Análisis del salario de jugadores en la NBA

Nombre: Martín Rodríguez Valiente

Contenido

Introducción

El siguiente trabajo tiene el objetivo de analizar los salarios de jugadores profesionales de baloncesto en la NBA (National Basketball Association). Se busca, mediante el uso de una muestra de jugadores y un método estadístico (Test de Bondad de Ajuste chi-cuadrado), realizar una inferencia sobre la distribución que presentan los salarios, planteando la hipótesis de que siguen una distribución Exponencial. A sí mismo, se realiza una regresión lineal múltiple, con la finalidad de encontrar variables que expliquen el comportamiento del salario, y analizar su significancia en el valor de los mismo.

Para tales objetivos, se obtuvieron datos de la situación contractual de cada jugador activo en la temporada 2020-2021, y sus rendimientos en la temporada anterior (2019-2020). Buscando si hay alguna relación en sus salarios actuales y sus desempeños previos.

Bondad de Ajuste

Para realizar la prueba de Bondad de Ajuste, se obtuvieron datos del salario de cada jugador de la NBA (En este trabajo todos los valores con "$" representan valores en moneda dólar estadounidense). Los mismos fueron extraídos del sitio online https://www.basketball-reference.com. El mismo recopila información detallada de las estadísticas individuales de cada jugador año a año, así como la situación contractual.

Se selecciona para el presente análisis, la información del año 2020-2021. Con un total de 490 jugadores.

Muestra

Para poder obtener la muestra, primero se establece el tamaño adecuado que debe tener la misma, de forma que los datos obtenidos sean representativos de la población.

Para calcular el tamaño de la muestra se usará la siguiente formula:

$$n = \frac{N \cdot \sigma^2 \cdot Z_\alpha^2}{e^2 \cdot (N-1) + \sigma^2 \cdot Z_\alpha^2}$$

Donde:

n = Tamaño adecuado de la muestra para nuestro análisis

N = Tamaño de la población

σ^2 = Varianza poblacional

Z_α : Valor obtenido mediante niveles de confianza, es un valor constante que surge de una distribución Normal Estándar, con media 0 (cero) y varianza 1 (uno).

e = Límite aceptable de error muestral

Para obtener la varianza poblacional se realiza la función "VAR.P()" en Excel, que realiza el siguiente procedimiento:

$\frac{\sum_{i=0}^{n}(x_i-\mu)^2}{n}$ = Var(x) = 83.491.485.479.184

Siendo "x_i" cada observación de x; y "μ" la media poblacional.

Ahora que disponemos de la varianza poblacional, nos queda determinar (Z_α), utilizaremos una confianza de 95% y utilizando una tabla de distribución normal estándar se obtiene un valor de 1,96. En cuanto al límite de error (e) se empleara un criterio de 10% del desvío poblacional, esto es el 10% de $\sqrt{Var\,(x)}$.

e = $0{,}1 \cdot \sqrt{83.491.485.479.184} = 913.737$

Por lo que ahora podemos obtener el tamaño de la muestra que vamos a utilizar:

$$n = \frac{N \cdot \sigma^2 \cdot Z_\alpha^2}{e^2 \cdot (N-1) + \sigma^2 \cdot Z_\alpha^2} = \frac{490 \cdot 83.491.485.479.184 \cdot 1.96^2}{913.737^2 \cdot (490-1) + 83.491.485.479.184 \cdot 1.96^2} \cong 216$$

Para obtener nuestras 216 observaciones, se realiza un método de muestreo aleatorio simple, que implica que cada elemento dentro de nuestra lista de salarios tiene la misma oportunidad de resultar seleccionado.

Este método nos asegura que cada una de las observaciones que resulte seleccionada será una variable aleatoria, cuya distribución de probabilidad es idéntica a la poblacional, a su vez, las mismas son independientes e idénticamente distribuidas, ya que ninguna observación se ve afectada por otra.

Realizaremos una numeración de cada salario, un conteo de 1 a 490, nuestra cantidad poblacional, y mediante la función "ALEATORIO.ENTRE()" seleccionaremos 216 números aleatorios en el rango antes mencionado, para posteriormente recolectar nuestras observaciones con la función "BUSCARX()".

Para agrupar los valores de nuestra muestra y preparar para la prueba de Bondad de Ajuste, se divide la misma en intervalos de $6.000.000, y se crea un cuadro con sus respectivas frecuencias observadas:

Al tener tal concentración en valores menores, se puede observar en la distribución una marcada asimetría positiva, es decir, presenta una mayor densidad de observaciones hacia la izquierda de la media.

Por este motivo, se decide plantear como Hipótesis que los datos siguen una distribución exponencial, dicha distribución es adecuada para modelizar variables continuas con forma asimétrica positiva y, en especial, suele ser muy utilizada para modelizar ingresos salariales.

La distribución exponencial es un caso particular de la distribución gama, donde *alpha* es igual a 1 y *beta* es un valor positivo.

