Análisis de Regresión

Trabajo Práctico - “Estadística en la NBA”

Grupo - “Los Heats”

NBA - Temporada 2010

Autores: Leguiza, Matias; Sanchez, Joaquin; Trost, Matias; Ziadi, Marcos.

Profesores: Noelia Castellana, Luciana Magnano

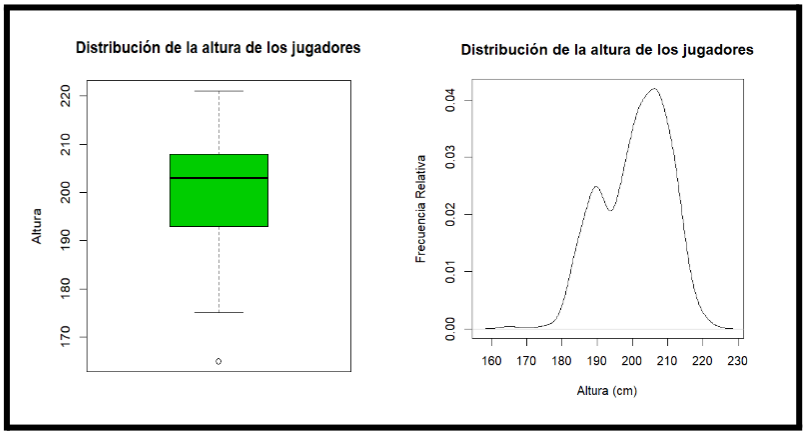
Licenciatura en Estadística.

Universidad Nacional de Rosario

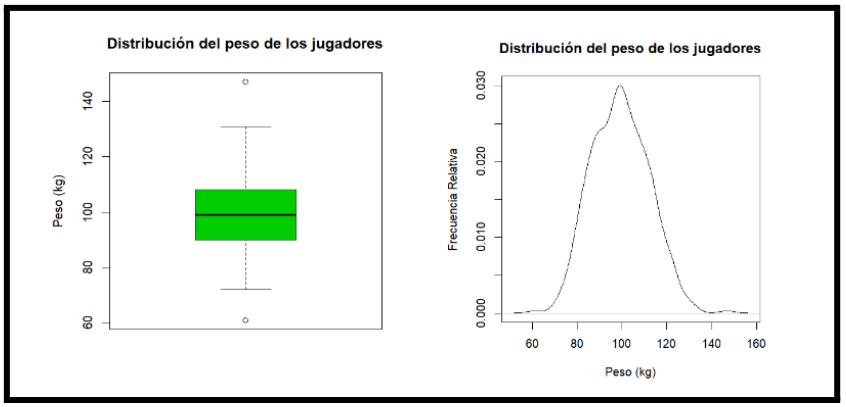
Junio, 2023.

1) ¿Qué características presentan los jugadores estudiados?

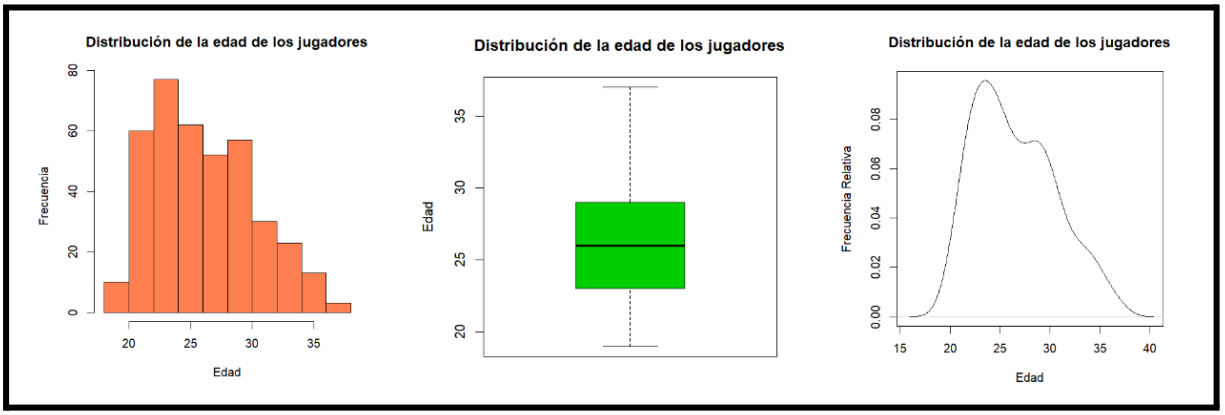
1. Realice un análisis descriptivo univariado de las variables relevadas. ¿Hay valores atípicos? ¿Hay valores perdidos?



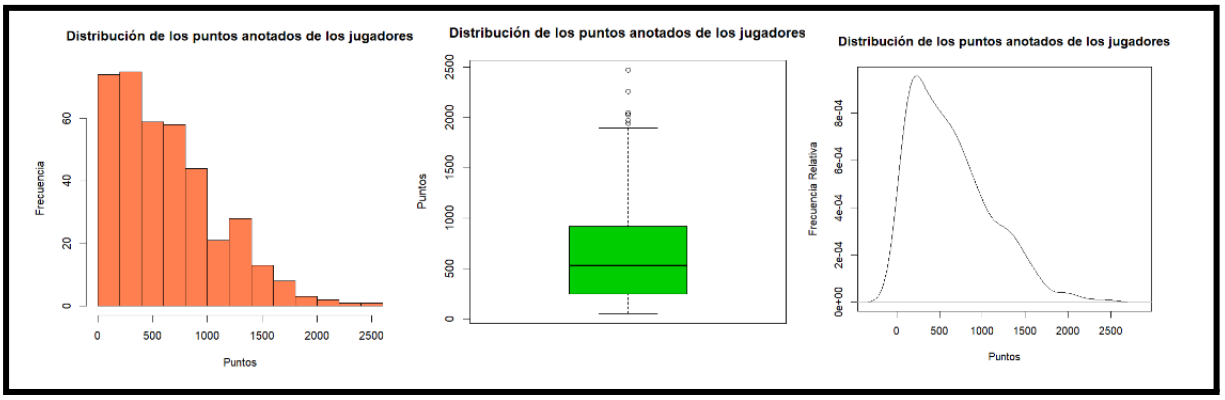
La variable *height* (Altura) tiene media 200.7 y un desvío estándar de 9.38. La distribución es simétrica, centrada en 205 cm, aunque hay bastantes jugadores que miden cerca de 190, siendo 165 cm el mínimo y 221 el máximo.



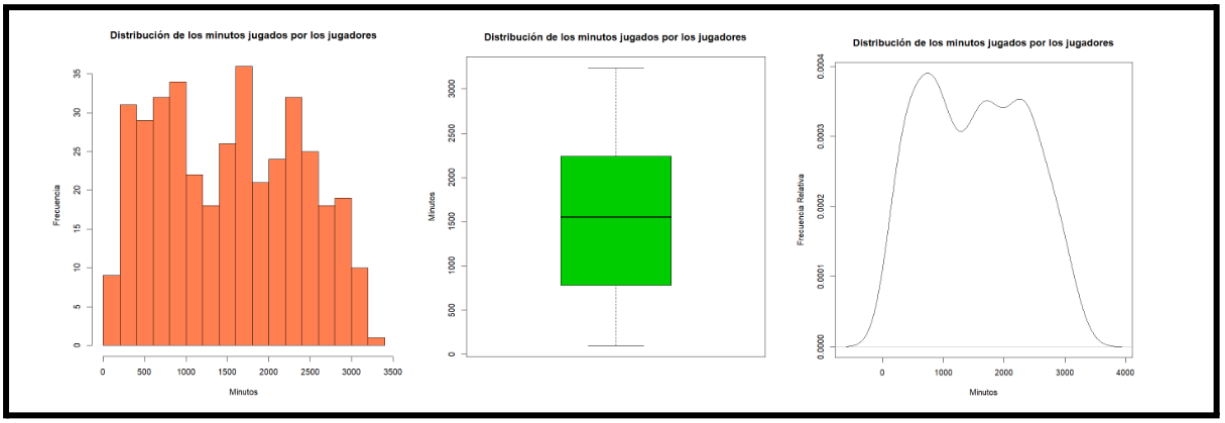
La variable *weight* (Peso) tiene media 99.29 y un desvío estándar de 12.7. Observamos una distribución simétrica, centrada en los 100 kg, con un mínimo de 61 kg y un máximo de 147 kg.



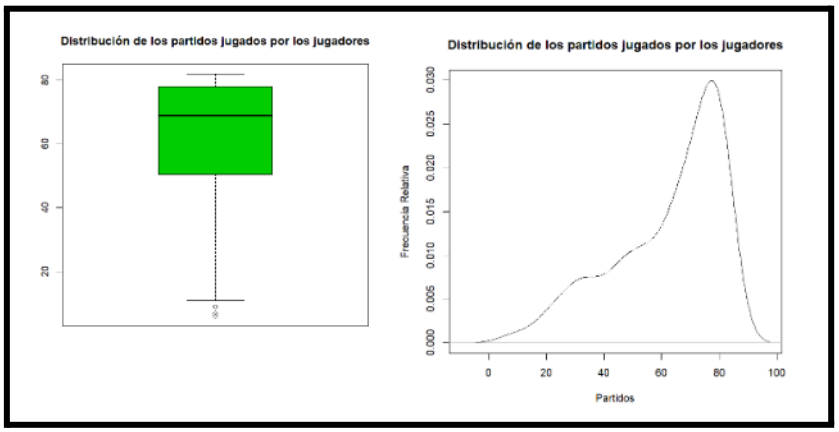
La variable *Age* (Edad) tiene media 26.51 y un desvío estándar de 4.07. Se puede ver que la distribución es simétrica, centrada en 25 años, siendo el mínimo de 19 años y 37 el máximo.



