{it} + \beta_3 Red\_card_{it} + \beta_4 Goals_{it} + \beta_5 Age_{it} + \beta_6 Local_{it} + \beta_7 Foreign_{it} + \beta_8 Altitude_{it} + \beta_9 Height_{it} + \beta_{10} Weight_{it} + \beta_{11}$$

$$\text{Possesion}_{it} + \beta_{12} \text{Substitution}_{it} +$$

$$u_{it}$$ 3.4.6. Descripción de variables

Tabla 3 - Estadística Descriptiva de Variables

Variable	Obs	Mean	Std. Dev.	Min	Max
Outcome	3,318	1	.8477746	0	2
Antiquity	3,318	52.32369	38.53954	4	118
Yellw_crd	3,318	2.509042	1.476295	0	8
Rd_crd	3,318	.2097649	.4606963	0	4
Goals	3,318	1.335744	1.189338	0	8
Age	3,318	24.77409	1.39174	21.06	27.96

Foreign	*3,318*	*3.535262*	*1.229152*	*0*	*8*
Height	*3,318*	*177.7586*	*1.354058*	*175*	*182*
Weight	*3,318*	*73.66365*	*1.492483*	*71*	*78*
Posses	*3,318*	*49.99397*	*5.977898*	*15*	*85*
Subst	*3,318*	*2.814346*	*.4841899*	*0*	*3*
Locall1v0	*3,318*	*.5*	*.5000754*	*0*	*1*

Tabla 4 - Correlación de Cualidad de Jugadores

	Age	***Height***	***Weight***
Age	***1.0000***		
Height	*-0.1322*	***1.0000***	
Weight	*0.1697*	*0.6483*	***1.0000***