Utilizaremos el término del parámetro escala $\beta = 1/\lambda$

De esta forma, la función de densidad queda definida como:

$$f_X(x, \beta) = \frac{1}{\beta}\, e^{-\frac{x}{\beta}}$$

Para todo $x \geq 0$

Donde:

La esperanza de x: E(x) = β

La varianza de x: Var(x) = β^2

Test de bondad de ajuste chi-cuadrado

Definiremos nuestra variable a estudiar.

X: Salario de jugadores NBA en 2020-2021

Se intentará inferir la distribución que siguen los salarios mediante una prueba que mide la discrepancia entre una distribución observada (en este caso, nuestra variable X) y otra teórica. Realizaremos una prueba, de igual contra mayor, para decidir si nuestros datos se apegan a una distribución exponencial.

Primero, plantemos nuestra hipótesis nula:

$$H_0: F(x) = F_0(x)$$

Donde $F_0(x)$ es nuestro modelo de probabilidad exponencial propuesto.

El estadístico de prueba es el siguiente:

$$D^2 = \sum_{i=0}^{k} \frac{(observada_i - teorica_i)^2}{teorica_i} \sim \chi_t^2(k\ -1-r)$$

Y nuestro criterio de decisión es el siguiente:

$$D^2 < \chi_t^2(k\ -1-r)$$

Donde los grados de libertad de la chi-cuadrado $(k\ -1-r)$ son:

k = cantidad de clases exhaustivas y mutuamente excluyentes

r = número de parámetros estimados

t representa el valor proporcionado por la distribución, según el nivel de significación elegido, en este caso elegiremos un 5% de significación (Error de Tipo 1).

Armamos la siguiente tabla para estimar nuestra media muestral, en orden de armar nuestras probabilidades siguiendo una distribución exponencial.

Clases	Inf	Sup	Punto medio	Frecuencia observada	xi * fi
0-6000000	-	6.000.000	3.000.000	136	408.000.000
6000000-12000000	6.000.000	12.000.000	9.000.000	28	252.000.000
12000000-18000000	12.000.000	18.000.000	15.000.000	17	255.000.000
18000000-24000000	18.000.000	24.000.000	21.000.000	10	210.000.000
24000000-30000000	24.000.000	30.000.000	27.000.000	12	324.000.000
30000000-36000000	30.000.000	36.000.000	33.000.000	10	330.000.000
36000000-42000000	36.000.000	42.000.000	39.000.000	2	78.000.000
42000000-48000000	42.000.000	48.000.000	45.000.000	1	45.000.000

$$\beta = \bar{x} = \frac{\sum x_i \cdot f_i}{n} = \frac{1.902.000.000}{216} = 8.805.556$$

Con nuestro parámetro *beta* estimado, podemos conseguir nuestras probabilidades siguiendo una distribución exponencial, utilizando la función "DISTR.GAMMA.N()" con *alpha* = 1 y *beta* = 8.805.556. Y luego obtener las frecuencias esperadas:

Clases	Distr. Exp.	Frec. Esperada
0-6000000	0,494	106,723
6000000-12000000	0,250	53,992
12000000-18000000	0,126	27,316
18000000-24000000	0,064	13,819
24000000-30000000	0,032	6,991
30000000-36000000	0,016	3,537
36000000-42000000	0,008	1,789
42000000-48000000	0,008	1,832

Al generar las frecuencias esperadas de cada clase, notamos que en tres de ellas el valor es bastante pequeño. Seguiremos un criterio conservador difundido en (Canavos, 1988) donde toda frecuencia esperada no puede ser menor que cinco. Para lograr esto combinaremos clases vecinas, para sumar sus frecuencias. Al hacer esto, perderemos un grado de libertad por cada par de clases agrupado. También debemos sumar las frecuencias observadas de esas clases para poder realizar nuestro estadístico de prueba.

Frec. Esperada	Frec. Observada	Est. prueba
106,7225135	136	8,032
53,99244458	28	12,513
27,31554923	17	3,896
13,81932668	10	1,056
6,991394836	12	3,588
7,158771189	13	4,766

La columna de "Est. prueba" se obtiene haciendo $\frac{(observada_i - teorica_i)^2}{teorica_i}$ donde la frecuencia teórica, es nuestra frecuencia esperada en el cuadro.

Con estos datos ya podemos conseguir nuestro valor empírico mediante el estadístico de prueba. A su vez, necesitamos un valor crítico para contrastar nuestra hipótesis, como nuestro Error de Tipo 1 es de 5%, entonces, nuestro valor critico es el valor que acumula un 5% de probabilidad en la cola derecha de una distribución chi-cuadrado con (k-1-r) grados de libertad. Como estimamos un parámetro β y agrupamos dos pares de clases, r es igual a 3 grados de libertad.