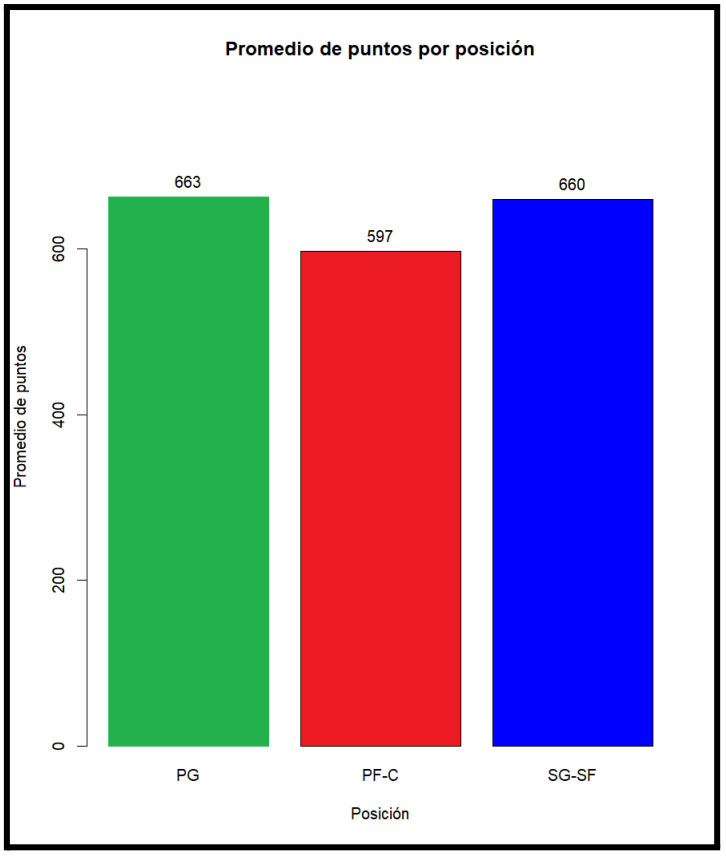
La variable *PTS* (Puntos anotados) tiene media 635.5 y un desvío estándar de 464.55. En el gráfico notamos una distribución asimétrica a la derecha, esto indica que la mayoría de jugadores tienen una cantidad de puntos totales cercana a los 500. La menor cantidad de puntos anotados es 51 y la máxima es 2472.



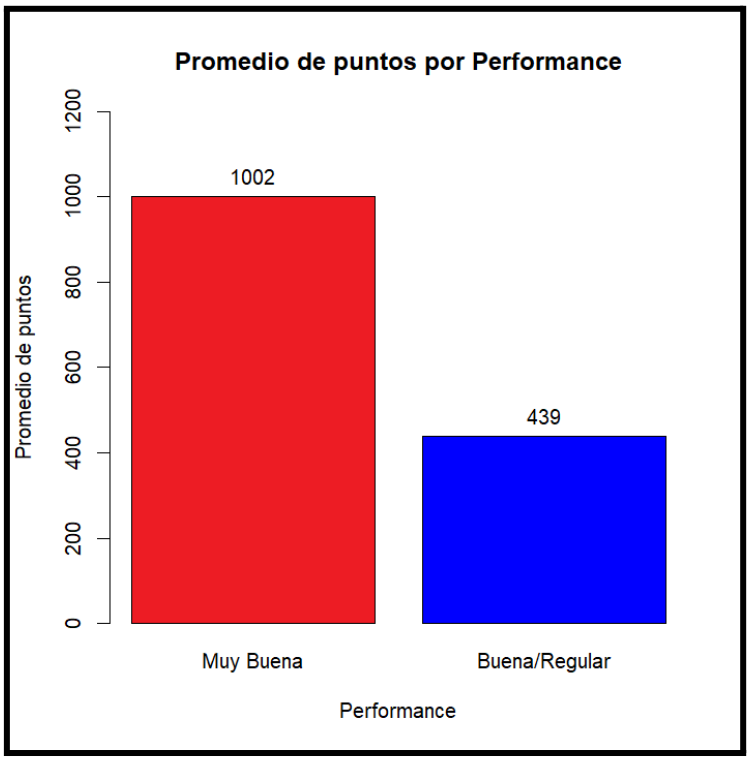
La variable *MP* (Minutos Jugados) tiene media 1522.4 y un desvío estándar de 845.8. En base al gráfico podemos observar que es uniforme entre 1000 y 3000 minutos, es decir, la gran mayoría de los jugadores jugaron entre 1000 y 3000 minutos. El jugador con menos minutos jugados jugó 96, mientras que el que más jugó disputó 3239 minutos.



La variable *G* (Partidos Jugados) tiene media 62.19 y un desvío estándar de 18.85. En el gráfico notamos una distribución asimétrica a la izquierda, esto indica que la mayoría de jugadores tienen una cantidad de partidos jugados cercana a 80. El máximo es de 82 partidos jugados y el mínimo de 6.



Se puede observar en el gráfico de barras que las posiciones SG-SF y PG tienen una cantidad de puntos promedios muy parecidos, mientras que la posición PF-C es la que tiene el menor promedio de puntos anotados.

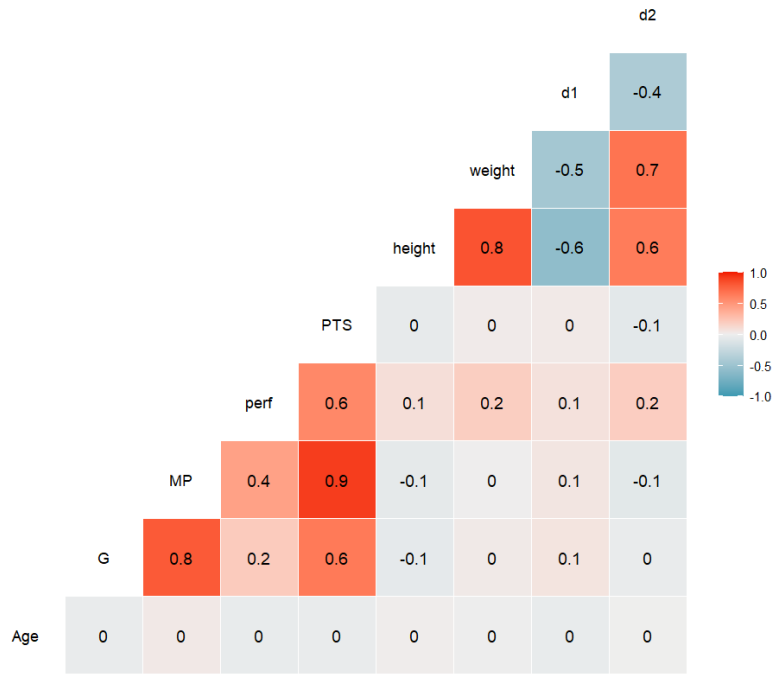


El promedio de puntos anotados por los jugadores que tuvieron una muy buena performance histórica es mucho mayor que el de los jugadores con una performance buena.

En todo el dataframe se encuentran tan solo 5 valores perdidos y estos son hallados en la columna de *perf* (Performance Histórica).

---

2. ¿Qué relación tienen las variables consideradas explicativas con los puntos totales anotadas? ¿Entre ellas presentan alguna relación?



En el gráfico se ve la relación que tienen todas las variables entre sí. Las que son de nuestro interés es la relación de todas las variables con la cantidad de puntos anotados.

Se ve como la edad, la altura y peso no tiene correlación. Por otro lado, se observa que la cantidad de minutos jugados tiene una correlación positiva del 90%, la más fuerte de todas.

Algunas correlaciones interesantes entre variables:

- Altura con Peso: 80%
- Minutos jugados con Partidos jugados: 80%
- Minutos jugados con Performance histórica: 40%

2) Modelo 1

En una primera instancia se desea estudiar la relación entre el total de minutos jugados y los puntos totales anotados. Para ello se decide ajustar un modelo de regresión lineal simple (Modelo 1).

---

1. Escribir la ecuación y los supuestos para el Modelo 1.

Modelo 1

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$$

X_1: $Cantidad\ total\ de\ minutos\ jugados.$

Supuestos:

- $\varepsilon_i \sim N\ \forall i = 1, 2, ...$
- $E(\varepsilon_i) = 0\ \forall i = 1, 2, ...$
- $V(\varepsilon_i) = \sigma^2\ \forall i = 1, 2, ...$
- $Cov(\varepsilon_i, \varepsilon_j) = 0\ \forall i \neq j$

---

2. ¿Existe regresión? Plantear la hipótesis correspondiente, escribir la estadística de prueba y su distribución y concluir en términos del problema.

Test de regresión

$H_0) \beta_1 = 0 \qquad H_1) \beta_1 \neq 0$

- Estadística de prueba:

$$t = \frac{b_1}{s(b_1)} \sim t_{n-2}$$

- Regla de decisión: Rechazo H_0 si $|t_{obs}| > t_{n-2,1-\frac{\alpha}{2}}$

$t_{crítico} = 1.966$

$t_{obs} = 43.478$

$\Rightarrow t_{obs} > t_{crítico}$ ∴ Rechazo H_0

∴ En base a la evidencia muestral y con un nivel de significación del 5% es de esperar que la cantidad total de minutos jugados aporte significativamente a la explicación de la cantidad total de puntos anotados.

---

3. Escribir la ecuación estimada e interpretar los parámetros estimados.

$$\widehat{y} = -\ 126,67\ +\ 0.5 x_1$$

- $\widehat{\beta}_0$ No tiene sentido realizar una interpretación de este coeficiente.
- $\widehat{\beta}_1$: A medida que la cantidad de minutos aumenta en una unidad, la cantidad de puntos anotados aumentará en un 0.5.

