15 de diciembre de 2021
Consuelo Calderón
Gastón Conejeros
Prof. Mauricio Bucca
Análisis de Datos Categóricos – SOC3070

‘¿Más grande, más peligroso?’: El efecto de la altura de los jugadores sobre la cantidad de tarjetas obtenidas

Pregunta de investigación

Se plantea que el ser más grande, en cuanto a altura, supone un mayor número de tarjetas, ya sean rojas o amarillas. Esto sucede porque, por un lado, ser más grande supone un mayor peligro visual desde la perspectiva del árbitro, por lo que tendería a sancionar al jugador en mayor medida y, por el otro, el hecho de ser más grande implica mayor torpeza en cuanto al movimiento de las extremidades, particularmente al hacer pressing.

- Al ejecutar un test de Chi-Cuadrado, se evidencia que la relación entre la cantidad de tarjetas obtenidas y la altura es significativa:

Test Chi - Cuadrado

χ^2 = 1278,7, df = 574, p-value < $2{,}2e^{-16}$

En consecuencia, interesa saber si la altura tiene un efecto en la tasa de tarjetas obtenidas, es decir, ¿ser más grande te hace más propenso a obtener alguna tarjeta como sanción, controlando por el peso y la posición del jugador?

El peso se consideraría como variable control, pues el hecho de ser más alto significa mayor peso, además de que esta variable también estaría aportando en términos del tamaño de los jugadores, por lo que la correlación entre la altura del jugador y el número de tarjetas obtenidas puede estar mediada por su peso.

Se considera el posicionamiento como variable control, pues en distintas posiciones las características físicas “óptimas” o “ideales” de los jugadores pueden variar (ej.: los defensas tienden a ser más altos, ya que necesitan detener el balón y “romper” la jugada de ataque, mientras que los volantes de creación les puede ser útil el hecho de ser más pequeños, al ser más difíciles de arrebatarles el balón en espacios acotados), por lo que la correlación entre la magnitud del jugador y el número de tarjetas obtenidas puede estar moderada por su posición[1].

[1] A pesar de que la variable ‘peso’ y ‘posición’ serán concebidas como variables mediadora y moderadora respectivamente, en un inicio el modelo de regresión de interés será construido como si estas variables fuesen

Datos y Métodos

La base de datos utilizada es relativamente grande, con un N de futbolistas de 2.053. Se decidió comenzar con un modelo de regresión OLS, con tal de analizar los resultados con el modelo más simple y de fácil interpretación:

$$y_{Tarjeta} = \beta_0 + \beta_1 x_{Altura} + \beta_2 x_{Peso} + \beta_3 x_{Posición} + e_i$$

$$\Rightarrow E(y|x_{Altura}, x_{Peso}, x_{Posición}) = \beta_0 + \beta_1 x_{Altura} + \beta_2 x_{Peso} + \beta_3 x_{Posición}$$

Luego, se decidió introducir una GLM con link logit (regresión logística), para así comparar conclusiones entre modelos, procurando robustez. Cabe destacar que para este modelo dicotomizamos la variable dependiente, entre 'no obtener tarjetas' y 'obtener tarjetas', para así poder expresar la variable en una Bernoulli:

$$\ln(y_{Tarjeta\ (i=1)}) = logit(p_i) = ln\frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{Altura} + \beta_2 x_{Peso} + \beta_3 x_{Posición} + e_i$$

$$\Rightarrow y_{Tarjeta\ (i=1)} = \frac{p_i}{1-p_i} = e^{\beta_0 + \beta_1 x_{Altura} + \beta_2 x_{Peso} + \beta_3 x_{Posición}}$$

$$\Rightarrow E(y_i|x_{Altura}, x_{Peso}, x_{Posición}) = e^{\beta_0 + \beta_1 x_{Altura} + \beta_2 x_{Peso} + \beta_3 x_{Posición}}$$