Grados de libertad = (k-1-r) = (8-1-3) = 4

Los resultados son los siguientes:

D^2	$33{,}85$
P-value	$8 \cdot 10^{-7}$
Valor Crítico	$9{,}49$
Región Crítica	$0{,}05$

El estadístico de prueba nos da un valor mayor a nuestro valor crítico, lo que nos indica *rechazar* nuestra H_0 sobre la forma funcional de la población. También podemos verlo mediante el P-value, que contiene la probabilidad asociada al estadístico de prueba, el mismo es menor a nuestra región critica de 0,05.

Regresión Lineal Múltiple

Para realizar un estudio sobre cuales aspectos del desempeño de los jugadores tiene relación con sus salarios, se obtiene una muestra, de igual proceso como con la Bondad de Ajuste. Como estamos analizando los salarios en 2020-2021 y el desempeño de los jugadores en la temporada anterior (2019-2020), la población se reduce, ya que hay jugadores que no continúan en la liga. Por lo que ahora nuestra población consiste en 414 jugadores.

Para obtener el tamaño de nuestra muestra, reutilizamos la fórmula:

$$n = \frac{N \cdot \sigma^2 \cdot Z_\alpha^2}{e^2 \cdot (N-1) + \sigma^2 \cdot Z_\alpha^2}$$

Donde:

$\sigma^2 = \frac{\sum_{i=0}^{n}(x_i - \mu)^2}{n} = 85.896.253.327.504$

$Z_\alpha = 1{,}96$

$e = 0{,}1 \cdot \sqrt{85.896.253.327.504} = 926.802$

Entonces:

$$n = \frac{N \cdot \sigma^2 \cdot Z_\alpha^2}{e^2 \cdot (N-1) + \sigma^2 \cdot Z_\alpha^2} = \frac{414 \cdot 85.896.253.327.504 \cdot 1.96^2}{926.802^2 \cdot (414-1) + 85.896.253.327.504 \cdot 1.96^2} \cong 200$$

Mediante un muestreo aleatorio simple, se obtienen 200 jugadores y su correspondiente salario y performance en la temporada, la misma se basa en estadísticas por juego, son métricas de popular uso para medir el desempeño de un jugador en esta liga, dentro de las cuales se encuentran “puntos por juego”, “asistencias por juego”, “minutos jugados por juego”, entre otras.

Para este análisis se seleccionaron 6 variables, que según criterio del investigador puede tener significatividad en la explicación de los ingresos de un jugador. Estas variables explicativas se utilizan para realizar un método de selección de variables denominado regresión por pasos (stepwise) en orden de obtener la mejor ecuación de regresión. El mismo realiza iteraciones y determina qué variables cumplen con el requisito de P-value exigido para quedarse en el modelo.

Existen dos versiones de este procedimiento: (1) selección hacia adelante y (2) eliminación hacia atrás.

(1): Esta variante comienza con una ecuación que no contiene variables de predicción e introduce variables conforme quedan como significativas en el modelo (en nuestro caso decimos que añada la variable mientras que su P-value sea menor de 0,05).

(2): Consiste en meter a todas las variables al principio y sustrae aquellas que pierden la significación (en nuestro caso decimos que retire la variable cuando su P-value sea mayor de 0,05).

Realizamos ambos procedimientos:

(1) selección hacia adelante

```
                      begin with empty model
p = 0.0000 <   0.0500  adding  PTS
p = 0.0000 <   0.0500  adding  Age
p = 0.0005 <   0.0500  adding  AST
```

Source	SS	df	MS		
Model	1.0737e+16	3	3.5791e+15	Number of obs	200
Residual	5.8492e+15	196	2.9843e+13	F(3, 196)	119.93
				Prob > F	0.0000
Total	1.6587e+16	199	8.3349e+13	R-squared	0.6474
				Adj R-squared	0.6420
				Root MSE	5.5e+06

Salary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
PTS	784896.5	96157.49	8.16	0.000	595260.4	974532.7
Age	569403.8	97578.42	5.84	0.000	376965.4	761842.2
AST	1106681	312070.3	3.55	0.000	491234.4	1722128
_cons	-1.67e+07	2470490	-6.76	0.000	-2.16e+07	-1.18e+07

(2) eliminación hacia atrás

```
                      begin with full model
p = 0.7080 >= 0.0500  removing TRB
p = 0.5595 >= 0.0500  removing BLK
p = 0.3809 >= 0.0500  removing STL
```

Source	SS	df	MS		
Model	1.0737e+16	3	3.5791e+15	Number of obs	200
Residual	5.8492e+15	196	2.9843e+13	F(3, 196)	119.93
				Prob > F	0.0000
Total	1.6587e+16	199	8.3349e+13	R-squared	0.6474
				Adj R-squared	0.6420
				Root MSE	5.5e+06

Salary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Age	569403.8	97578.42	5.84	0.000	376965.4	761842.2
PTS	784896.5	96157.49	8.16	0.000	595260.4	974532.7
AST	1106681	312070.3	3.55	0.000	491234.4	1722128
_cons	-1.67e+07	2470490	-6.76	0.000	-2.16e+07	-1.18e+07

Ambas variantes coinciden en la selección de variables y, a su vez, con nuestro criterio de identificar a las variables (Age): Edad del jugador, (PTS): Puntos por partido, (AST): Asistencias por partido, como relevantes en nuestro análisis.