---

4. Comentar acerca del valor del coeficiente de determinación.

Coeficiente de correlación

$$R^2 = \frac{SCR_m}{SCT_m} = 1 - \frac{SCE}{SCT_m}$$

$$\Rightarrow R^2 = 0.83$$

El 83% de la variabilidad total de la cantidad de puntos anotados es explicada por la Relación lineal entre la cantidad de minutos jugados y los puntos anotados.

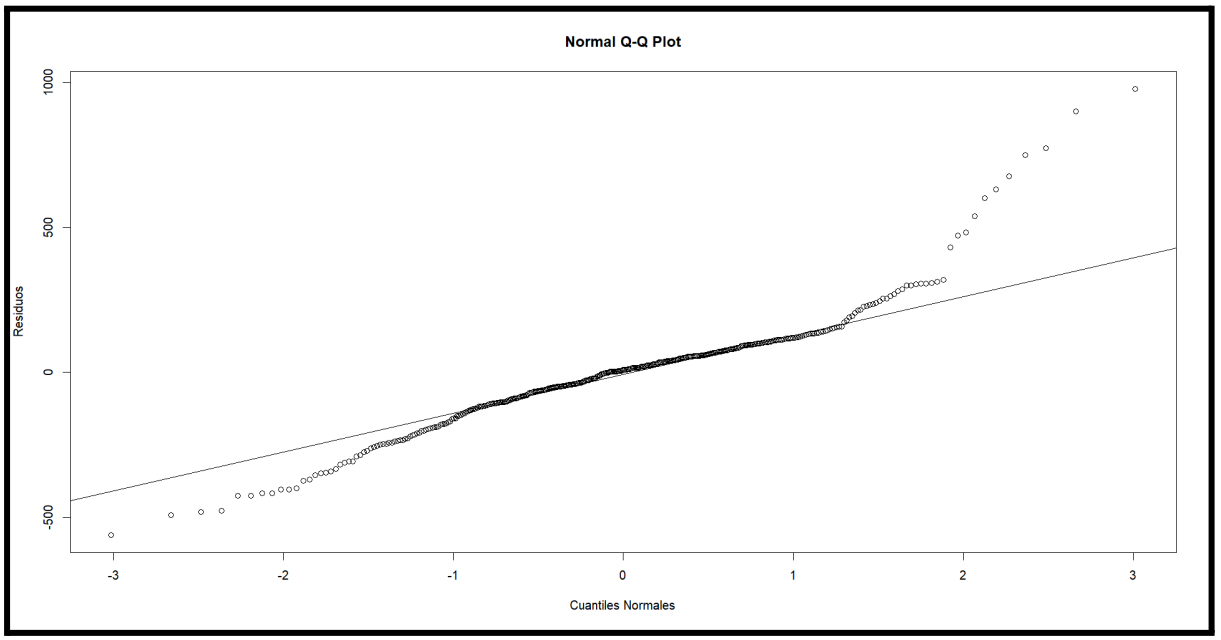
---

5. Evaluar si se cumplen los supuestos del modelo. ¿Hay observaciones atípicas? ¿Corresponde a algún jugador particular?

Análisis de residuos

- **Normalidad**:¿Los errores se distribuyen normalmente?

Grafico probabilistico normal:



Test de bondad:

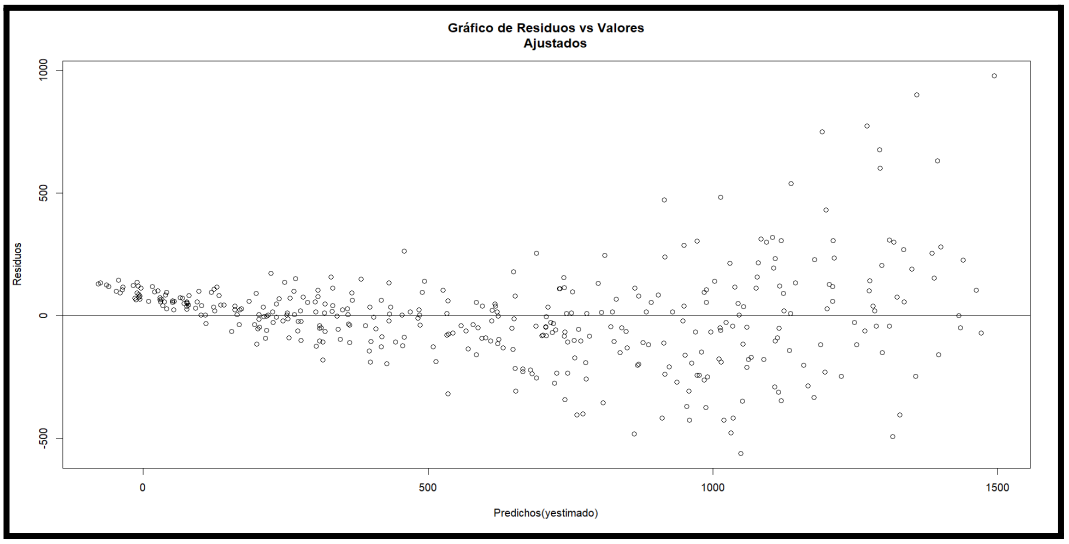
H_0) Los errores se distribuyen normal.

H_1) Los errores no se distribuyen normal.

p_value=0,00000000000000022<0,05

$\therefore$ Mediante el gráfico probabilístico normal y el test de bondad, llegamos a la conclusión de que los errores no están distribuidos normalmente.

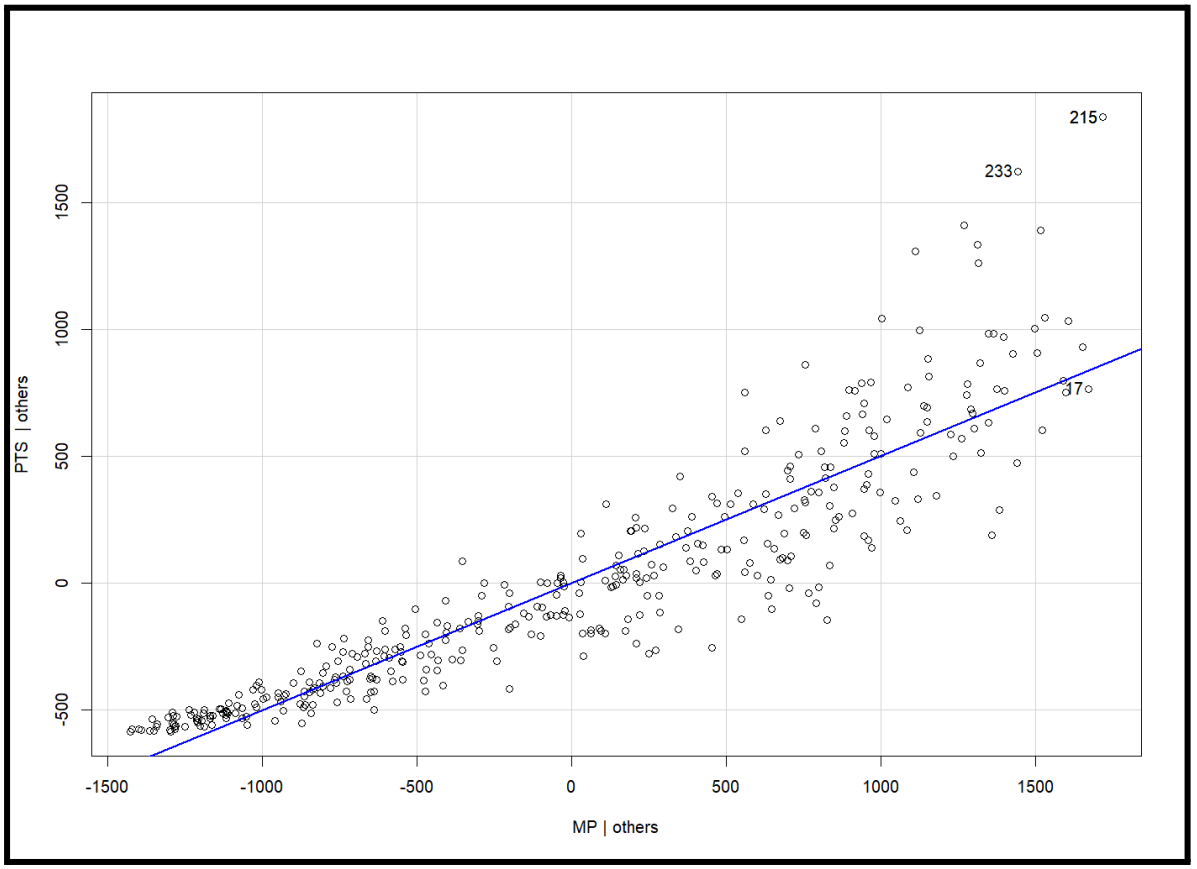
- **Variancia constante**



Realizando un gráfico de residuos vs valores ajustados, vemos que no se puede encerrar a todos los puntos por una banda horizontal, por lo que el supuesto no se cumple.

- **Correlación de los errores**: ¿Los errores están correlacionados?
  No se puede analizar este supuesto ya que el orden en el que fueron recolectados los datos es desconocido.

- **Linealidad del regresor**



En este gráfico de regresión podemos observar que hay una relación lineal positiva entre la variable respuesta y el regresor. Por lo tanto, el supuesto de linealidad se cumple.

