Análisis del Desempeño de los equipos de la NFL en 1976

Carlos Enrique Ponce Villagran

Facultad de Ciencias Físico-Matemáticas

25 de noviembre de 2019

Contexto del Problema

El modelo de regresión considera las regresoras $x_2 :=$*Yardas por aire (temporada)*, $x_7 :=$*Porcentaje de Carreras (jugadas por tierra/jugadas totales)* y $x_8 :=$*Yardas por tierra del contrario (temporada)* y se desea saber si existe una relación con $Y :=$*Juegos ganados (por temporada de 14 juegos)*

Datos del Problema

n	y	x_2	x_7	x_8	n	y	x_2	x_7	x_8
1	10	1985	59.7	2205	15	6	2140	59.2	1901
2	11	2855	55	2096	16	5	1730	54.4	2288
3	11	1737	65.6	1847	17	5	2072	49.6	2062
4	13	2905	61.4	1903	18	5	2929	54.3	2861
5	10	1666	66.1	1457	19	6	2268	58.7	2411
6	11	2927	61	1848	20	4	1982	51.7	2289
7	10	2341	66.1	1564	21	3	1792	61.9	2203
8	11	2737	58	1821	22	3	1606	52.7	2592
9	4	1414	57	2577	23	4	1492	57.8	2053
10	2	1838	58.9	2476	24	10	2835	59.7	1979
11	7	1480	67.5	1984	25	6	2416	54.9	2048
12	10	2191	57.2	1917	26	8	1638	65.3	1786
13	9	2229	58.8	1761	27	2	2649	43.8	2876
14	9	2204	58.6	1709	28	0	1503	53.5	2560

Gráfica de Probabilidad Normal

Para el análisis gráfico de los residuales se consideraros los Residuales (e_i), Residuales Estandarizados (d_i) y los Residuales Estudentizados (r_i) y se obtuvieron los siguientes datos:

Gráfica de Probabilidad Normal

n	e_i	d_i	r_i
1	3.70382102	2.23070825	2.45276788
2	1.9644307	1.22727746	1.24101022
3	2.72657876	1.70079039	1.7754083
4	1.60858433	1.02818774	1.02946799
5	0.01213948	0.00791378	0.00774716
6	-0.6566606	-0.419395	-0.41207744
7	-1.90514129	-1.2073225	-1.21951571
8	0.48350328	0.30156663	0.29577807
9	2.07339399	1.33702894	1.36052785
10	-2.30942598	-1.44344412	-1.47869499
11	-0.06085327	-0.04022722	-0.03938156
12	2.0670156	1.25384078	1.26973229
13	-0.13114062	-0.08052695	-0.07884211
14	-0.25174324	-0.15659696	-0.15337819

Gráfica de Probabilidad Normal

n	e_i	d_i	r_i
15	-2.21647025	-1.33309262	-1.35620152
16	1.05513548	0.64790941	0.63988855
17	-0.3248969	-0.22151142	-0.21706951
18	-0.49235925	-0.3699733	-0.36322079
19	-0.13114518	-0.0813189	-0.0796177
20	-0.32068925	-0.19956457	-0.19552503
21	-3.04006094	-1.87132777	-1.98224182
22	1.29315427	0.81920998	0.8134148
23	-0.87961837	-0.5477576	-0.53960816
24	-0.44260879	-0.27704613	-0.27164765
25	-1.66544826	-1.01497424	-1.01564064
26	-0.15120631	-0.09454439	-0.092571
27	-0.36261807	-0.25758302	-0.25250889
28	-1.64567034	-1.04665174	-1.04883101

Gráfica de Probabilidad Normal

Para los residuales se obtuvo la siguiente gráfica:

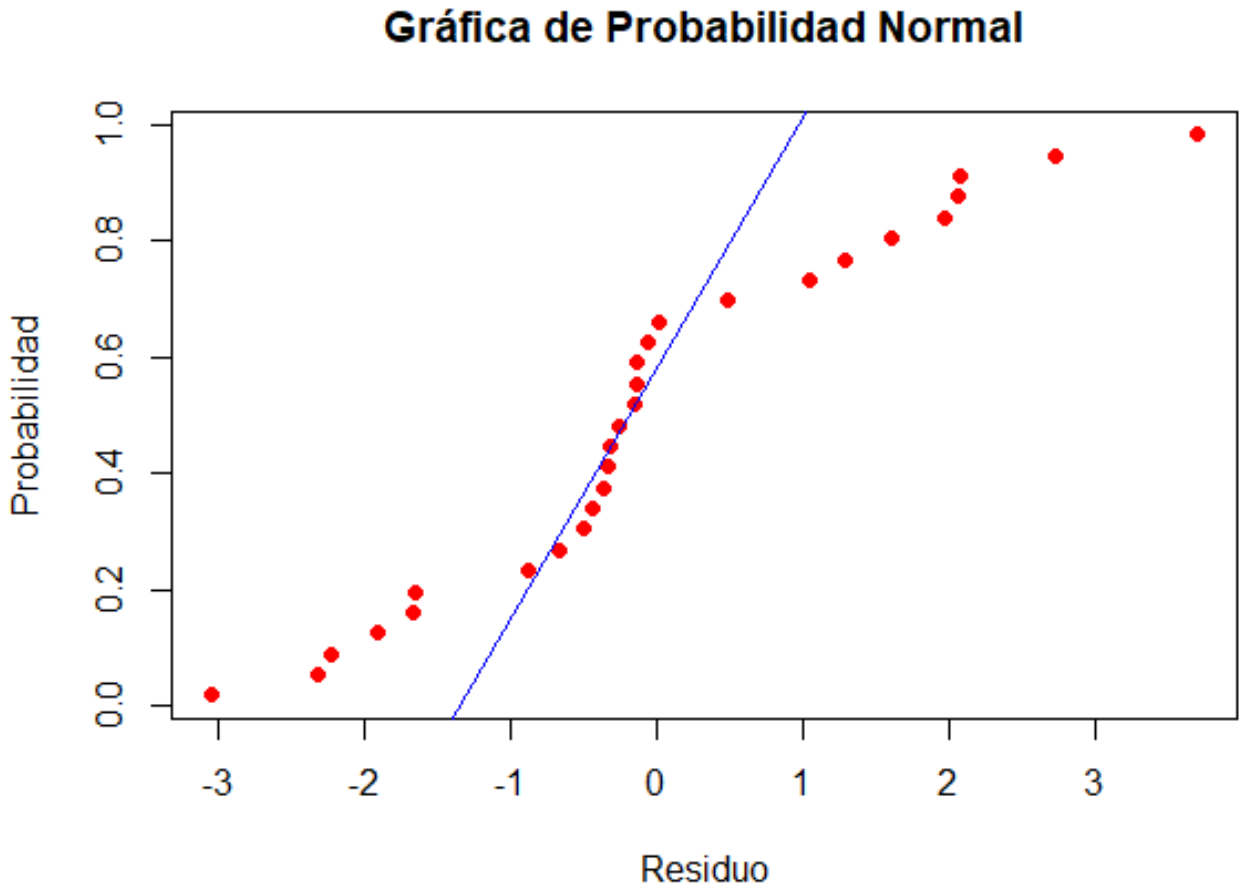


Figura: Gráfica de los Residuales

Gráfica de Probabilidad Normal

Para los residuales estandarizados:

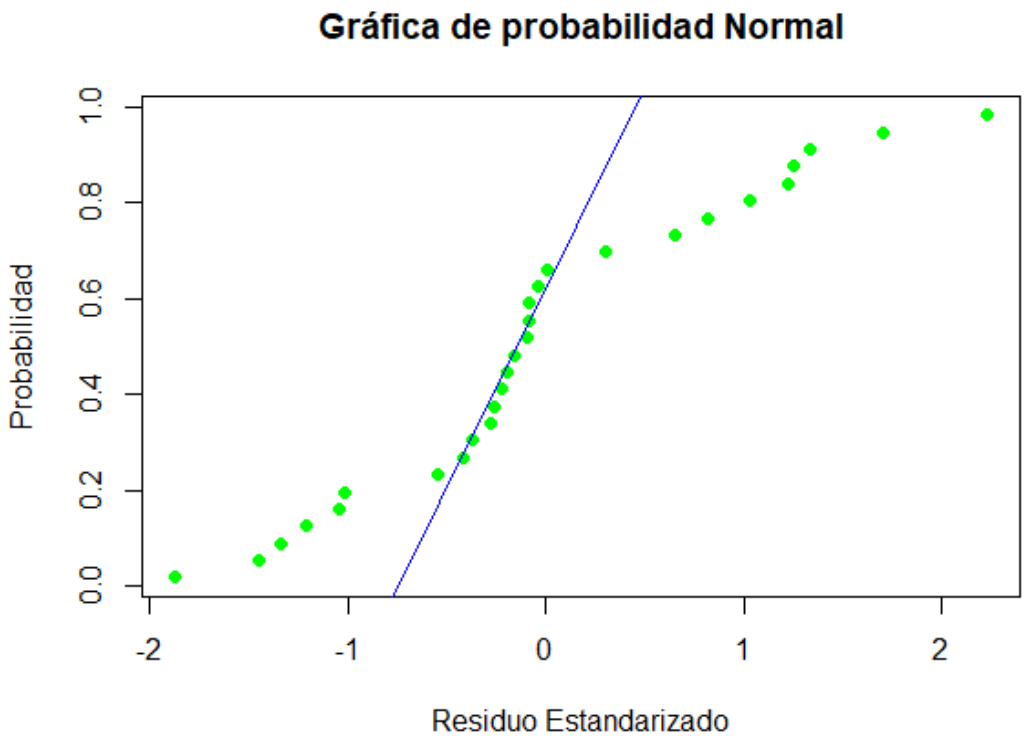


Figura: Gráfica de los Residuales Estandarizados

Gráfica de Probabilidad Normal

Para los residuales estudentizados:

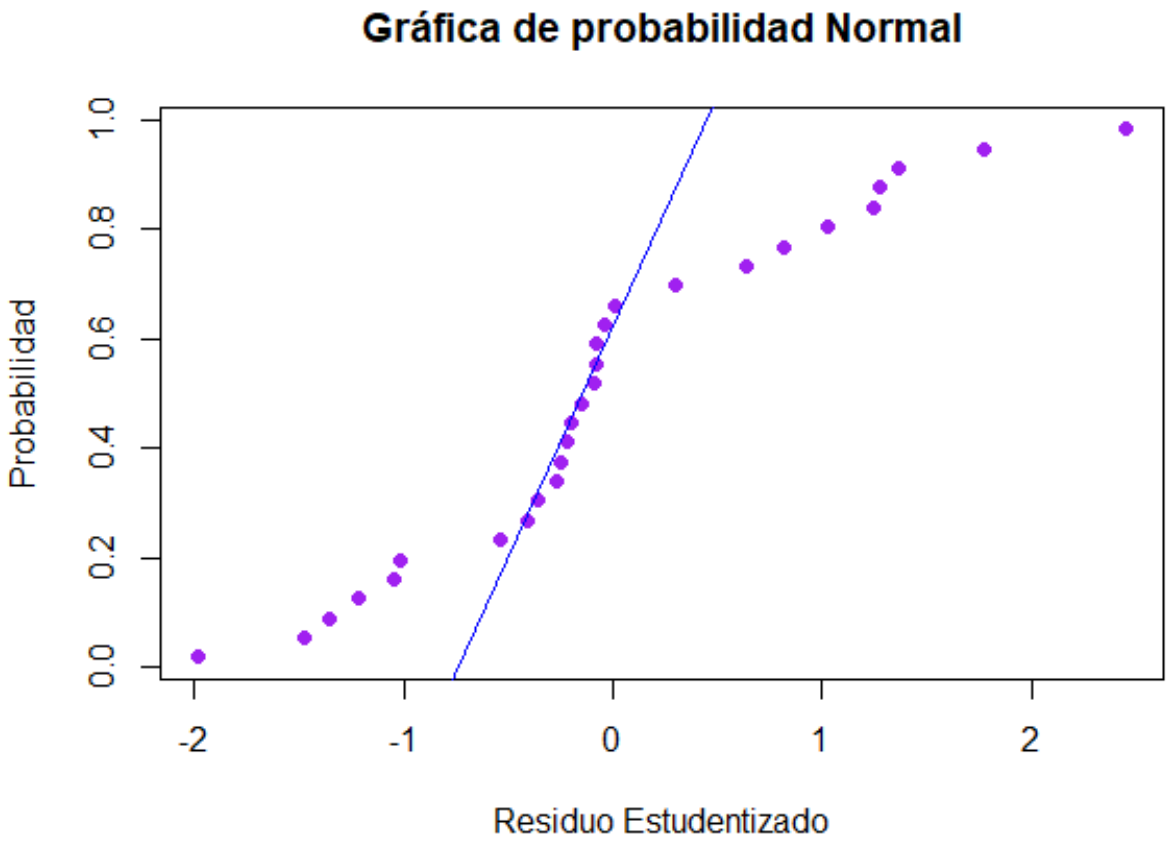
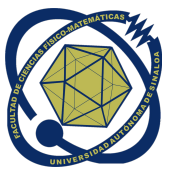