Finalmente, se configuró un modelo de regresión Poisson (el central en nuestro análisis), dado que la variable dependiente *y* (obtener tarjeta) adquiere valores enteros, desde cero hasta infinito positivo, por lo que la relación se podría escribir como la "cantidad esperada de tarjetas dada la complexión física y la posición del jugador". Por ende, tiene sentido modelar una regresión Poisson para el conteo del evento de interés:

$$\ln(\mu_i) = \beta_0 + \beta_1 x_{Altura} + \beta_2 x_{Peso} + \beta_3 x_{Posición} + \ln(n_i)$$

$$\Rightarrow E(y_i|x_{Altura}, x_{Peso}, x_{Posición}) = \mu_i = e^{\beta_0 + \beta_1 x_{Altura} + \beta_2 x_{Peso} + \beta_3 x_{Posición}}$$

- Donde $y_i \sim Poisson(\mu_i)$

Asimismo, como se mencionó en un inicio, la variable sobre la posición del jugador es concebida como una variable moderadora de la correlación entre altura y cantidad de tarjetas obtenidas, por lo que tendría sentido introducir una interacción entre la posición y la altura al modelo de regresión Poisson:

$$\ln(\mu_i) = \beta_0 + \beta_1 x_{Altura} + \beta_2 x_{Peso} + \beta_3 x_{Posición} + \beta_4 x_{altura} x_{Posición} + \ln(n_i)$$

$$\Rightarrow E(y_i|x_{Altura}, x_{Peso}, x_{Posición}) = \mu_i = e^{\beta_0 + \beta_1 x_{Altura} + \beta_2 x_{Peso} + \beta_3 x_{Posición} + \beta_4 x_{altura} x_{Posición}}$$

- Donde $y_i \sim Poisson(\mu_i)$

Por último, se decidió explorar la necesidad de una regresión Quasi-Poisson para el mejoramiento del modelo con interacción:

---

solamente de control, pero en un segundo modelo se integrará una interacción entre la variable independiente de interés ('altura') y la variable moderadora ('posición').

$$y_i \sim quasi - Poisson(u_i = e^{\beta_0+\beta_1 x_{Altura}+\beta_2 x_{Peso}+\beta_3 x_{Posición}+\beta_4 x_{altura} x_{Posición}}, \sigma = \sqrt{\omega \cdot e^{\beta_0+\beta_1 x_{Altura}+\beta_2 x_{Peso}+\beta_3 x_{Posición}+\beta_4 x_{altura} x_{Posición}}})$$

Resultados

Tabla 1: Modelos de regresión

	Regresión OLS	**Regresión logística**	**Modelo Poisson**	**Modelo Poisson con efecto interacción**	**Modelo de regresión Quasi-Poisson con efecto interacción**
Intercepto	0,561***	-1,057***	-0,902***	7,267***	7,267***
	(0,081)	(0,219)	(0,152)	(1,088)	(1,385)
Altura	-0,004***	-0,010***	-0,010***	-0,054***	-0,054***
	(0,001)	(0,002)	(0,001)	(0,006)	(0,007)
Peso	0,004***	0,011***	0,011***	0,012***	0,012***
	(0,001)	(0,001)	(0,001)	(0,001)	(0,001)
Posición: Defensa	0,350***	1,247***	1,230***	-5,735***	-5,735***
	(0,009)	(0,030)	(0,025)	(1,105)	(1,406)
Posición: Mediocampo	0,327***	1,202***	1,183***	-8,945***	-8,945***
	(0,009)	(0,032)	(0,026)	(1,105)	(1,406)
Posición: Delantero	0,178***	0,827***	0,800***	-6,432***	-6,432***
	(0,010)	(0,032)	(0,027)	(1,123)	(1,429)
Altura*Posición: Defensa				0,037***	0,037***