En la base de datos se encontraron una gran cantidad de jugadores con observaciones atípicas:

- Haciendo un análisis de residuos internamente estudentizados, los jugadores que presentan outliers son:

	Player	Year	Age	Pos	Tm	G	MP	perf	PTS	height	weight
	Amar'e Stoudemire	2010	27	PF-C	PHO	82	2838	1	1896	208	111
	Carmelo Anthony	2010	25	SG-SF	DEN	69	2634	1	1943	203	108
	Dirk Nowitzki	2010	31	PF-C	DAL	81	3039	1	2027	213	111
	Dwyane Wade	2010	28	SG-SF	MIA	77	2792	1	2045	193	99
	Kevin Durant	2010	21	SG-SF	OKC	82	3239	1	2472	206	108
	Kobe Bryant	2010	31	SG-SF	LAL	73	2835	1	1970	198	96
	LeBron James	2010	25	SG-SF	CLE	76	2966	1	2258	203	113

- Haciendo un análisis de residuos externamente estudentizados, los jugadores que presentan outliers son:

	Player	Year	Age	Pos	Tm	G	MP	perf	PTS	height	weight
	Amar'e Stoudemire	2010	27	PF-C	PHO	82	2838	1	1896	208	111
	Anthony Parker	2010	34	SG-SF	CLE	81	2289	0	595	198	97
	Ben Wallace	2010	35	PF-C	DET	69	1975	1	381	206	108
	Boris Diaw	2010	27	PF-C	CHA	82	2906	0	925	203	113
	Carmelo Anthony	2010	25	SG-SF	DEN	69	2634	1	1943	203	108
	Chris Bosh	2010	25	PF-C	TOR	70	2526	1	1678	211	106
	Chris Duhon	2010	27	PG	NYK	67	2072	0	494	185	83
	Chuck Hayes	2010	26	PF-C	HOU	82	1773	0	358	198	108
	Corey Maggette	2010	30	PF-C	GSW	70	2081	1	1387	198	98
	Danny Granger	2010	26	SG-SF	IND	62	2278	1	1497	206	100
	Dirk Nowitzki	2010	31	PF-C	DAL	81	3039	1	2027	213	111
	Dwyane Wade	2010	28	SG-SF	MIA	77	2792	1	2045	193	99
	Earl Watson	2010	30	PG	IND	79	2322	0	619	185	88
	Jared Jeffries	2010	28	SG-SF	TOT	70	1794	0	372	211	104
	Jason Kidd	2010	36	PG	DAL	80	2881	1	824	193	92
	Kevin Durant	2010	21	SG-SF	OKC	82	3239	1	2472	206	108
	Kobe Bryant	2010	31	SG-SF	LAL	73	2835	1	1970	198	96
	LeBron James	2010	25	SG-SF	CLE	76	2966	1	2258	203	113
	Marcus Camby	2010	35	PF-C	TOT	74	2314	1	556	211	99
	Monta Ellis	2010	24	SG-SF	GSW	64	2647	1	1631	190	83
	Shane Battier	2010	31	SG-SF	HOU	67	2168	0	534	203	99
	Thabo Sefolosha	2010	25	SG-SF	OKC	82	2348	0	489	201	99

- Jugadores que se presentaron con valores extremos:

	Player	Year	Age	Pos	Tm	G	MP	perf	PTS	height	weight
1	Andre Iguodala	2010	26	SG-SF	PHI	82	3193	1	1401	198	97
50	Brook Lopez	2010	21	PF-C	NJN	82	3027	1	1542	213	124
96	David Lee	2010	26	PF-C	NYK	81	3019	1	1640	201	102
112	Dirk Nowitzki	2010	31	PF-C	DAL	81	3039	1	2027	213	111
136	Gerald Wallace	2010	27	SG-SF	CHA	76	3119	1	1386	201	97
181	Jeff Green	2010	23	PF-C	OKC	82	3043	0	1239	196	88
215	Kevin Durant	2010	21	SG-SF	OKC	82	3239	1	2472	206	108
285	O.J. Mayo	2010	22	SG-SF	MEM	82	3113	0	1432	201	99
323	Rudy Gay	2010	23	SG-SF	MEM	80	3175	1	1567	203	104
348	Stephen Jackson	2010	31	SG-SF	TOT	81	3129	1	1667	190	83
385	Zach Randolph	2010	28	PF-C	MEM	81	3051	1	1681	206	117

Hay dos jugadores que son outliers y valores extremos, estos son: Dirk Nowitzki y Kevin Durant, los cuales son jugadores que anotaron muchos puntos y jugaron muchos minutos. Kevin Durant fue el máximo anotador de la NBA en esa temporada.

Los valores extremos se pueden dividir en 2 grupos, el de “jugadores muy anotadores” y “jugadores poco anotadores”, y esto se ve definido porque no alcanzaron una cierta cantidad de puntos por la cantidad de minutos jugados. Muchos de estos jugadores, como Andre Iguodala y Brook Lopez, son buenos jugadores (ambos son campeones de la NBA en la actualidad), pero su función principal no es anotar puntos, ya que son normalmente jugadores defensivos. Algo para destacar es que ninguno de estos jugadores son bases (Point Guards) esto denota como el base tiene una función en el juego en la cual debe anotar bastantes puntos.

En los outliers se ve algo parecido, ya que hay jugadores que cumplen el rol de asistidores o reboteadores (por ejemplo Jason Kidd quedó entre los máximos asistidores de la liga en esa temporada).

En este apartado también se pueden ver los mejores jugadores de la liga como Lebron, Bosh, Durant, Kobe, etc.

3) Modelo 2

---

1. En una segunda instancia se desea estudiar la relación entre todas las variables relevadas y el total de puntos anotados. Para ello se decide ajustar un modelo de regresión lineal múltiple (Modelo 2).

Modelo 2

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 P_{i1} + \beta_7 D_{i1} + \beta_8 D_{i2} + \varepsilon_i$$

- X_1: $Edad\ del\ jugador.$
- X_2: $Cantidad\ total\ de\ partidos\ jugados.$
- X_3: $Cantidad\ total\ de\ minutos\ jugados.$
- X_4: $Altura(cm).$
- X_5: $Peso(kg).$
- P_1: $Performance.$
- D_1: $Posición\ (D_1 = 1\ si\ la\ posición\ es\ Point\ Guard,\ D_1 = 0\ si\ no\ lo\ es)$
- D_2: $Posición\ (D_2 = 1\ si\ la\ posición\ es\ Power\ Forward - Center,\ D_2 = 0\ si\ no\ lo\ es)$

Supuestos:

- $\varepsilon_i \sim N\ \forall i = 1, 2, \ldots$
- $E(\varepsilon_i) = 0\ \forall i = 1, 2, \ldots$
- $V(\varepsilon_i) = \sigma^2\ \forall i = 1, 2, \ldots$
- $Cov(\varepsilon_i, \varepsilon_j) = 0\ \forall i \neq j$

---

2. ¿Existe regresión? Plantee las hipótesis correspondientes, arme el cuadro ANOVA e interprete en términos del problema.

Cuadro anova:

```
Analysis of Variance Table

Response: PTS
           Df   Sum Sq  Mean Sq   F value                Pr(>F)
Age         1    77699    77699    2.9774               0.08526 .
G           1 34666749 34666749 1328.3967 < 0.00000000000000022 ***
MP          1 35967525 35967525 1378.2413 < 0.00000000000000022 ***
height      1    13749    13749    0.5268               0.46839
weight      1    87083    87083    3.3369               0.06854 .
perf        1  1902438  1902438   72.8996 0.0000000000000003502 ***
d1          1   144860   144860    5.5509               0.01899 *
d2          1   163238   163238    6.2551               0.01281 *
Residuals 373  9734063    26097
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Test de regresión

$H_0)\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$

$H_1)\beta_j \neq 0,\ j = \overline{1,8}$

- Estadística de prueba:

$$F = \frac{CMR_m}{CME} \sim F_{p,\, n-(p+1)}$$

- Regla de decisión: Rechazo H_0 si $F_{obs} > F_{8,\, n-(p+1),\, \alpha}$

$F_{crítico} = 2,1224$

$F_{obs} = 349,76$

$\Rightarrow F_{obs} = 349,76 > 2,1224 = F_{crítico}$ ∴ Rechazo H_0

∴ En base a la evidencia muestral y con un nivel de significación del 5% es de esperar que al menos una de las variables aporte significativamente a la explicación de la cantidad total de puntos anotados.

---

3. Estime los coeficientes del modelo y realice los test parciales. ¿Qué sugieren estos test?

Modelo estimado:

$$\widehat{y_i} = 383,67 - 6.71\, x_{i1} - 3.58\, x_{i2} + 0.51\, x_{i3} - 1.28\, x_{i4} + 0.95 x_{i5} + 198.19 p_{i1} - 68.19 d_{i1} - 60.11 d_{i2}$$

- x_1: $Edad\ del\ jugador.$
- x_2: $Cantidad\ total\ de\ partidos\ jugados.$
- x_3: $Cantidad\ total\ de\ minutos\ jugados.$
- x_4: $Altura(cm).$
- x_5: $Peso(kg).$
- p_1: $Performance\ (p_1 = 1\ si\ la\ performance\ fue\ muy\ buena,\ p_1 = 0\ si\ fue\ regular)$
- d_1: $Posición\ (D_1 = 1\ si\ la\ posición\ es\ Point\ Guard,\ D_1 = 0\ si\ no\ lo\ es)$
- d_2: $Posición\ (D_2 = 1\ si\ la\ posición\ es\ Power\ Forward - Center,\ D_2 = 0\ si\ no\ lo\ es)$

Test parciales:

- Hipótesis:

$H_0)\beta_j = 0 \quad j = \overline{1,8}$

$H_1)\beta_j \neq 0 \quad j = \overline{1,8}$

- Estadística de prueba:

$$t = \frac{b_j}{s(b_j)} \sim t_{n-(p+1)} \qquad j = \overline{1,8}$$

- Regla de decisión: Rechazo H_0 si $|t_{obs}| > t_{n-(p+1),1-\frac{\alpha}{2}}$

- Conclusión: Una vez realizados los test parciales, llegamos a la conclusión de que todas las variables son significativas cuando las otras están en el modelo menos el peso y la altura.

4. ¿Existe multicolinealidad?