Gráfico 2 - Edad, Peso y Altura Promedio

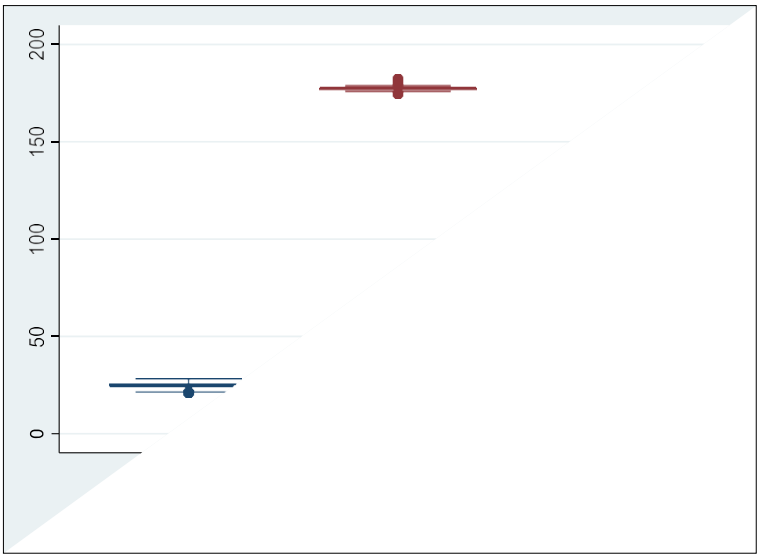


Gráfico 1 - Amarillas, Rojas, Goles, Substituciones y Posesión

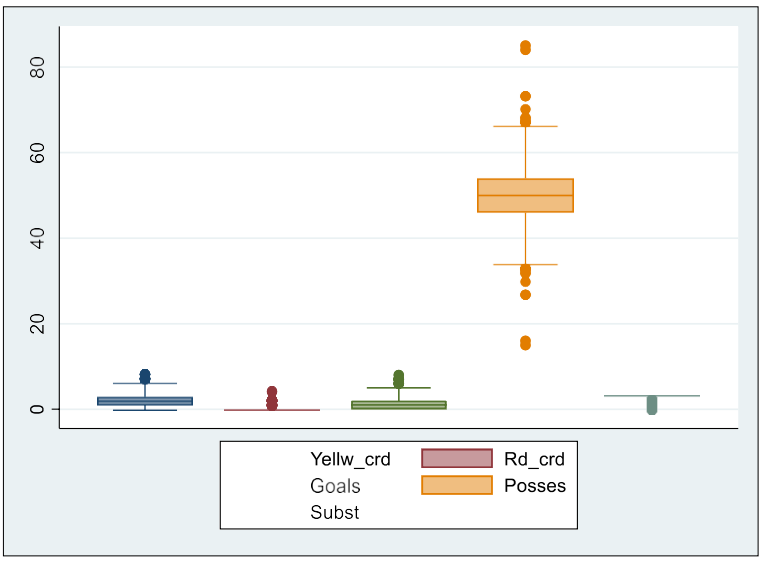


Gráfico 3 - Altitud del estadio

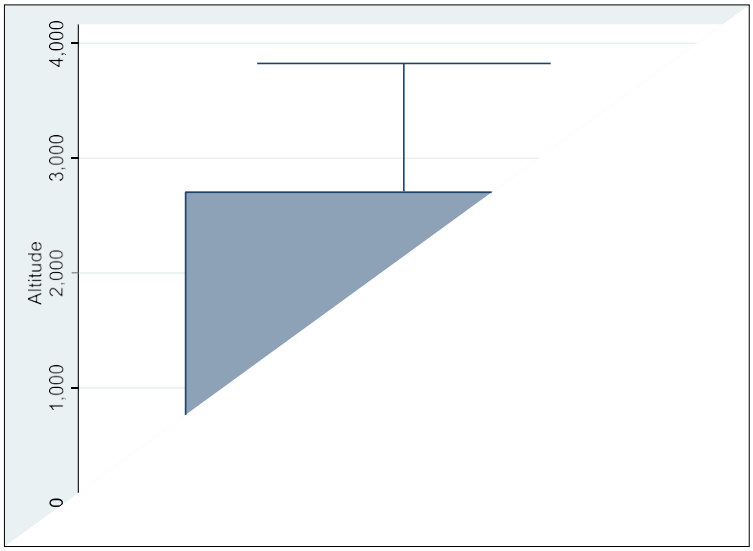
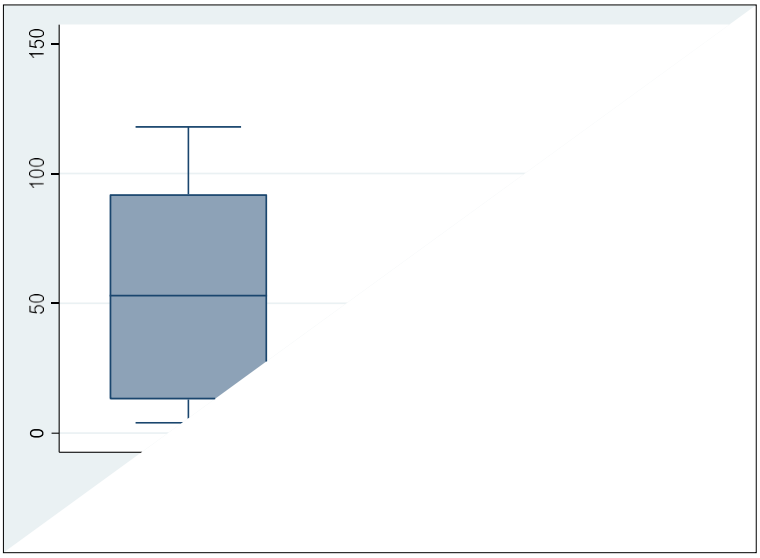


Gráfico 4 - Antigüedad, Local y Extranjeros



RESULTADOS

A partir del modelo escogido y tomando como referencia una distribución Chi-2, se pudieron determinar diferentes resultados que serán presentados a continuación:

Tabla 5 - Resultados Regresión Logit Ordenado

Outcome	***Coef.***	***Rbst Std. Err.***	***z***	***P > |z|***
Rd_crd	*-.3203332*	*.0818479*	*-3.91*	*0.000**
Goals	*1.401829*	*.0437862*	*32.02*	*0.000**
Age	*.0640336*	*.0302156*	*2.12*	*0.034***
Height	*.1351031*	*.0408761*	*3.31*	*0.001**
Weight	*-.067778*	*.0356534*	*-1.90*	*0.057****
Posses	*.0412988*	*.0069585*	*5.94*	*0.000**
Locall1v0	*1.00173*	*.0791036*	*12.66*	*0.000**
Antiquity	*.0014347*	*.0011011*	*1.30*	*0.193*
Yellw_crd	*.054989*	*.0259574*	*2.12*	*0.034***
Altitude	*-1.26e-06*	*.0000284*	*-0.04*	*0.965*
Subst	*.0638736*	*.0769259*	*0.83*	*0.406*
Foreign	*.0510133*	*.0313311*	*1.63*	*0.103*

*Nota: *,**,*** denota niveles de significancia de 1, 5 y 10%,*

Se obtuvo que la variable *Antiquity* no posee significancia en el modelo para distintos niveles de confianza (0.193), indicando que no tiene influencia en la probabilidad de ganar un partido de fútbol. *Yellw_crd* no arroja los datos esperados (a pesar de ser significativo). Si se tomará en cuenta dicha variable se estaría afirmando que, para aumentar las probabilidades de ganar, los equipos tendrían que cometer más faltas, lo cual no tendría mucho sentido. Este resultado es similar al de Celis (2013), en donde la variable "*tarj_amar"* arroja, de igual forma, datos no relevantes dentro del modelo. *Altitude* tampoco resulta significativo (0.965), lo cual refuerza los argumentos de Chumacero (2009) en donde, a través de una modelación Probit ordenada, se obtiene que la altitud del estadio no es un factor determinante en el resultado de un partido. Sin embargo, otros factores como la temperatura y humedad si mostraron ser relevantes. Se recomienda incluir dichas variables climatológicas en futuras investigaciones. A su vez, las variables *Subst* y *Foreign* no poseen significancia en el modelo para distintos niveles de confianza (0.406 y 0.103, respectivamente). Nuevamente, el resultado no significativo de *Foreign* es similar al de Celis (2013) en donde la variable "*num_extranj*" no posee significancia incluso al 10% (0.314).

Por lo tanto, se continuará el análisis volviendo a realizar la regresión logit ordenada con las variables que resultaron significativas:

Tabla 6 - Resultados Regresión Logit Ordenado #2

Variable	***Coef.***	***Rbst Std. Err.***	***z***	***P > |z|***
Rd_crd	*-.2690099*	*.0792909*	*-3.39*	*0.001**
Goals	*1.409789*	*.043681*	*32.27*	*0.000**
Age	*.0552235*	*.0287326*	*1.92*	*0.055****
Height	*.1531683*	*.0386506*	*3.96*	*0.000**
Weight	*-*	*.0353653*	*-1.94*	*0.052****

	.0687609			
Posses	*.0415805*	*.0068873*	*6.04*	*0.000**
Locall1v0	*.9748933*	*.078042*	*12.49*	*0.000**

*Nota: *, **, *** denota niveles de significancia de 1, 5 y 10%,*

La variable *Rd_crd* muestra ser significativa para diferentes niveles mostrando que, el aumento en una tarjeta roja, disminuye en 0.269 la probabilidad de obtener un resultado victorioso. Para *Weight* la interpretación es análoga (con la diferencia de que es significativa solo al 10%). *Goals* arroja que el incremento de una unidad en la cantidad de goles aumenta en 1.409 la probabilidad de ganar un partido. Respecto al porcentaje de posesión de balón y la altura promedio de los jugadores, los datos de *Posses* y *Height* indican que el incremento en una unidad de estas variables aumenta la probabilidad de ganar en 0.042 y 0.153, respectivamente. La dummy que diferencia a los equipos locales y visitantes resulta significativa, indicando que la probabilidad de obtener una victoria es 0.975 veces mayor para los equipos que juegan en casa que los que juegan como visitantes. Este resultado se refuerza en el hecho de que diferentes autores, que toman en cuenta dicha variable dicotómica en su modelación econométrica, obtienen datos similares al del presente trabajo. Finalmente, la variable que denota la edad promedio de los jugadores indica que su incremento en una unidad aumenta la probabilidad de ganar en 0.055. Cabe resaltar, que este resultado se interpreta desde el punto de vista de la experiencia del jugador. Es decir, que el tener jugadores de mayor edad (y por ende más experimentados) incrementa las chances de obtener la victoria en un partido de fútbol. Sin embargo, es de conocimiento del autor que el tener jugadores más jóvenes en el partido incrementa también la probabilidad de ganar.