Tabla 1: Modelos de regresión

	Regresión OLS	**Regresión logística**	**Modelo Poisson**	**Modelo Poisson con efecto interacción**	**Modelo de regresión Quasi-Poisson con efecto interacción**
				(0,006)	(0,007)
Altura*Posición: Mediocampo				0,054***	0,054***
				(0,006)	(0,007)
Altura*Posición: Delantero				0,038***	0,038***
				(0,006)	(0,008)
Factor de dispersión					1,620
R^2	0,018				
R^2 Ajustado	0,018				
Observaciones	126947	126947	126947	126947	126947
AIC		147624,709	223587,636	223402,544	
BIC		147683,218	223646,145	223490,308	
Log Likelihood		-73806,354	-111787,818	-111692,272	
Deviance		147612,709	144579,936	144388,845	144388,845

***p < 0,001; **p < 0,01; *p < 0,05

Fuente: Elaboración propia a partir de Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results" (Silberzahn et at 2018)

Dado que los coeficientes observados en un modelo de regresión Poisson corresponden a un logaritmo natural, para hacer más comprensible e intuitiva la interpretación, se transformaron los coeficientes beta de regresión a su expresión exponenciada:

- Interpretación de los coeficientes del modelo de regresión Poisson simple:
    - Altura → independientemente del resto de variables independientes, en promedio la altura supone un 0,010 menos en el logaritmo natural de la cantidad de recuentos de tarjetas obtenidas. En otras palabras, un cambio de un centímetro en la altura del jugador multiplica el recuento de tarjetas recibidas por 0,9896.
    - Peso → con independencia del resto de covariables, el peso, en promedio, significa un 0,011 más en el logaritmo natural del número de tarjetas obtenidas. Es decir, un cambio de una unidad del jugador multiplica el recuento de tarjetas recibidas por 1,0112.
    - Posición: Defensa → independientemente de las otras variables, la posición de defensa supone, en promedio, un 1,230 más en el logaritmo natural de la cantidad de tarjetas recibidas, en contraste con la posición de arquero. Consecuentemente, el hecho de que el jugador sea defensa multiplica el recuento de la cantidad de tarjetas recibidas en 3,4225, en contraste con la posición de arquero.
    - Posición: Mediocampo → con independencia del resto de covariables, la posición de mediocampo significa un 1,183 más en el logaritmo natural del total de recuentos de tarjetas recibidas, en comparación con ser aquero. Por ende, el hecho de que el jugador sea mediocampista multiplica el recuento de la cantidad de tarjetas recibidas en 3,2625, en contraste con la posición de arquero.
    - Posición: Delantero → independientemente de las demás variables, la posición de delantero implica un aumento del 0,800 en el logaritmo natural del número de tarjetas recibidas, en promedio y comparando con la posición de arquero. Correlativamente, el hecho de que el jugador sea delantero multiplica el recuento de la cantidad de tarjetas recibidas en 2,2265, en contraste con la posición de arquero.
- Interpretación de los coeficientes del modelo de regresión quasi-Poisson con efecto interacción:
    - Se evidencia que la interacción entre la posición de defensa y la altura implicaría, en promedio, 5,752 menos en el logaritmo natural de la cantidad de tarjetas obtenidas, en contraste con la posición de arquero y con independencia del resto de variables independientes. Implicando que, en el grupo de los defensas, el aumento de un centímetro de altura multiplica el recuento esperado de tarjetas recibidas por 0,0032 en comparación con los arqueros.
    - Se aprecia que los mediocampistas, en interacción con la altura, implica un 8,945 menos en el logaritmo natural del total de tarjetas recibidas por el jugador, en promedio y en contraposición a los arqueros, independientemente de las demás covariables. Conllevando a que un aumento de un centímetro de altura multiplica el recuento esperado de tarjetas recibidas por 0,0001 para los mediocampistas, en comparación con los arqueros.
    - Se observa que la posición de delantero en interacción con la altura, con independencia de las otras variables, conlleva a un 6,448 menos en el logaritmo

natural del número de tarjetas obtenidas, comparándola con la posición de arquero. Es decir, en el grupo de los delanteros el aumento de un centímetro de altura en el jugador multiplica el recuento esperado de tarjetas recibidas por 0,0016 en comparación con los arqueros.