```
     Age        G       MP   height   weight     perf       d1       d2
1.020251 3.174984 3.875908 3.911289 3.737134 1.556945 1.589002 2.021756
```

Utilizando la función vif() en el modelo, podemos observar que ninguna de las variables está correlacionada con otra, por lo que no hay multicolinealidad.

4) Modelo 3

Según el análisis de multicolinealidad y el resultado de los test parciales plantee un nuevo modelo llamado Modelo 3. Incorporar la variable dicotómica performance. Evaluar incorporar interacciones.

Modelo 3

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 P_{i1} + \beta_5 D_{i1} + \beta_6 D_{i2} + \beta_7 X_{i1} P_{i1} + \beta_8 X_{i1} D_{i1} + \beta_9 X_{i1} D_{i2} + \beta_{10} X_{i2} P_{i1} + \beta_{11} X_{i2} D_{i1} + \beta_{12} X_{i2} D_{i2} + \beta_{13} X_{i3} P_{i1} + \beta_{14} X_{i3} D_{i1} + \beta_{15} X_{i3} D_{i2} + \beta_{16} P_{i1} D_{i1} + \beta_{17} P_{i1} D_{i2} + \varepsilon_i$$

- X_1: $Edad\ del\ jugador.$
- X_2: $Cantidad\ total\ de\ partidos\ jugados.$
- X_3: $Cantidad\ total\ de\ minutos\ jugados.$
- P_1: $Performance\ (P_1 = 1\ si\ la\ performance\ fue\ muy\ buena,\ P_1 = 0\ si\ fue\ regular)$
- D_1: $Posición\ (D_1 = 1\ si\ la\ posición\ es\ Point\ Guard,\ D_1 = 0\ si\ no\ lo\ es)$
- D_2: $Posición\ (D_2 = 1\ si\ la\ posición\ es\ Power\ Forward - Center,\ D_2 = 0\ si\ no\ lo\ es)$

Supuestos:

- $\varepsilon_i \sim N\ \forall i = 1, 2, ...$
- $E(\varepsilon_i) = 0\ \forall i = 1, 2, ...$
- $V(\varepsilon_i) = \sigma^2\ \forall i = 1, 2, ...$
- $Cov(\varepsilon_i, \varepsilon_j) = 0\ \forall i \neq j$

Test de paralelismo:

$H_0) \beta_j = 0,\ j = \overline{7, 17}$ $\qquad$ $H_1) Al\ menos\ un\ \beta_j \neq 0,\ j = \overline{7, 17}$

- Estadística de prueba:

$$F = \frac{R(\beta_7\beta_8\beta_9\beta_{10}\beta_{11}\beta_{12}\beta_{13}\beta_{14}\beta_{15}\beta_{16}\beta_{17}/\beta_0\beta_1\beta_2\beta_3\beta_4\beta_5\beta_6)/11}{CME_{MC}} \sim F_{11,\, n-(p+1)}$$

- Regla de decisión: Rechazo H_0 si $F_{obs} > F_{11,\, n-(p+1),\, \alpha}$

  $F_{crit} = 1.815$

  $F_{obs=} = 9,396$

  $\Rightarrow F_{obs} = 9,396 > 1.815 = F_{crítico}$ $\therefore$ Rechazo H_0

$\therefore$ En base a la evidencia muestral y con un nivel de significación del 5% se llega a la conclusión de que las rectas no son paralelas.

Luego, realizamos test parciales para conocer qué interacciones son significativas:

Test parciales:

- Hipótesis:
  $H_0) \beta_j = 0 \quad j = \overline{7,17}$
  $H_1) \beta_j \neq 0 \quad j = \overline{7,17}$
- Estadística de prueba:
  $t = \frac{b_j}{s(b_j)} \sim t_{n-(p+1)} \qquad j = \overline{1,8}$
- Regla de decisión:
  Rechazo H_0 si $|t_{obs}| > t_{n-(p+1),1-\frac{\alpha}{2}}$
- Conclusión:
  En base a los test parciales llegamos a la conclusión de que las interacciones significativas son: MP:d1 y MP:perf. Por lo que el modelo resulta:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i3} D_{i1} + \beta_5 P_{i1} X_{i3} + \varepsilon_i$$

- X_1: $Edad\ del\ jugador$
- X_2: $Cantidad\ total\ de\ partidos\ jugados$
- X_3: $Cantidad\ total\ de\ minutos\ jugados$
- P_1: $Performance (P_1 = 1\ si\ la\ performance\ fue\ muy\ buena,\ P_1 = 0\ si\ fue\ regular)$
- D_1: $Posición\ (D_1 = 1\ si\ la\ posición\ es\ Point\ Guard,\ D_1 = 0\ si\ no\ lo\ es)$

---

1. Estime el nuevo Modelo 3 y evalúe si se cumplen los supuestos del modelo mediante un análisis de residuos. ¿Hay alguna observación atípica? Realice los gráficos de regresión parcial. Comente.

Modelo estimado:

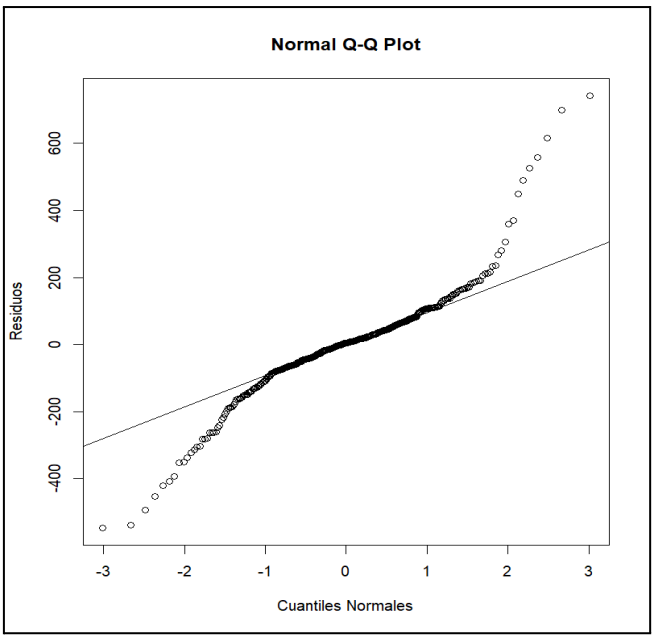
$$y_i = 260,98 - 7,84 x_{i1} - 3,33 x_{i2} + 0,47 x_{i3} - 0,031 x_{i3} d_{i1} + 0,11 x_{i3} p_{i1}$$

- x_1: $Edad\ del\ jugador$
- x_2: $Cantidad\ total\ de\ partidos\ jugados$
- x_3: $Cantidad\ total\ de\ minutos\ jugados$
- p_1: $Performance\ (p_1 = 1\ si\ la\ performance\ fue\ muy\ buena,\ p_1 = 0\ si\ fue\ regular)$
- d_1: $Posición\ (d_1 = 1\ si\ la\ posición\ es\ Point\ Guard,\ d_1 = 0\ si\ no\ lo\ es)$

Análisis de residuos:

- **Normalidad:** ¿Los errores se distribuyen normalmente?

Grafico probabilistico normal:



Test de bondad:

H_0) Los errores se distribuyen normal.

H_1) Los errores no se distribuyen normal.

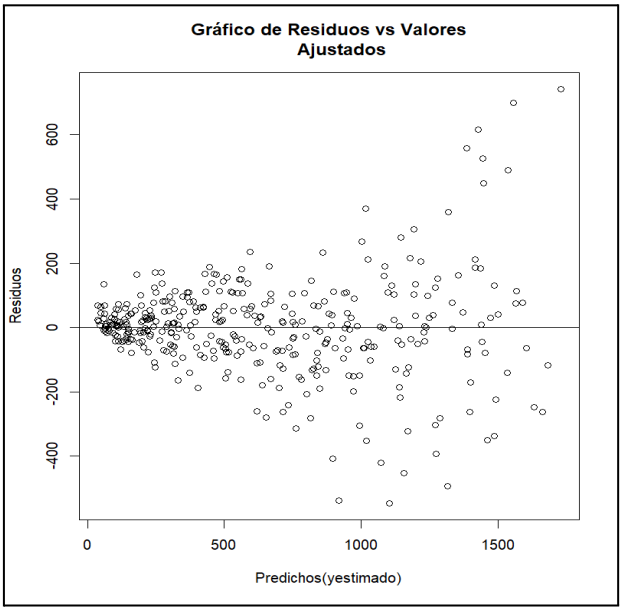
p_value<0,05

∴ Mediante el gráfico probabilístico normal y el test de bondad, llegamos a la conclusión de que los errores no están distribuidos normalmente.

- **Correlación de los errores:** ¿Los errores están correlacionados?

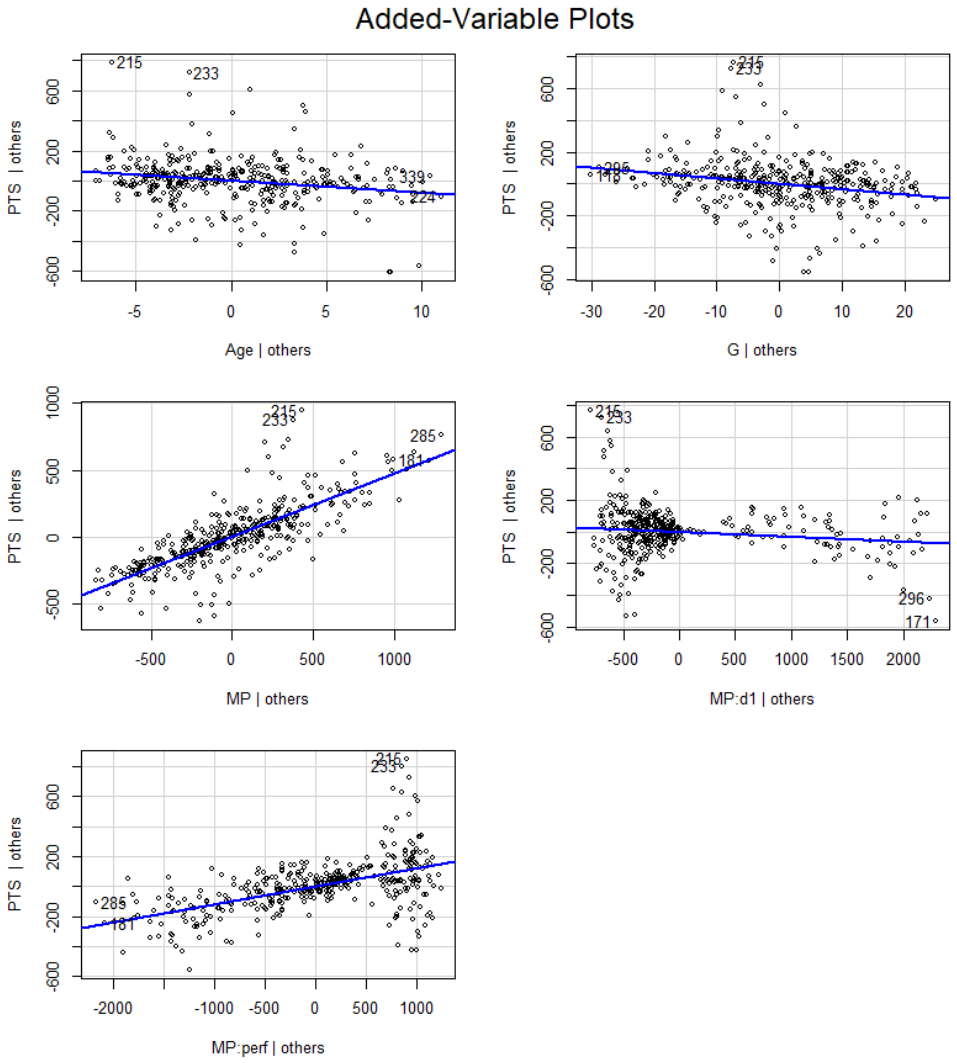
No se puede analizar este supuesto ya que el orden en el que fueron recolectados los datos es desconocido.

- **Variancia constante**



Realizando un gráfico de residuos vs valores ajustados, vemos que no se puede encerrar a todos los puntos por una banda horizontal, por lo que el supuesto no se cumple.

- **Linealidad de los regresores**:



Se puede observar que la edad tiene una leve linealidad negativa con la cantidad de puntos anotados, al igual que la cantidad de partidos.