Planteamos nuestra ecuación de regresión, que involucra una variable respuesta (y_i) como función de varias variables explicativas (x_i).

Ecuación de regresión múltiple:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

Donde:

y_i = Salario

x_{i1} = Edad

x_{i2} = Asistencias por partido

x_{i3} = Puntos por partido

ε_i = Error aleatorio no observable asociado con y_i

β_0 = Valor de la respuesta media cuando todas las variables de predicción valen cero

$\beta_j \; siendo \; j \; = 1, 2, 3$

Este modelo supone el caso de la teoría basada en el modelo normal, donde las observaciones y_i son variables aleatorias independientes, normalmente distribuidas con:

$E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 \, x_{i2} + \beta_3 \, x_{i3}$

$Var(y_i) = \sigma^2$

con $i = 1, 2, 3, \ldots, n$

Para estimar nuestros parámetros β usamos el método de mínimos cuadrados, que consistete en minimizar la suma de los errores cuadraticos.

La función de error cuadrático está dada por:

$$S(\beta_0, \beta_1, \beta_2, \beta_3) \; = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \; \beta_2 x_{i2} - \beta_3 x_{i3})^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

Para minimizar la función, se realiza una derivación de la misma respecto a cada parámetro y se iguala a cero.

$$\frac{\partial \sum_{i=1}^{n} \varepsilon_i^2}{\partial \beta_j} = 0$$

Resolver este sistema de forma analítica es complicado, por lo que, el resultado de este sistema de ecuaciones se expresa en forma matricial como:

$$(X^t \cdot X)B = X^t \cdot Y$$

$$\text{donde: } \mathrm{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad B = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \quad \mathrm{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{13} \\ 1 & x_{21} & \ldots & x_{23} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{n3} \end{pmatrix} \quad X^t = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_{11} & x_{21} & \ldots & x_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{13} & x_{23} & \ldots & x_{n3} \end{pmatrix}$$

Si la matriz inversa $(X^t \cdot X)^{-1}$ existe, se multiplica en ambos lados:

$$(X^t \cdot X)^{-1} \cdot (X^t \cdot X)B = (X^t \cdot X)^{-1} \cdot (X^t \cdot Y)$$

entonces el estimador por mínimos cuadrados está dado por:

$$B = (X^t \cdot X)^{-1} \cdot (X^t \cdot Y)$$

Realizando la regresión, obtenemos los coeficientes[1] de nuestros parámetros β:

Parámetro	*Coeficientes*
Intercepción (β_0)	-16.697.631
Edad (β_1)	569.404
Asistencias por partido (β_2)	1.106.681
Puntos por partido (β_3)	784.897

Al ser un modelo lineal, podemos interpretar a los coeficientes ($\beta_1, \beta_2, \beta_3$) como la cantidad en la que cambia nuestra variable respuesta (Salario) cuando una variable explicativa cambia en una unidad, permaneciendo el resto de las variables constante. En nuestro caso, vemos que el ingreso crece $569.404 con un año más de edad, el resto constante; que un aumento de un punto por partido, ceteris paribus, aumenta en $784.897 el salario; y que un aumento de una asistencia por partido, ceteris paribus, incrementa el ingreso en $1.106.681.

En cuanto al intercepto (β_0), representa la respuesta media de la variable respuesta, cuando las explicativas son todas nulas, no tiene ningún significado en particular como un término separado del modelo de regresión.

Análisis de la Varianza

Suma cuadrados totales = Suma cuadrados tratamientos + Suma cuadrados residuos

SCT: Suma de cuadrados totales (variabilidad total)

$SCT = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 16.586.530.326.258.400$

SCTra: Suma de cuadrados de los tratamientos (variabilidad explicada por el modelo)

$SCTra = \sum_{i=1}^{n}(\hat{y} - \bar{y})^2 = 10.737.314.824.269.800$

SCR[2]: Suma de cuadrados de los residuos (variabilidad no explicada)

$SCR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\varepsilon_i^2 = 5.849.215.501.988.630$

[1] Podemos verificar que son los mismos valores que obtuvimos en la regresión por pasos (stepwise) anteriormente.
[2] Nótese que el objetivo de la estimación por el método de mínimos cuadrados es la elección de parámetros que hacen que la suma de los cuadrados de los residuos sea lo más chica posible.

$$SCT = SCTra + SCR$$

$$16.586.530.326.258.400 = 10.737.314.824.269.800 + 5.849.215.501.988.630$$

Bondad de ajuste del modelo

Un indicador útil para cuantificar si nuestro modelo ajusta bien, es decir, tiene buena capacidad explicativa y predictiva, es el coeficiente de determinación R^2. Es una medida relativa del grado de asociación lineal entre las variables explicativas y la variable respuesta. Representa la proporción de la variabilidad que explica el modelo, sobre la variabilidad total.

$$R^2 = \frac{SCTra}{SCT}$$

Los valores que toma R^2 están siempre en el intervalo $0 \leq R^2 \leq 1$.