A continuación, se procederá a comprobar el supuesto de regresión paralela del modelo. Como se mencionó anteriormente, la regresión logit ordenada asume que los coeficientes que describen la relación entre la categoría más baja frente a la más alta de la variable categórica son los mismos que los que describen la relación entre la siguiente categoría más baja frente a la más alta y así sucesivamente. Para

el caso del presente trabajo de investigación, se comprobará si los coeficientes que describen la relación entre la categoría 0 (*perder*) frente a la 1 (*empatar*) son los mismos que los que describen la relación entre la categoría 1 frente a la 2 (*ganar*). Para comprobar el supuesto se emplearán las siguientes pruebas: Razón de Verosimilitud, Wald, de Puntuación (Score Test), Wolfe-Gould (o de Razón de Verosimilitud Aproximada) y Brant (o Prueba Aproximada de Wald). Utilizando el comando "*oparallel*" del software Stata, se obtienelo siguiente:

Tabla 7 - Supuesto de Regresión Paralela

Prueba	***Chi2***	***df***	***P > Chi2***
Wolfe Gould	*69.59*	*7*	*0.000*
Brant	*69.43*	*7*	*0.000*
Score	*75.71*	*7*	*0.000*
LikelhdRatio	*69.48*	*7*	*0.000*
Wald	*89.92*	*7*	*0.000*

Siendo la hipótesis nula (*Ho*) que el supuesto no se viola, la significancia de todas las pruebas indica que el supuesto de regresión paralela si se viola. Esto refuerza el argumento de Long y Fresee (2014) de que los supuestos del modelo logit ordenado son con frecuencia violados. Por tanto, se empleará un logit ordenado generalizado debido a que dicha modelación relaja de forma selectiva los supuestos del logit ordenado produciendo resultados que no poseen los problemas de un logit ordenado y siendo, a suvez, más fácil de interpretar.

Por lo tanto, los resultados de realizar una regresión logística ordenada generalizada con las variables significativas antes mencionadas son los siguientes:

Tabla 8 - Resultados Regresión Logit Ordenado Generalizado

Outcome	Coef.	Rbst Std. Err.	z	P > |z|
0				
Rd_crd	-.2589464	.0803748	-3.22	0.001
Goals	1.166826	.0508253	22.96	0.000
Age	.0538922	.0286948	1.88	0.060
Height	.153426	.03818	4.02	0.000
Weight	-.0673517	.0349738	-1.93	0.054
Posses	.0417166	.0067513	6.18	0.000
Locall1v0	.9833268	.0782329	12.57	0.000
_cons	-26.75831	5.573385	-4.80	0.000
1				
Rd_crd	-.2589464	.0803748	-3.22	0.001
Goals	1.655772	.0569447	29.08	0.000

Age	*.0538922*	*.0286948*	*1.88*	*0.060*
Height	*.153426*	*.03818*	*4.02*	*0.000*
Weight	*-.0673517*	*.0349738*	*-1.93*	*0.054*
Posses	*.0417166*	*.0067513*	*6.18*	*0.000*
Locall1v0	*.9833268*	*.0782329*	*12.57*	*0.000*
_cons	*-29.23229*	*5.578358*	*-5.24*	*0.000*