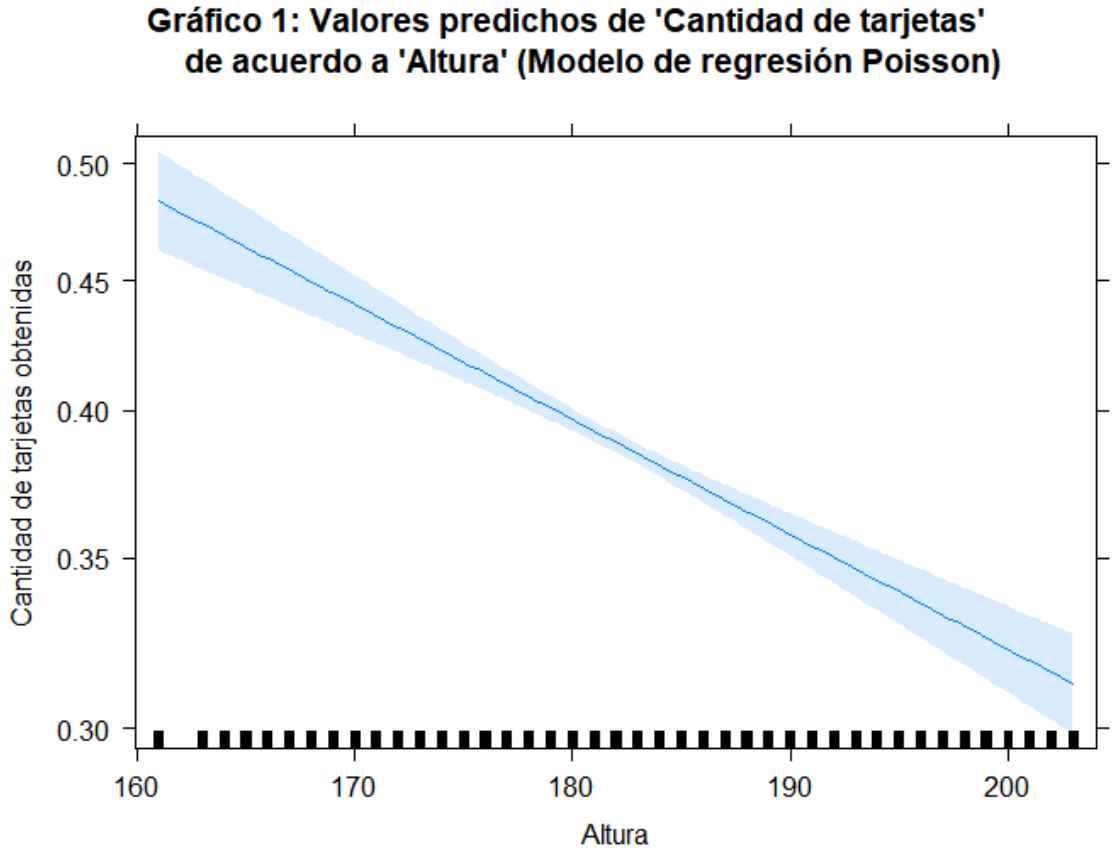
- El peso significa un 0,012 más en el logaritmo natural de las tarjetas recibidas por el jugador en promedio, independientemente de las demás covariables. Esto quiere indicar que el aumento de una unidad del peso del jugador multiplica el recuento esperado de tarjetas recibidas por 1,0116.