En regresiones múltiples existe un inconveniente con este coeficiente, dado que al agregar variables explicativas al modelo sólo puede aumentar el R^2 y nunca reducirlo.

Por este motivo, se utiliza un R^2 ajustado, dividiendo cada suma de cuadrados por sus correspondientes grados de libertad, de la siguiente forma:

$$R^2(ajustado) = 1 - \frac{\frac{SCR}{n-k-1}}{\frac{SCT}{n-1}}$$

En nuestra regresión obtenemos el siguiente resultado:

$$R^2(ajustado) = 1 - \frac{\frac{5.849.215.501.988.630}{196}}{\frac{16.586.530.326.258.400}{199}} = 0{,}642$$

Test de significación Global

Otra prueba que también podría considerarse como una medida del ajuste del modelo es la prueba de significación global F.

Esta prueba consiste en comprobar si se cumple la hipótesis nula de que todas las estimaciones de los parámetros son cero, H_0: $\hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3 = 0$. Si el parámetro F es significativamente distinto de 0 se rechaza la H_0 y se puede afirmar que alguna $\hat{\beta}_j$ es distinta de 0, por lo que se rechaza la hipótesis de que el modelo no explica nada.

El estadístico de prueba es el siguiente:

$$\frac{\frac{VE}{k}}{\frac{VnoE}{n-k-1}} \sim F_{(k;\, n-k-1)}$$

Al ser *VE* y *VnoE* distribuciones chi-cuadrado con k, y $(n-k-1)$ grados de libertad respectivamente, el cociente entre ambas chi-cuadrado genera una distribución F de Snedecor con k grados de libertad del numerador, y $(n-k-1)$ grados de libertad del denominador.

Esta prueba es una de igual contra mayor. Elegiremos un Error de Tipo 1 de 5% para contrastar la hipótesis.

Nuestro estadístico resulta:

$$\frac{\frac{10.737.314.824.269.800}{3}}{\frac{5.849.215.501.988.630}{196}} = 119{,}93$$

Y luego buscamos nuestro valor crítico, en una F de Snedecor, el valor que acumula un 5% en la cola derecha, con 3 grados de libertad del numerador y 196 grados de libertad del denominador.

F de Snedecor	119,93
P-value	$3{,}9 \cdot 10^{-44}$
Valor critico	2,65

Al ser el estadístico mayor al valor crítico, cae en la región critica, y *rechazamos* la hipótesis de que el modelo no explica nada.

Test de significación Individual

Esta prueba de hipótesis nos permite determinar que variables son significativas en nuestro modelo. Es una prueba de igual contra distinto con una distribución t de Student.

Las hipótesis son:

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

El estadístico de prueba es el siguiente:

$$\frac{\hat{\beta}_j - \beta_j}{S(\beta_j)} \sim t_{n-k-1}$$

Se realiza el estadístico de cada parámetro:

$$\frac{\hat{\beta}_1 - 0}{S(\beta_1)} = 5{,}84$$

$$\frac{\hat{\beta}_2 - 0}{S(\beta_2)} = 3{,}55$$

$$\frac{\hat{\beta}_3 - 0}{S(\beta_3)} = 8{,}16$$

En el siguiente cuadro se observan los estadísticos y sus probabilidades asociadas:

Parámetro	*Coeficientes*	*Error típico*	*Estadístico t de student*	*P-value*
Age	569.404	97.578	5,835	2,1906E-08
AST	1.106.681	312.070	3,546	0,00048876
PTS	784.897	96.157	8,163	3,9129E-14

Si establecemos un Error de tipo 1 de 5%, vemos que nuestras tres variables explicativas son significativas, cada una con un P-value menor a 0,05, por lo que *rechazamos* la H_0.

Supuestos del modelo

En esta sección se realiza un análisis para buscar violaciones de los supuestos en nuestra regresión o deficiencias del modelo.

Multicolinealidad

Cuando existe una correlación muy fuerte entre variables predictivas, los resultados de la predicción serán ambiguos, especialmente con respecto a los valores de los coeficientes de regresión estimados. Esto constituye lo que se conoce como *Multicolinealidad*.

Se utiliza la matriz de correlación entre variables predictoras:

	Age	AST	PTS
Age	1.0000		
AST	0.2983	1.0000	
PTS	0.2232	0.7370	1.0000

Podemos ver que existe una correlación entre Asistencias y Puntos por partido de 0.74. Por lo que se realiza una prueba para determinar si existe colinealidad.