Nota: Se emplea el comando "autofit" para simplificar la identificación de modelos de proporción de probabilidad parcial que se ajusten mejor a los datos

Se puede interpretar los resultados del logit ordenado generalizado como coeficientes de un modelo logit binario, donde las categorías de la variable dependiente colapsan a solo dos categorías. Por tanto, el tener coeficientes positivos significa que un incremento en el valor de las covariables aumenta, también, las probabilidades de estar en el nivel más alto de la dependiente categórica y viceversa. Tomando como base la categoría 2 (*ganar*), en el primer panel de coeficientes se puede observar que un incremento en la cantidad de tarjetas rojas y el peso promedio de los jugadores disminuye la probabilidad de pasar de la categoría 0 (*perder*) a la 2. A su vez, un incremento en la edad y altura promedio de los jugadores aumenta la probabilidad de pasar de perder a ganar. Para la variable que denota el porcentaje de posesión de balón la interpretación es análoga y, respecto a la dummy que diferencia a los equipos locales y visitantes, los datos muestran que el incremento de esta variable (de 0 a 1) aumenta la probabilidad de pasar de la categoría 0 a la 2. El coeficiente de *Goals* es positivo e incrementa ligeramente a través de los puntos de corte, lo cual significa que hay menores efectos en pasar de perder a ganar y mayores efectos en pasar de la

categoría 1 (empatar) a la 2. La interpretación del segundo panel de coeficientes es análoga al del primero, con la diferencia de que ahora se analiza tanto los incrementos como reducciones en la probabilidad de pasar de la categoría 1 a la 2 (tomando como base, de igual forma, la categoría 2).Entonces, se observa que la interpretación es más directa para aquellas variables que cumplen el supuesto de regresión paralela (similar a la modelación logit ordenada) mientras que para las variables que incumplen el supuesto la interpretación cambia levemente. Por otro lado, los resultados de los efectos marginales promedio, para las distintas categorías de la variable de respuesta, se muestran a continuación:

Tabla 9 - Efectos Marginales Promedio

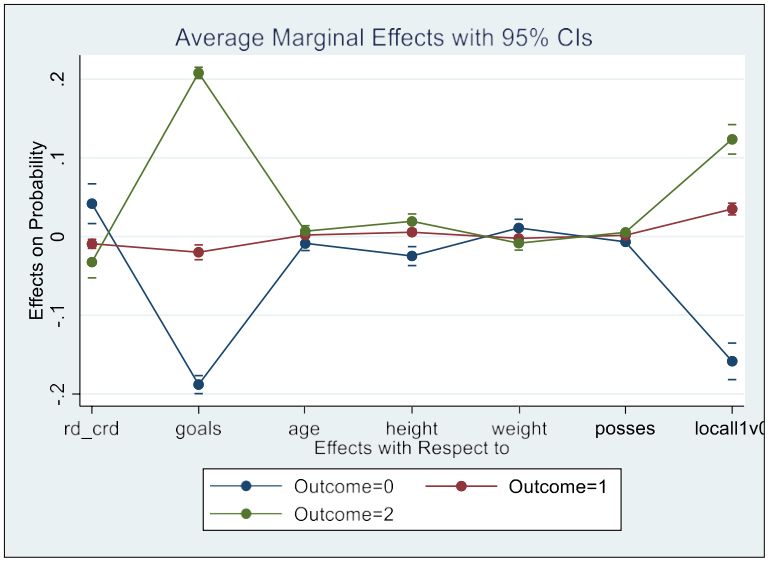
	dy/dx	***Std. Err.***	***z***	***P > |z|***
Rd_crd				
_predict				
1	*.0417339*	*.012901*	*3.23*	*0.001**
2	*-.0091969*	*.0029364*	*-3.13*	*0.002**
3	*-.0325369*	*.0100799*	*-3.23*	*0.001**
Goals				
_predict				
1	*-.1880549*	*.0058407*	*-32.20*	*0.000**
2	*-.0199948*	*.0047964*	*-4.17*	*0.000**
3	*.2080497*	*.0036311*	*57.30*	*0.000**
Age				