Notamos que todos los coeficientes de los cinco modelos de regresión son estadísticamente significativos al 99,9% nivel de confianza. Más aún, se observa, como conclusión general de todos los modelos, que los defensas son aquellos más propensos a recibir una tarjeta por el árbitro, seguido de los mediocampistas y, al final, por los delanteros, al compararlo con la posición de arquero. Por otro lado, se destaca que el BIC más pequeño para los GLM corresponde a la regresión logística, lo que estaría significando que este sería el modelo con el mejor ajuste. No obstante, como nuestro interés radica en la regresión Poisson, se evidencia que aquel con efecto interacción es el que tiene mejor ajuste (menor BIC) y, más aún, el modelo de regresión quasi-Poisson con efecto interacción culminaría por ser el "más indicado" para efectuar la interpretación, dado que corrige por la inferencia del modelo. Asimismo, se evidencia que contrariamente a la hipótesis inicial, una mayor altura supondría una leve menor cantidad de tarjetas recibidas por el jugador. No obstante, el peso estaría aportando, aunque de manera muy pequeña, a una mayor cantidad de tarjetas obtenidas. Sin embargo, esta conclusión podría estar siendo afectada por la presencia de variables no observadas que, en su momento, no se midieron, como la tendencia del jugador a presentar *fairplay*, así como también la sinergia que cada jugador tiene con su equipo en cuestión, por ejemplo. Asimismo, alguna variable independiente podría estar experimentando un comportamiento diferente al modelado tal que no haya sido explorado, como por ejemplo la altura podría estar presentando un efecto cuadrático y este no haya sido tomado en cuenta.

En cuanto al modelo de regresión quasi-Poisson, se aprecia que el factor de dispersión ω es extremadamente cercano a uno, lo que implica que los coeficientes de regresión del modelo quasi-Poisson sean casi iguales a los del modelo Poisson con efecto interacción, indicando que no existirán diferencias estrictas al analizar estas correlaciones utilizando cualquiera de los dos modelos, pues las conclusiones van a ser las mismas, a grandes rasgos.

Apéndice: Análisis suplementarios

En primera instancia, se exigió un gráfico con tal de explorar visualmente la relación entre altura y cantidad de tarjetas obtenidas en base al modelo de regresión Poisson sin efecto interacción, en cuanto a los valores predichos que se pueden calcular. En este, es factible concluir que, contrariamente a lo que se proponía en un inicio, una mayor altura estaría implicando un mayor número de tarjetas recibidas por algún jugador:

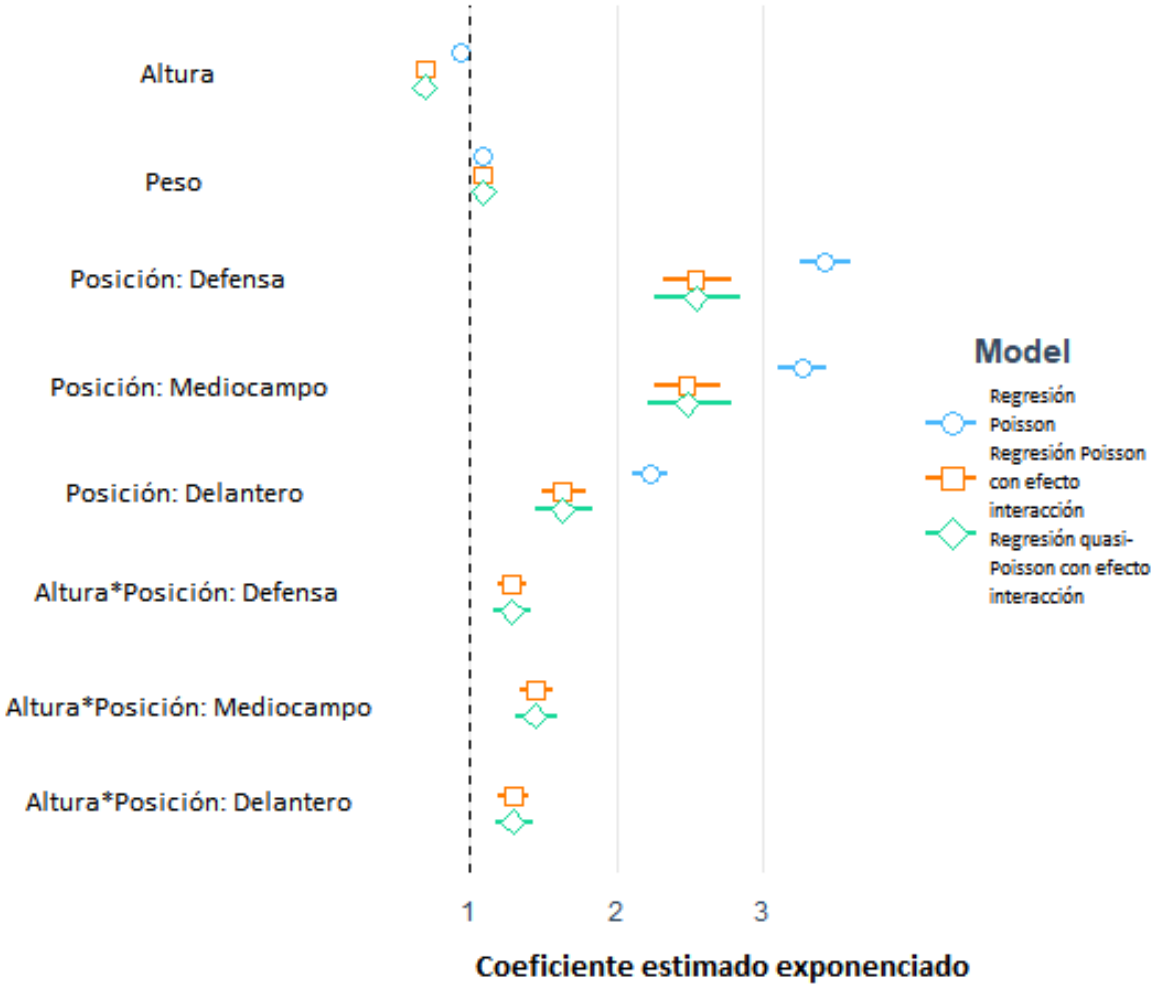


Fuente: Elaboración propia a partir de Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results" (Silberzahn et at 2018)

Cabe destacar que los intervalos de confianza de las estimaciones en cuanto a la cantidad de tarjetas obtenidas por el jugador según su altura son más amplios en los valores extremos sobre la altura, mientras que tienden a angostarse en los valores intermedios, lo que puede indicar que la distribución de la altura se da de forma normal, pues habría un mayor número de observaciones en los valores intermedios y, por ende, implicarían estimaciones más exactas, explicando, recursivamente, el porqué de las expresiones de los intervalos de confianza en el Gráfico 2.

En segundo lugar, se construyó un gráfico tal que contenga y compare los coeficientes de regresión estimados tanto para el modelo de regresión Poisson 'normal' como para el con efecto interacción, incluyendo sus intervalos de confianza correspondientes:

Gráfico 2: Coeficientes estimados para el Modelo de regresión Poisson simple, el Modelo de regresión Poisson con efecto interacción y el Modelo de regresión quasi-Poisson con efecto interacción



Fuente: Elaboración propia a partir de Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results" (Silberzahn et at 2018)

Del gráfico, es posible observar que todos los coeficientes son estadísticamente significativos a un 95% nivel de confianza (pues no entrecruzan el 1, que refiere cuando el coeficiente beta de regresión igual a cero es exponenciado). Sin embargo, pareciese que el peso y la altura no suponen tendencia a obtener tarjetas de mayor magnitud que indica la variable sobre la posición, al comparar en referencia a los arqueros. Asimismo, cabe destacar que es la altura la única variable del modelo que implica una menor cantidad de tarjetas obtenidas a medida que aumenta la altura, reforzando la conclusión mencionada anteriormente. Más aún, se aprecia que para el modelo de regresión Poisson simple no hay diferencias sustanciales en cuanto a ser defensa o mediocampista al momento de recibir alguna tarjeta por el árbitro, en referencia a ser arquero,