Hay una linealidad positiva entre la cantidad de minutos jugados y la cantidad de puntos.

<u>Análisis de observaciones atípicas:</u>

En la base de datos se encontraron una gran cantidad de jugadores “outliers”:

- Haciendo un análisis de residuos internamente studentizados, los jugadores que presentan outliers son:

Player	Year	Age	Pos	Tm	G	MP	perf	PTS	height	weight
Ben Wallace	2010	35	PF-C	DET	69	1975	1	381	206	108
Carmelo Anthony	2010	25	SG-SF	DEN	69	2634	1	1943	203	108
Dirk Nowitzki	2010	31	PF-C	DAL	81	3039	1	2027	213	111
Dwyane Wade	2010	28	SG-SF	MIA	77	2792	1	2045	193	99
Jason Kidd	2010	36	PG	DAL	80	2881	1	824	193	92
Kevin Durant	2010	21	SG-SF	OKC	82	3239	1	2472	206	108
Kobe Bryant	2010	31	SG-SF	LAL	73	2835	1	1970	198	96
LeBron James	2010	25	SG-SF	CLE	76	2966	1	2258	203	113
Marcus Camby	2010	35	PF-C	TOT	74	2314	1	556	211	99

- Haciendo un análisis de residuos externamente studentizados, los jugadores que presentan outliers son:

Player	Year	Age	Pos	Tm	G	MP	perf	PTS	height	weight
Al Horford	2010	23	PF-C	ATL	81	2845	1	1148	208	111
Amar'e Stoudemire	2010	27	PF-C	PHO	82	2838	1	1896	208	111
Anderson Varejao	2010	27	PF-C	CLE	76	2166	1	651	208	123
Ben Wallace	2010	35	PF-C	DET	69	1975	1	381	206	108
Brendan Haywood	2010	30	PF-C	TOT	77	2355	1	704	213	121
Carmelo Anthony	2010	25	SG-SF	DEN	69	2634	1	1943	203	108
Chris Andersen	2010	31	PF-C	DEN	76	1695	1	448	208	111
Chris Bosh	2010	25	PF-C	TOR	70	2526	1	1678	211	106
Corey Maggette	2010	30	PF-C	GSW	70	2081	1	1387	198	98
Danny Granger	2010	26	SG-SF	IND	62	2278	1	1497	206	100
Dirk Nowitzki	2010	31	PF-C	DAL	81	3039	1	2027	213	111
Dwyane Wade	2010	28	SG-SF	MIA	77	2792	1	2045	193	99
Emeka Okafor	2010	27	PF-C	NOH	82	2370	1	850	208	114
Jason Kidd	2010	36	PG	DAL	80	2881	1	824	193	92
Joakim Noah	2010	24	PF-C	CHI	64	1925	1	687	211	104
Kevin Durant	2010	21	SG-SF	OKC	82	3239	1	2472	206	108
Kobe Bryant	2010	31	SG-SF	LAL	73	2835	1	1970	198	96
Lamar Odom	2010	30	PF-C	LAL	82	2585	1	882	208	99
LeBron James	2010	25	SG-SF	CLE	76	2966	1	2258	203	113
Marcus Camby	2010	35	PF-C	TOT	74	2314	1	556	211	99
Rajon Rondo	2010	23	PG	BOS	81	2963	1	1110	185	84
Raymond Felton	2010	25	PG	CHA	80	2643	1	968	185	92
Samuel Dalembert	2010	28	PF-C	PHI	82	2124	1	667	211	115
Thabo Sefolosha	2010	25	SG-SF	OKC	82	2348	0	489	201	99

- Jugadores que se presentaron con valores extremos:

Player	Year	Age	Pos	Tm	G	MP	perf	PTS	height	weight
Aaron Brooks	2010	25	PG	HOU	82	2919	1	1604	183	73
Andre Miller	2010	33	PG	POR	82	2500	1	1146	206	102
Baron Davis	2010	30	PG	LAC	75	2523	1	1145	190	94
Brandon Jennings	2010	20	PG	MIL	82	2671	0	1270	185	77
Chauncey Billups	2010	33	PG	DEN	73	2490	1	1427	190	91
Derek Fisher	2010	35	PG	LAL	82	2227	0	615	185	90
Deron Williams	2010	25	PG	UTA	76	2802	1	1419	190	90
Derrick Rose	2010	21	PG	CHI	78	2871	1	1619	190	86
Earl Watson	2010	30	PG	IND	79	2322	0	619	185	88
George Hill	2010	23	PG	SAS	78	2276	0	964	190	85
Grant Hill	2010	37	SG-SF	PHO	81	2430	0	912	203	102
Jason Kidd	2010	36	PG	DAL	80	2881	1	824	193	92
Jeff Green	2010	23	PF-C	OKC	82	3043	0	1239	196	88
Jonny Flynn	2010	20	PG	MIN	81	2339	0	1094	183	83
Mike Conley	2010	22	PG	MEM	80	2569	0	959	185	79
O.J. Mayo	2010	22	SG-SF	MEM	82	3113	0	1432	201	99
Raja Bell	2010	33	SG-SF	TOT	6	180	0	71	196	92
Rajon Rondo	2010	23	PG	BOS	81	2963	1	1110	185	84
Raymond Felton	2010	25	PG	CHA	80	2643	1	968	185	92
Russell Westbrook	2010	21	PG	OKC	82	2813	1	1322	190	90
Stephen Curry	2010	21	PG	GSW	80	2896	1	1399	190	86
Steve Nash	2010	35	PG	PHO	81	2660	1	1333	190	88

El único jugador que es outlier y valor extremo es Jason Kidd qué es un jugador que tiene una baja cantidad de puntos anotados en relación a la cantidad de minutos jugados. Otro punto a tener en cuenta es su edad, ya que es uno de los jugadores más longevos de la temporada.

Entre los valores extremos encontramos jugadores de gran edad (como Grant Hill), jugadores de corta edad (Jonny Flynn), jugadores con muchos partidos jugados (Aaron Brooks, Russel Westbrook), un jugador con muy pocos partidos jugados (Raja Bell), jugadores que disputaron una gran cantidad de minutos (O. J. Mallo) y un jugador con muy pocos minutos en cancha (Raja Bell).

En los outliers encontramos jugadores que anotaron pocos puntos siendo que jugaron muchos partidos y minutos (Ben Wallace), jugadores que anotaron muchos puntos (Kevin Durant), etc.

---

2. Si no se cumplen los supuestos plantear y realizar transformaciones para intentar solucionar el problema.

Recurrimos a una transformación logarítmica:

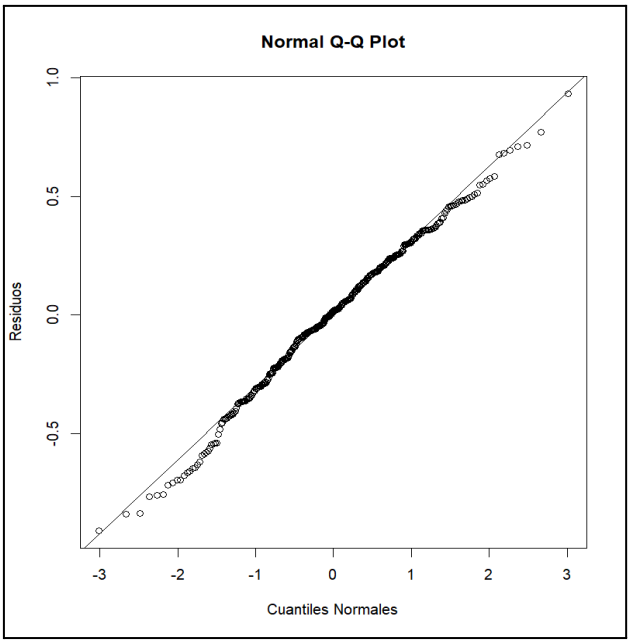
$$ln(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i3} d_{i1} + \beta_5 p_{i1} x_{i3} + \varepsilon_i$$

Ahora realizamos un análisis de residuos de este nuevo modelo:

Análisis de residuos:

- **Normalidad**

Grafico probabilistico normal:



Test de bondad:

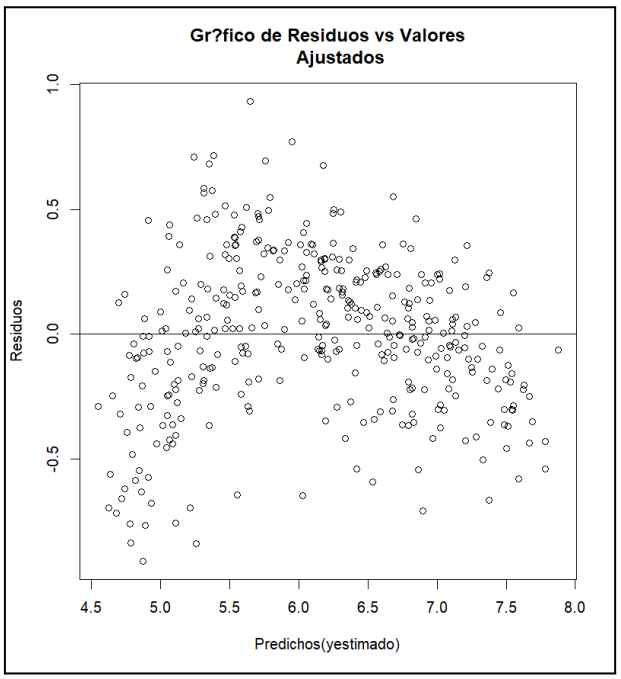
H_0) Los errores se distribuyen normal.

H_1) Los errores no se distribuyen normal.

p_value>0,05