_predict				
1	*-.0086857*	*.0046207*	*-1.88*	*0.060****
2	*.0019141*	*.0010327*	*1.85*	*0.064****
3	*.0067716*	*.0036021*	*1.88*	*0.060****
Height				
_predict				
1	*-.0247274*	*.0061188*	*-4.04*	*0.000**
2	*.0054492*	*.00142*	*3.84*	*0.000**
3	*.0192782*	*.0047833*	*4.03*	*0.000**
Weight				
_predict				
1	*.0108549*	*.0056298*	*1.93*	*0.054****
2	*-.0023921*	*.0012562*	*-1.90*	*0.057****
3	*-.0084628*	*.0043916*	*-1.93*	*0.054****
Posses				
_predict				
1	*-.0067234*	*.0010727*	*-6.27*	*0.000**

2	*.0014816*	*.0002649*	*5.59*	*0.000**
3	*.0052417*	*.0008425*	*6.22*	*0.000**
Locall1v0				
_predict				
1	*-.1584807*	*.0118385*	*-13.39*	*0.000**
2	*.0349245*	*.0038283*	*9.12*	*0.000**
3	*.1235562*	*.0095557*	*12.93*	*0.000**

Se puede observar que, en promedio, el incremento en una unidad en la cantidad de tarjetas rojas aumenta en 0.042 la probabilidad de *perder* y disminuye la probabilidad de *empatar* y *ganar* en 0.009 y 0.033 respectivamente. El mismo análisis se aplica para la variable asociada al peso promedio de los jugadores, donde su incremento en una unidad aumenta en 0.011 la probabilidad de *perder* y disminuye la probabilidad de *empatar* y *ganar* en 0.002 y 0.008. A su vez, un incremento, en promedio, en la cantidad de goles anotados disminuye la probabilidad de *perder* y *empatar* en 0.188 y 0.019 y, también, aumenta en 0.208 la probabilidad de *ganar*. Respecto a la variable que denota la edad promedio de los jugadores, los datos muestran que, en promedio, un incremento en una unidad de dicha variable disminuye en 0.009 la probabilidad de *perder* y aumenta la probabilidad de *empatar* y *ganar* en 0.002 y 0.007 respectivamente. La interpretación antes mencionada es similar, también, para el caso de *Height* y *Posses*. Para la variable dicotómica podemos observar que, en promedio, su incremento en una unidad (es decir, de 0 a 1) disminuye en 0.158 la probabilidad de *perder* y aumenta la probabilidad de *empatar* y *ganar* en 0.035 y 0.124 respectivamente.

Gráfico 5 - Efectos Marginales Promedio



DISCUSION: CONCLUSIONES Y RECOMENDACIONES

El presente trabajo de investigación emplea la metodología econométrica para determinar los factores que influyen en la probabilidad de ganar un partido de fútbol parael caso de la Liga 1 peruana en el periodo 2015-2019. La evidencia muestra:

1. Que los años de antigüedad que posee un equipo no influye de manera significativa en la probabilidadde obtener un resultado victorioso.

2. A pesar de que la variable asociada a las tarjetas amarillas resulte significativa, esta no arroja los resultados esperados. Como se mencionó anteriormente, si se incorporará dicha variable se estaría afirmando que, para incrementar la probabilidad de ganar, los equipos deben de cometer más faltas contra los jugadores del equipo rival. Esto último no guarda mucha lógica y, además, se estaría incentivando comportamientos antideportivos.

3. Se demuestra, también, que tanto las sustituciones como la cantidad de jugadores extranjeros en el partido no influye de manera significativa en la probabilidad de ganar. Se busco evidencia literaria acerca de la no influencia de la cantidad de sustituciones, pero solamente se encontraron artículos asociados a los factores que determinan los patrones de sustitución. Cabe resaltar, que la no significancia de los jugadores extranjeros es mínima, por lo que si se podría argumentar que el tener jugadores de otros países influye en la probabilidad de salir victorioso (en términos de calidad de jugadores). Se podría analizar, en futuras investigaciones, el efecto de tener jugadores de una nacionalidad específica en la probabilidad de ganar un partido. Contrario a la creencia popular, la altitud del estadio no resulta un factor determinante.