no obstante, la posición de delantero supone una menor cantidad de tarjetas obtenidas que estas otras dos posiciones, aunque obviamente presenta una propensión mayor a obtener tarjetas que los arqueros. Al visualizar los modelos con efecto interacción, se evidencia que este efecto es estadísticamente el mismo entre todas las posiciones, puesto que los intervalos de confianza se entrecruzan, además de ser positivos (coeficiente exponenciado mayor a 1), lo que indica que una mayor altura supondrá una menor tendencia a obtener tarjetas para los defensas, mediocampistas y delanteros.

Por otro lado, con tal de reforzar la argumentación de que la altura tiene incidencia en la predicción/explicación de la variable sobre la cantidad de tarjetas para un modelo de regresión quasi-Poisson con efecto interacción, se configuró un Bootstrap en torno a esta variable independiente de interés, para así conseguir el Average Marginal Effect y estimar su significancia estadística. Se obtiene que el intervalo de confianza de este estimador, a un 95% nivel de confianza, se encontrará en el rango {-812,350; -3,360}, significando que el efecto marginal promedio de la altura sobre el número de tarjetas obtenidas por un jugador será estadísticamente significativo, pues el intervalo no entrecruza el cero, además de ser negativo, lo que implica que a mayor altura, el jugador tenderá a obtener una menor cantidad de tarjetas rojas o amarillas.

Luego, se realizó un *cross-validation* para poner a prueba nuestro modelo. El modelo de 'entrenamiento' se realiza en una submuestra correspondiente al 80% de la muestra original, dejando el otro 20% como la submuestra de 'testeo'. En teoría, si es que el modelo está correcto, la proporción de personas que reciben cada cantidad de tarjetas en una submuestra con respecto a la otra deberían ser similares entre sí.

Submuestra 80%	Conteo de tarjetas	0	1	2	3	4	5	6	7	8	9	10	11	12	13	15	Total
	Número de jugadores	73237	20429	4832	1767	714	319	136	71	32	10	6	2	1	1	1	101558
	Porcentaje	72,113%	20,116%	4,758%	1,740%	0,703%	0,314%	0,134%	0,070%	0,032%	0,010%	0,006%	0,002%	0,001%	0,001%	0,001%	100%
Submuestra 20%	Conteo de tarjetas	0	1	2	3	4	5	6	7	8	9	10	11	12	13	15	Total
	Número de jugadores	18309	5107	1208	442	178	80	34	18	8	2	2	1	0	0	0	25389
	Porcentaje	72,114%	20,115%	4,758%	1,741%	0,701%	0,315%	0,134%	0,071%	0,032%	0,008%	0,008%	0,004%	0%	0%	0%	100%

Fuente: Elaboración propia a partir de Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results" (Silberzahn et at 2018)

Tal como se puede observar, los porcentajes se respetan en ambas muestras de manera bastante certera, con lo que se comprueba una validación cruzada de la variable sobre la obtención de tarjetas.

Por último, con tal de indagar en la preocupación final de que alguna de las variables independientes del modelo esté expresándose para con la variable dependiente de alguna manera no lineal, se decidió por probar esta hipótesis para la altura, concibiéndola como un efecto cuadrático:

$$\ln(\mu_i) = \beta_0 + \beta_1 x_{Altura} + \beta_2 x_{Altura}^2 + \beta_3 x_{Peso} + \beta_4 x_{Posición} + \ln(n_i)$$

$$\Rightarrow E(y_i | x_{Altura}, x_{Peso}, x_{Posición}) = \mu_i = e^{\beta_0 + \beta_1 x_{Altura} + \beta_2 x_{Altura}^2 + \beta_3 x_{Peso} + \beta_4 x_{Posición}}$$