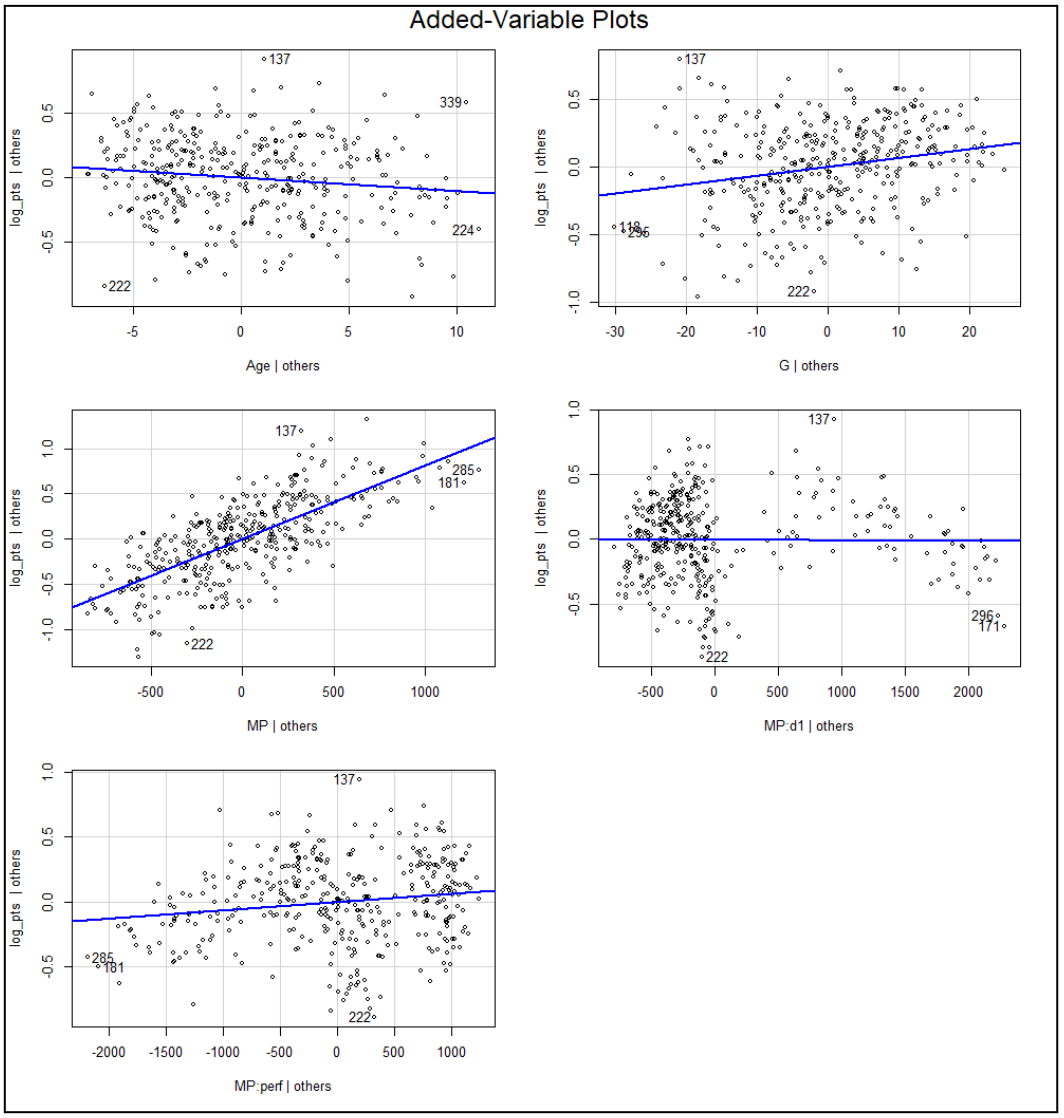
∴ Mediante el gráfico probabilístico normal y el test de bondad, llegamos a la conclusión de que los errores están distribuidos normalmente.

- **Correlación de los errores:** ¿Los errores están correlacionados?
  No se puede analizar este supuesto ya que el orden en el que fueron recolectados los datos es desconocido.
- **Variancia constante**



Realizando un gráfico de residuos vs valores ajustados, vemos que se puede encerrar a todos los puntos por una banda horizontal, por lo que el supuesto se cumple. Los errores tienen variancia constante.

- **Linealidad de los regresores**



Se puede observar que la edad tiene una leve linealidad negativa con la cantidad de puntos anotados.

Hay una linealidad positiva entre la cantidad de minutos jugados y la cantidad de puntos, al igual que la cantidad de partidos.

Todas las variables parecen ajustarse mejor a una recta que en el modelo sin ser transformado, por lo que se cumple el supuesto

5) Modelo final

---

1. Escriba la ecuación estimada del Modelo final en forma general y según la variable performance. Interpretar los coeficientes en términos del problema.

$$ln(y_i) = 4,71 - 0,01x_{i1} + 0,006x_{i2} + 0,0008x_{i3} - 0,000004x_{i3}d_{i1} + 0.00006x_{i3}p_{i1}$$

- Si perf = 0:
  $\Rightarrow ln(y_i) = 4,71 - 0,01x_{i1} + 0,0065x_{i2} + 0,0008x_{i3} - 0,0000047x_{i3}d_{i1}$
- Si perf = 1:
  $\Rightarrow ln(y_i) = 4,71 - 0,01x_{i1} + 0,0065x_{i2} + (0,0008 + 0.00006)x_{i3} - 0,0000047x_{i3}d_{i1}$

- x_1: $Edad\ del\ jugador.$
- x_2: $Cantidad\ total\ de\ partidos\ jugados.$
- x_3: $Cantidad\ total\ de\ minutos\ jugados.$
- p_1: $Performance\ (p_1 = 1\ si\ la\ performance\ fue\ muy\ buena,\ p_1 = 0\ si\ fue\ regular)$
- d_1: $Posición\ (d_1 = 1\ si\ la\ posición\ es\ Point\ Guard,\ d_1 = 0\ si\ no\ lo\ es)$

Interpretación de los coeficientes:

- β_1= -0,01: A medida que la edad aumenta en una unidad, el logaritmo de la cantidad de puntos anotados disminuye en 0,01.
- β_2= 0,0065: A medida que la cantidad de partidos jugados aumenta en una unidad, el logaritmo de la cantidad de puntos anotados aumenta en 0,0065.
- Si perf = 0: β_3= 0,0008: A medida que la cantidad de minutos jugados aumenta en una unidad, el logaritmo de la cantidad de puntos anotados aumenta en 0,0008.
- Si perf = 1: β_3= 0,00086: A medida que la cantidad de minutos jugados aumenta en una unidad, el logaritmo de la cantidad de puntos anotados aumenta en 0,00086.
- β_4= -0,0000047: A medida que la cantidad de minutos jugados, dependiendo la posición del jugador, aumenta en una unidad, el logaritmo de la cantidad de puntos anotados disminuye en 0,0000047.

---

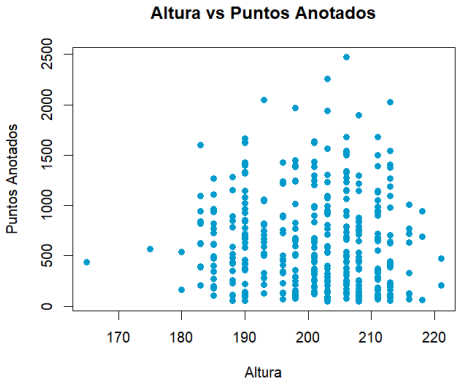
2. Elija un coeficiente e interprete su intervalo de confianza del 95%.

   El intervalo de confianza para la variable "Edad" indica que con un 95% de confianza podemos asegurar que el verdadero coeficiente de regresión de "Edad" se encuentra dentro del intervalo -0.0183 y -0.0026 (Teniendo en cuenta la transformación logarítmica de la variable respuesta). Esto significa que, al nivel de confianza del 95%, por cada aumento unitario en la edad del jugador, se espera que los puntos anotados disminuyan entre -0.0183 y -0.0026.

3. Responda a las preguntas que se realizaron los periodistas según el modelo final estimado.

¿A mayor altura del jugador, mayor cantidad de puntos anotados?

En base a los resultados de nuestro estudio, obtuvimos que la altura no tiende a tener un efecto significativo en la cantidad de puntos que un jugador anota. Por lo que no hay una correlación considerable entre la altura y la cantidad de puntos que un jugador anota.

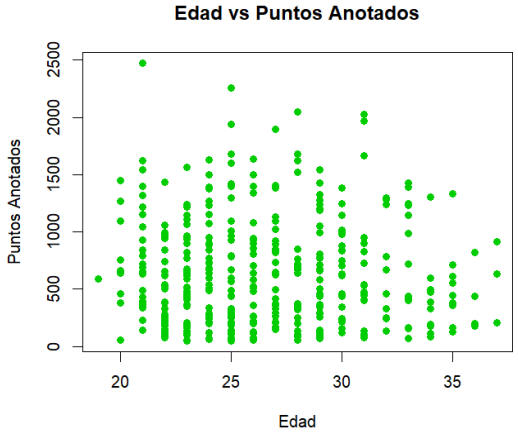


¿La posición de juego influye en los puntos anotados?

La posición de juego influye en los puntos anotados. Los jugadores que juegan en la posición Point Guard, en base a nuestro modelo, mientras más minutos jugados tienen, menos puntos realizan al final de la temporada en comparación a jugadores que juegan otras posiciones y poseen los mismos minutos jugados.

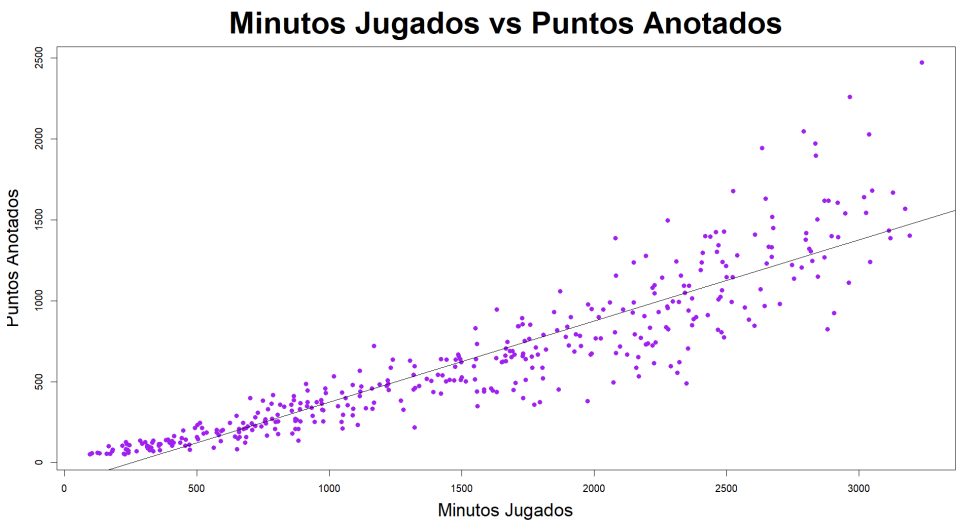
¿Cómo influye la edad en la cantidad de puntos anotados?

Llegamos a la conclusión de que la edad es un factor significativo a la hora de estimar la cantidad de puntos anotados. Pudimos observar que mientras más años tiene un jugador, menos puntos anota.



¿Influye en los puntos anotados la cantidad de minutos que el jugador estuvo en el campo de juego?

En base a los resultados de nuestro análisis, obtuvimos que la cantidad de minutos jugados tiene una correlación positiva con la cantidad de puntos que un jugador anota.



¿La performance histórica del jugador influye en el total de puntos anotados en esa temporada?

Concluimos que la cantidad de minutos que jugó un jugador en base a la performance histórica del jugador tuvo un efecto en la cantidad de puntos que este anotó durante la temporada. Aquellos jugadores que tienen una performance histórica Muy Buena anotaron más puntos que aquellos que no la tuvieron.