4. Sin embargo, a pesar de que dicha variable resultara no significativa, otros trabajos muestran que la temperatura y humedad si muestran ser relevantes. Se podría incluir dichas variables en futuras investigaciones.

5. A su vez, el presente trabajo permitió corroborar el efecto negativo que generan las tarjetas rojas en la probabilidad de empatar y ganar un partido.

6. Por lo tanto, se recomienda a los equipos no cometer, en la medida de lo posible, faltas contra los jugadores del equipo contrincante. No obstante, el juicio del

árbitro influye bastante en ello (se pueden generar situaciones donde se comete una clara falta y el árbitro decide no colocar una tarjeta rojao viceversa).

7. Asimismo, se muestra que el tener jugadores con mayor peso corporal disminuye la probabilidad de empatar y ganar. Los planteles técnicos deben de mantener a sus jugadores en la mejor condición posible (ejercicios de condicionamiento, de resistencia, dieta balanceada y saludable, etc.) a fin de que puedan tener un mejor desempeño durante el partido.

8. Se evidencia, también, que el tener un mayor porcentaje de posesión durante el partido y tener jugadores más altos incrementa la probabilidad de obtener un resultado de empate o victoria. Los equipos deben de elaborar, en conjunto con los planteles técnicos, tácticas y estrategias de pase con el objetivo de mantener un mayor porcentaje de posesión del balón y así aumentar sus chances de anotar más goles y, además, contratar jugadores que posean una estatura ligeramente mayor al promedio.

9. Respecto a la variable asociada a la edad, los datos muestran que tener jugadores, con una edad levemente mayor al promedio, incrementa la probabilidad de empatar y ganar. Nuevamente, lo antes mencionado posee sentido desde el punto de vista de la experienciadel jugador.

10. Finalmente, se ratifica los efectos positivos que genera el anotar una mayor cantidad de goles y jugar en la modalidad de local.

REFERENCIAS BIBLIOGRÁFICAS

Paredes, O. (2002). El deporte como juego: un análisis cultural. San Vicente de Raspeig, Universidad de Alicante.

Mesa, R. y Arboleda, R. (2007). Aproximaciones teóricas al estudio de la relación economía y deporte. Distrito Federal, Universidad Autónoma Metropolitana: Unidad Azcapotzalco.

Coremberg, A., Sanguinetti, J. y Wierny, M. (2016). El fútbol en la economía argentina: números sin pasiones. Observatorio de economía del fútbol, Universidad deBuenos Aires.

Panfichi, A., Vila, G., Chávez, N. y Saravia, S. (2018). El otro partido: La disputa por el gobierno del fútbol peruano. Fondo Editorial, Pontificia Universidad Católica del Perú.

Contreras, J. y Muñoz, C. (2015). Estrategias en la Premier League: Evidencia empírica. Econometría y decisiones tácticas en el fútbol inglés. Chía, Universidad de La Sabana.

Celis, S. (2013). Transferencia de conocimiento como spillover en el rendimiento de un equipo deportivo: Evidencia para el fútbol nacional colombiano. Chía, Universidad de La Sabana.

Chumacero, R. (2009). Altitude or Hot Air. Journal of Sports Economics, Central Bank of Chile and University of Chile. Fiallo, N. (2017). Determinantes del desempeño deportivo y de los ingresos de los equipos profesionales de fútbol de Colombia – Categoría A – 2011 a 2012. Universidadde Santo Tomás, Bogotá.

Borland, J. (2006). Production functions for Sporting teams. Department of Economics, University of Melbourne, Australia.

Pindyck, R. y Rubinfeld, D. (2009). Microeconomía – Seventh Edición. Pearson Education S.A., Madrid.