Utilizamos el método Factor de inflación de la varianza (FIV), que realiza una regresión que tiene a una variable explicativa como una función de las demás variables explicativas. Por ejemplo, para el caso de Edad, sería:

$x_{i1} = \beta_0 + \beta_2\, x_{i2} + \beta_3\, x_{i3} + \varepsilon_i$

y utiliza el R_i^2 de esta regresión auxiliar para formar el (FIV):

$$FIV_i = \frac{1}{1 - R_i^2}$$

Cuanto mayor sea R_i^2, significa que hay un problema de colinealidad, ya que una variable predictora se explica por las demás variables predictoras. A su vez, cuanto mayor es R_i^2 mayor es (FIV). Gujarati (2010) afirma que si el FIV de una variable es superior a 10 (esto sucede si R_i^2 excede de 0.90), se dice que esa variable es muy colineal.

Los resultados del método son:

Variable	VIF	1/VIF
AST	2.28	0.438041
PTS	2.19	0.456863
Age	1.10	0.911019
Mean VIF	1.86	

Observamos que el FIV[3] de cada variable se encuentra en valores pequeños y cercanos a 1, que es el menor valor que puede tener. De esta forma, se descarta la presencia de un problema de colinealidad.

Análisis de residuos

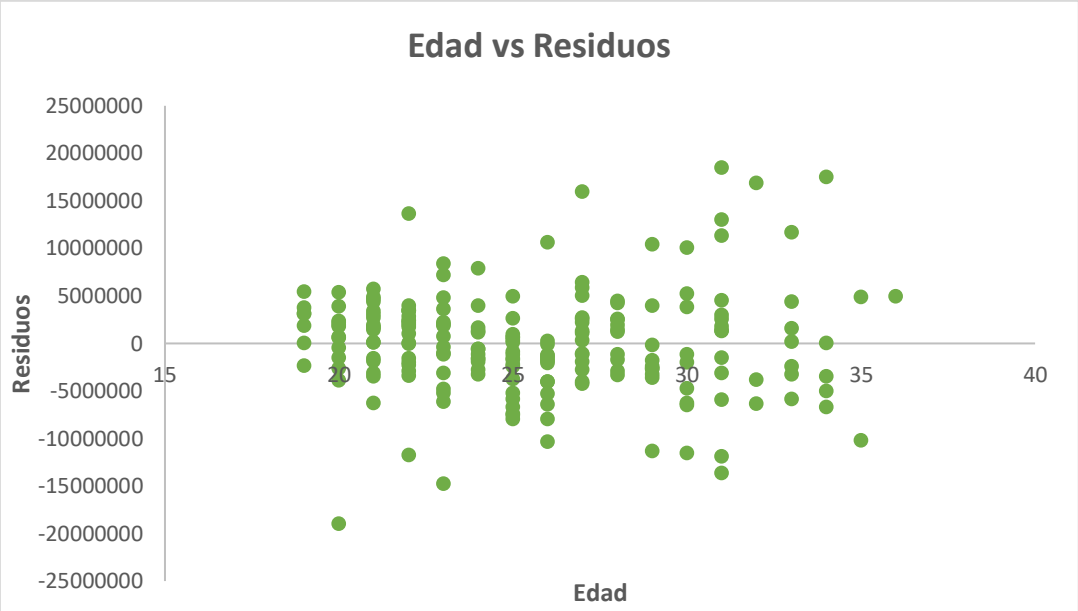
Se puede utilizar gráficos de los residuos contra los correspondientes valores de cada una de las variables de predicción, para verificar si el modelo es lineal o no. A su vez, se puede determinar problemas sobre la varianza del error, graficando los residuos contra los correspondientes valores estimados de la respuesta.

Uno de los supuestos de Gauss-Márkov es la varianza constante de los errores.

$Var(e_i) = \sigma^2$

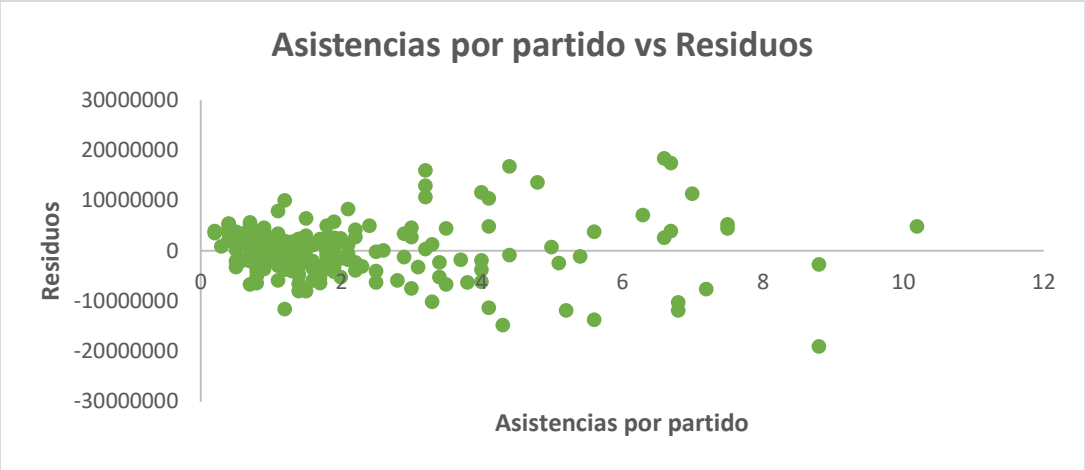
La existencia de *heterocedasticidad*, es decir, la tendencia a aumentar o disminuir los residuos al aumentar los valores estimados de la respuesta, no hace que los estimadores pierdan su propiedad de insesgados, pero ya no son eficientes.