---

4. El entrenador del equipo Miami Heat quiere utilizar el modelo para estimar la cantidad de puntos que va a anotar al final de la temporada 2023 un jugador nuevo llamado Nicolas Sánchez. Este jugador mide 1.95mts, pesa 96 kg., su posición de juego es PG (Point Guard), presentó una performance regular en las temporadas previas y se cree que va a jugar 29 partidos con un total de 510 minutos. Presentar la estimación puntual y por intervalo del 95%. (Edad: 22 años)

- Estimación puntual:

En base a todos los datos dados y utilizando nuestro modelo, Nicolas Sánchez se estima que va a anotar un total de 161 puntos durante la temporada 2023.

- Estimación por intervalo del 95%:

Con un nivel de confianza del 95%, se estima que la cantidad de puntos que Nicolas Sánchez anotará durante la temporada 2023 estará dentro del rango (86, 302).

En otras palabras, se espera que Nicolas anote por lo menos 86 puntos y como máximo 302 puntos durante esta temporada.

29 JUNIO 2011

La Capital

EDITION #1

29 JUNIO 2011

NBA | TEMPORADA 2010 CURIOSIDADES

En la NBA, las estadísticas juegan un papel crucial tanto para los equipos como para los aficionados. Desde evaluar el rendimiento de los jugadores hasta analizar estrategias y tomar decisiones informadas, las estadísticas proporcionan una visión detallada del desempeño individual y colectivo en diversos aspectos del juego. En este artículo, exploraremos los resultados de un análisis estadístico, centrándonos en los factores que influyeron en la cantidad de puntos anotados por los jugadores durante la temporada de 2010.

¿Influye en los puntos anotados la altura de los jugadores?

Después de un analisis cuidadoso de los datos, tomando en cuenta la altura, el peso, cantidad de partidos jugados, cantidad de minutos jugados, posición y el desempeño histórico de multiples jugadores, se llego a la conclusión de que la altura no es tan importante como se podría pensar. En otras palabras, que un jugador sea más alto no implica que vaya a anotar más puntos. Hay otros factores más influyentes en la cantidad de puntos que un jugador anota.

¿Y la posición de los jugadores?

Examinando la relación entre la posición del jugador y la cantidad de puntos anotados, se obtuvo que, en general, la posición del jugador es un factor importante para predecir la cantidad de puntos anotados: los Point Guard tienden a anotar puntos con menos frecuencia que aquellos que juegan en otras posiciones.

¿Qué tal la edad?

En cuanto a la edad del jugador, se llego a un resultado lógico. Se descubrió que la edad es un factor extremadamente relevante al predecir la cantidad de puntos anotados. Este hallazgo nos indica que el factor de la edad juega un papel clave en el desempeño de los jugadores, parece que no solo se trata de habilidades y talento, sino también de la capacidad física y resistencia, las cuales parecen verse altamente afectadas a medida que los jugadores se hacen mayores.

¿Y que se puede decir de la cantidad de minutos que el jugador estuvo en campo?

La cantidad de minutos es un factor altamente relevante al predecir la cantidad de puntos anotados. Esta relación tiene lógica si consideramos que cuanto más tiempo un jugador está en el campo, más oportunidades tendrá para crear jugadas ofensivas, para encontrar espacios, recibir pases y anotar puntos, etc. Además, los jugadores que reciben más minutos suelen ser considerados piezas clave del equipo. Por lo tanto, no es sorprendente que haya una fuerte correlación entre la cantidad de minutos y la cantidad de puntos anotados por un jugador en esta temporada de la NBA.

Performance Histórica: ¿Algo que vale la pena tener en cuenta?

Se descubrió que el rendimiento individual se convierte en un factor clave cuando se combina con la cantidad de minutos que el jugador ha jugado. Si bien el desempeño individual es relevante, su impacto en la cantidad de puntos anotados se maximiza cuando se le asigna una mayor participación en los partidos.

¿Hubieron jugadores únicos esta temporada?

Sí, Dirk Nowitzki y Kevin Durant. Ambos son jugadores que anotaron una inmensa cantidad de puntos y jugaron drasticamente más minutos que el jugador promedio. Kevin Durant fue el máximo anotador de la temporada, anotando 2472 puntos.

Jugadores como Andre Iguodala o Brook Lopez tienen una cantidad de minutos jugados muy similares a los de Durant, pero no tienen los mismos puntos anotados. Esto muy probablemente sea debido a que tienen una labor mucho mas defensiva en el campo de juego.

Tambien se vieron a jugadores como Jason Kidd, que tiene una gran cantidad de minutos jugados, pero pocos puntos anotados. En el caso de Kidd, aunque no destaque como anotador, si lo hace como asistidor, termino siendo uno de los maximos asistidores de la temporada, promediando 9.05 asistencias por partido.