Leeds, M. y Von Allen, P. (2016). The Economics of Sports – Fifth Edition. Editorial Routledge, Nueva York.

Rengifo, E. (2019). Teoría de la empresa. Facultad de Ciencias Económicas y de Negocios. Universidad Nacional de la Amazonía Peruana.

Arzubi, A. (2003). Análisis de Eficiencia sobre Exploraciones Lecheras de la Argentina. Departamento de Economía, Sociología y Políticas Agrarias. Universidad de Cordoba.

Mendieta, J. (2005). Apuntes de Microeconomía II: Teoría del Consumidor, Teoría del Productor, Teoría de Juegas y Competencia Imperfecta. Facultad de Economía. Universidad de Los Andes.

Bairam, E., J. Howells and G. Turner (1990), Production functions in cricket: Australian and New Zealand cricket, Applied Economics, 22, 871-79.

Carmichael, F. and D. Thomas (1995), Production and efficiency in team sports: An investigation of rugby league football, Applied Economics, 27, 859-69.

Scully, G. (1974), Pay and performance in major league baseball', American Economic Review, 64, 915-30.

Roggiero, L. (2012). El negocio no es redondo: Los determinantes del desempeño deportivo y financiero de los equipos de fútbol profesional del Ecuador. Facultad Latinoamericana de Ciencia Sociales. Sede Ecuador.

Pedrosa, R. y Salvador, J. (2003). El impacto del deporte en la economía: problemasde medición. Universidad de Valladolid.

De La Rosa, C. (2016). Introducción a modelos de datos de panel. Facultad de Ciencias Económicas y Empresariales. Universidad de Valladolid.

Labra, R. y Torrecillas, C. (2014). Guía Cero para datos de panel: Un enfoque práctico. Universidad Autónoma de Madrid.

Arellano, M y Bover, O. (1990). La econometría de datos de panel. Investigaciones Económicas (Segunda Época). Volumen XIV, Pág. 3 – 45.

Carrasco, R. (2001). Modelos de elección discreta para datos de panel y modelos deduración: una revisión de la literatura. Universidad Carlos III de Madrid.

Burdisso, T. (1997). Estimación de una función de costos para los bancos privados argentinos utilizando datos de panel. Banco Central de la República Argentina.

Strauss, A (1987). Qualitative analysis for social scientifics. New York: CambridgeUniversity de Press.

Carbajal, Y., Carrillo, B. y De Jesús, L. (2018). Dinámica productiva del sector automotriz y la manufactura en la frontera norte de México: Un análisis con datos de panel, 1980 – 2014. Frontera Norte, Vol. 30, Núm. 30, PP. 29 – 56.

Mayorga, M. y Muñoz, E. (2000). La técnica de datos de panel: Una guía para su uso e interpretación. Departamento de Investigaciones Económicas. Banco Central deCosta Rica.

Wooldridge, J. (2015). Introducción a la Econometría. Quinta Edición. Cengage Learning.

Wooldridge, J. (2001). Econometric Analysis of Cross Section and Panel Data. The MIT Press. Cambridge, Massachusetts.

Salisu, A. (2016). The Proportional Odds Assumption in Ordered Logit/Probit Models. CBN International Training Institute. Nigeria.

Grilli, L. y Rampichi, C. (2014). Ordered Logit Model. University of Florence. Italy.

Williams, R. (2016). Understanding and interpreting generalized ordered logit models. The Journal of Mathematical Sociology.

Long, J. S., & Freese, J. (2006). Regression models for categorical dependent variables using stata (2nd ed.). College Station, TX: Stata Press.

Long, J. S., & Freese, J. (2014). Regression models for categorical dependent variables using stata (3rd ed.). College Station, TX: Stata Press.

Williams, R. (2010). Fitting heterogeneous choice models with oglm. The Stata Journal, 10(4), 540–567.

Fullerton, A. S., & Dixon, J. C. (2010). Generational conflict or methodological artifact? Reconsidering the relationship between age and policy attitudes in the U.S., 1984–2008. Public Opinion Quarterly, 74(4), 643–673.