Viendo el grafico de los residuos contra la variable predictora Edad, se puede apreciar una nube de puntos sin clara forma. No parece presentar problemas.

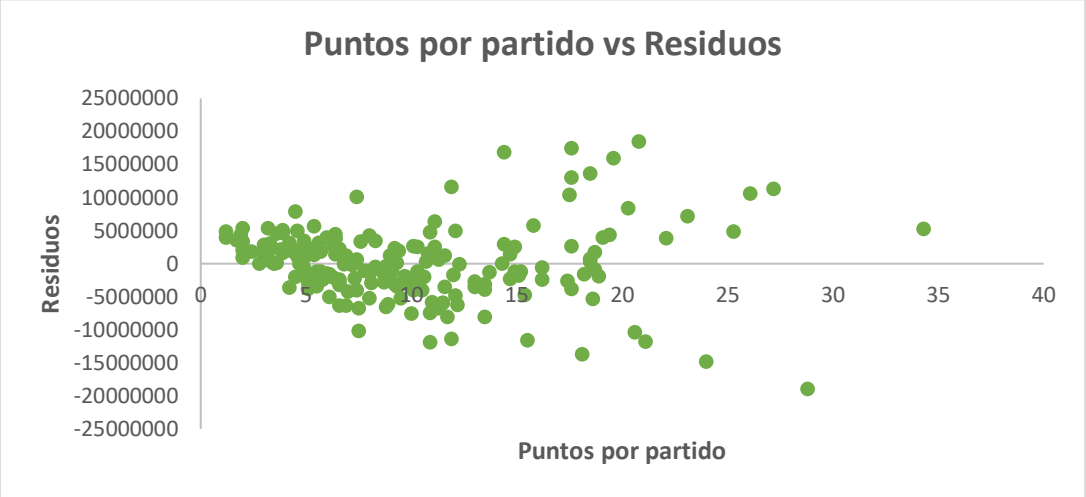


En los residuos contra las asistencias, podemos ver que toma cierta forma, y que crece a medida que aumenta el valor de Asistencias, implicando una varianza no constante.

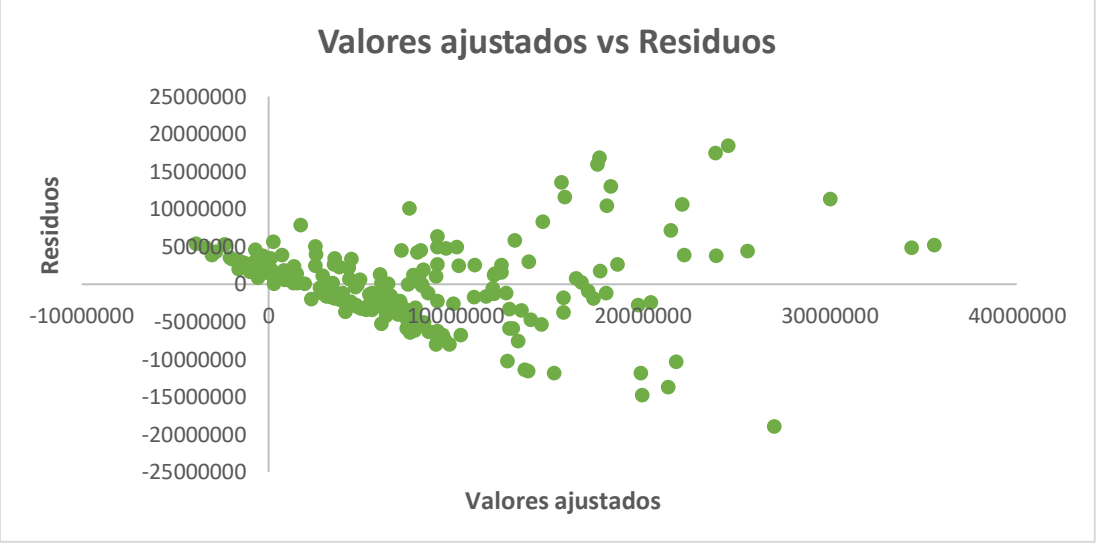
[3] Al mismo tiempo, la columna 1/VIF nos indica (1 - R_i^2).



En la última variable, Puntos por partido, se aprecia una tendencia marcada.



Para comprobar una situación de heterocedasticidad, graficamos los valores esperados de la variable respuesta contra sus residuos.



Los residuos presentan una tendencia, y no vemos una nube de puntos sin forma, como se desearía.