- Donde $y_i \sim Poisson(\mu_i)$

Tabla 2: Modelos de regresión Poisson

	Modelo Poisson con efecto cuadrático	**Modelo Poisson**
Intercepto	-35,912***	-0,902***
	(2,771)	(0,152)
Altura	0,377***	-0,010***
	(0,031)	(0,001)
Altura^2	-0,001***	
	(0,000)	
Peso	0,012***	0,011***
	(0,001)	(0,001)
Posición: Defensa	1,185***	1,230***
	(0,025)	(0,025)
Posición: Mediocampo	1,133***	1,183***
	(0,026)	(0,026)
Posición: Delantero	0,753***	0,800***
	(0,027)	(0,027)
AIC	223422,717	223587,636
BIC	223490,977	223646,145
Log Likelihood	-111704,358	-111787,818
Deviance	144413,017	144579,936

Tabla 2: Modelos de regresión Poisson

	Modelo Poisson con efecto cuadrático	Modelo Poisson
Observaciones	126947	126947

$^{***}p < 0,001$; $^{**}p < 0,01$; $^{*}p < 0,05$

Fuente: Elaboración propia a partir de Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results" (Silberzahn et at 2018)

Se evidencia que los coeficientes de regresión cambiaron visiblemente, lo que puede constituir el primer indicio de que el efecto de la altura es cuadrático. Asimismo, se explicita que el coeficiente que acompaña al término cuadrático sobre la altura es estadísticamente significativo, a un 99,9% nivel de confianza. Por último, el BIC que contiene al efecto cuadrático es levemente menor al que lo excluye, implicando que el ajuste del modelo de regresión Poisson que concibe a la altura como cuadrática en su relación con la cantidad de tarjetas obtenidas es mejor que el otro por lo que es factible afirmar que el comportamiento de la altura no es lineal. Como el término cuadrático con respecto a la altura es negativo, se infiere que su comportamiento es cóncavo, es decir, cuando la altura es baja, una mayor altura significará una mayor propensión a obtener tarjetas en partidos de fútbol, sin embargo, se llega a un punto crítico donde esta relación se invierte, es decir, una mayor altura supone un menor número de tarjetas recibidas.

Es posible determinar aquel punto crítico derivando la expresión e igualándola a cero, obteniendo así la recta tangente a la curva donde ocurre el cambio de signo de la pendiente, controlando por las demás variables independientes:

$$\mu_i = e^{\beta_0+\beta_1 x_{Altura}+\beta_2 x^2_{Altura}}$$

$$\mu'_i = (\beta_1 + 2 \cdot \beta_2 x_{Altura}) e^{\beta_0+\beta_1 x_{Altura}+\beta_2 x^2_{Altura}}$$

Al igualarlo a cero, se evidencia que es imposible que la expresión exponenciada sea igual a cero, por lo que inevitablemente la expresión lineal será quien sea igual a cero:

$$\beta_1 + 2 \cdot \beta_2 x_{Altura} = 0 \Rightarrow x_{Altura} = \frac{-\beta_1}{2 \cdot \beta_2}$$

Reemplazando:

$$x_{Altura} = \frac{-0,377}{2 \cdot -0,001} = 188,5$$

Introduciendo el valor encontrado de x para la altura cuando la derivada de la función es igual a cero en la ecuación original:

$$\mu_i = e^{-35,912+0,377 \cdot 188,5-0,001 \cdot 188,5^2} = e^{-2,866} = 0,057$$

$$\ln(\mu_i) = -2,866$$

En consecuencia, desde la altura más baja hasta los 188,5 centímetros, la cantidad de tarjetas que un jugador tenderá a obtener serán mayores a medida que su altura aumenta. No obstante, cuando se llega a los 188,5 o, en otras palabras, a las 0,057 tarjetas recibidas (que esto en la práctica es imposible, pero funciona en términos analíticos) esta relación se invierte, implicando que una mayor altura significa una menor propensión a obtener más tarjetas.