Vamos a realizar una prueba de hipótesis para determinar si existe heterocedasticidad de forma analítica. Utilizaremos el Test de White.

Donde las hipótesis son:

H_0: ${\sigma^2}_{\varepsilon_i} = {\sigma^2}_{\varepsilon}$ (Homocedasticidad)

H_1: ${\sigma^2}_{\varepsilon_i} = C \cdot x_{ij}^2$ (Heterocedasticidad)

Para realizar esta prueba, elevamos los residuos al cuadrado para poder realizar una regresión auxiliar con los $\hat{\varepsilon}_i^2$ como variable respuesta, y como variables predictoras tomamos las variables explicativas que utilizamos en la regresión inicial, sus cuadrados, y los términos cruzados $(x_{i1} \cdot x_{i2}), (x_{i1} \cdot x_{i3}), (x_{i2} \cdot x_{i3})$.

El estadístico de prueba es el siguiente:

$$n \cdot R^2 \sim \chi_p^2$$

Donde:

R^2 : coeficiente de determinación de la regresión auxiliar

p : número de variables explicativas de la regresión auxiliar

El resultado del test es el siguiente:

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(9)      =     66.13
         Prob > chi2  =    0.0000
```

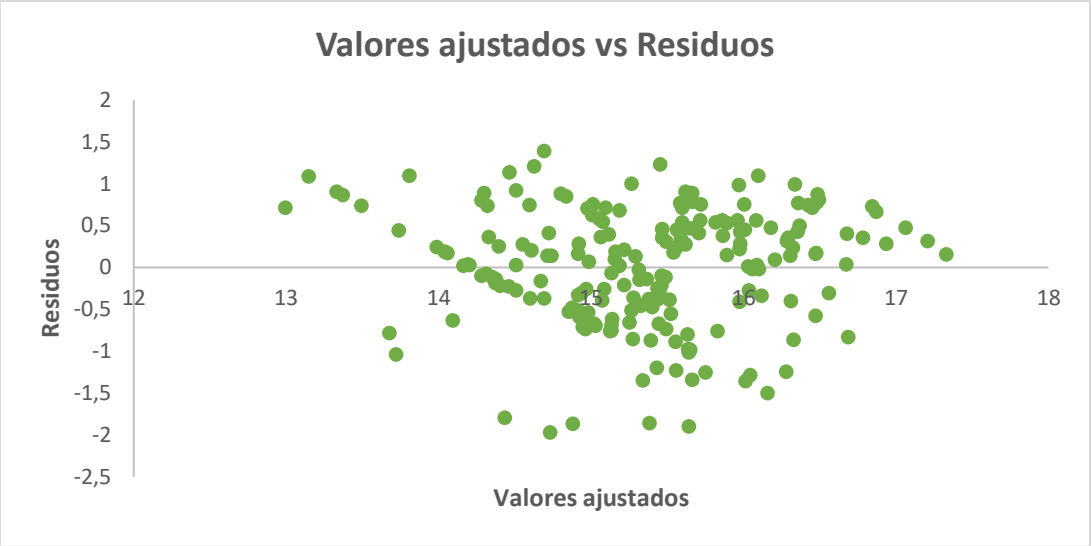
El P-value de la prueba es de $8{,}7 \cdot 10^{-11}$, menor a 0,05 por lo que se *rechaza* la H_0.

Ante la presencia de *Heterocedasticidad*, una posible solución es realizar una transformación logarítmica (modelo Log-Log).

Por lo que nuestra nueva ecuación de regresión es:

$$lny_i = \beta_0 + \beta_1 lnx_{i1} + \beta_2\, lnx_{i2} + \beta_3\, lnx_{i3} + \varepsilon_i$$

Realizamos la estimación de los parámetros y comparamos los valores esperados de la variable respuesta con sus residuos.



Realizamos nuevamente la prueba de White y el resultado es el siguiente:

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(9)      =      7.38
         Prob > chi2  =    0.5980
```

Al ser el P-value > 0,05 *no rechazamos* la H_0 de Homocedasticidad. Vemos que la transformación logarítmica reduce la heterocedasticidad, esto sucede porque comprime las escalas en las cuales se miden las variables.

A su vez, ofrece otra ventaja analítica, al tratarse de un modelo de doble logaritmo, todos los coeficientes de las pendientes son elasticidades.

En nuestra regresión, los coeficientes son los siguientes:

Parámetro	*Coeficientes*
lnAge (β_1)	1,62
lnAST (β_2)	0,33
lnPTS (β_3)	0,75

Por lo que, por ejemplo, podemos interpretar que un incremento del 1% en los Puntos por partido está asociado a un cambio en el Salario de 0,75%.

Referencias

Canavos, G. (1988). *"Probabilidad y Estadística. Aplicaciones y Métodos".*

Gujarati, D. N. (2010). *"Econometría".* 5ta edición.

Link de Excel:

https://drive.google.com/drive/folders/1bodEtvowORNz-IHsyFLGIuMUbMFyh8PA?usp=sharing