Figura: Gráfica de los Residuales Estudentizados

Como podemos notar de las 3 gráficas los residuales no siguen del todo la forma de una linea, más que nada tal parece que siguen la forma de una distribución con colas delgadas por lo que podría haber más de un punto atípico y apoyándonos de los residuales calculados podemos notar que los puntos a investigar serian e_1 y e_{21}.



Gráfica de residuales en función de los valores ajustados $\hat{y}_i$

Continuando nuestro análisis para encontrar inadecuaciones del modelo continuaremos con las gráficas de Residuales vs. Valores ajustados, con los tres residuales con los que hemos estado trabajando:

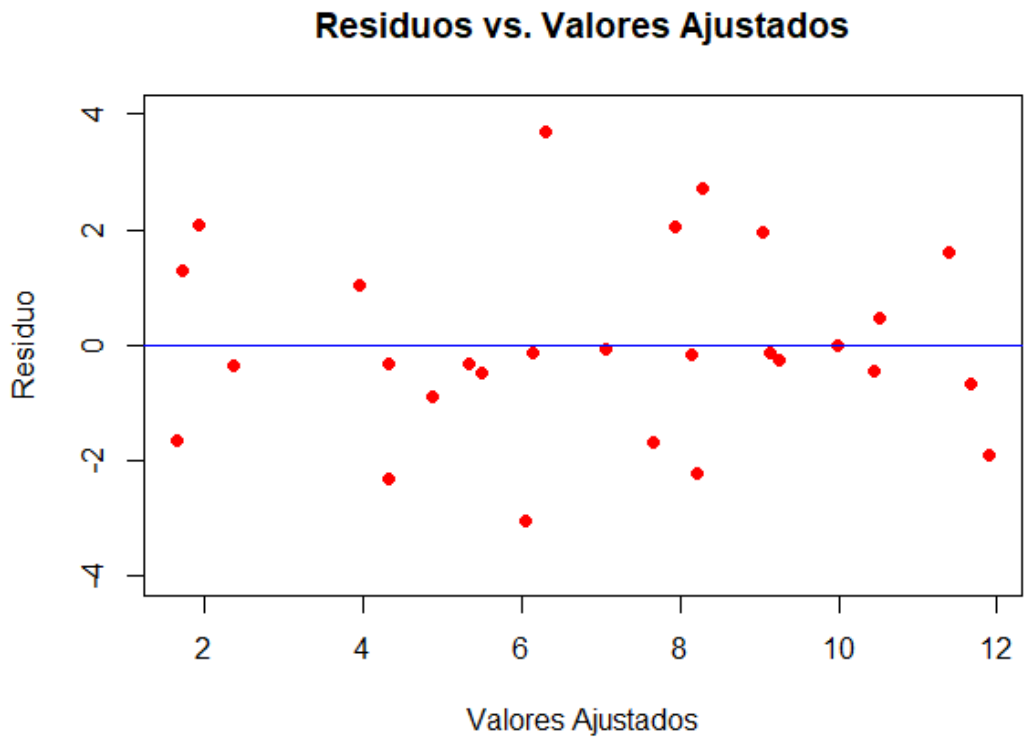


Figura: Gráfica e_i vs. $\hat{y}_i$

Gráfica de residuales en función de los valores ajustados $\hat{y}_i$

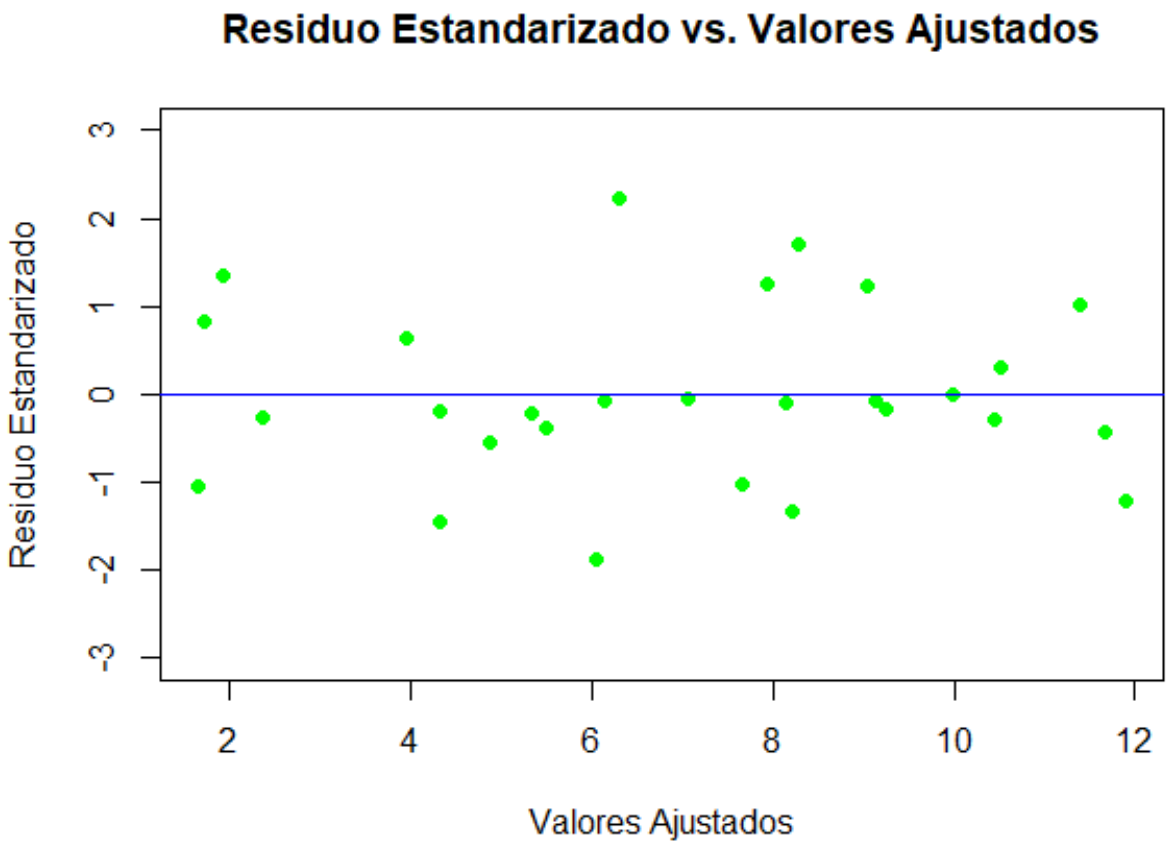


Figura: Gráfica d_i vs. $\hat{y}_i$

Gráfica de residuales en función de los valores ajustados $\hat{y}_i$

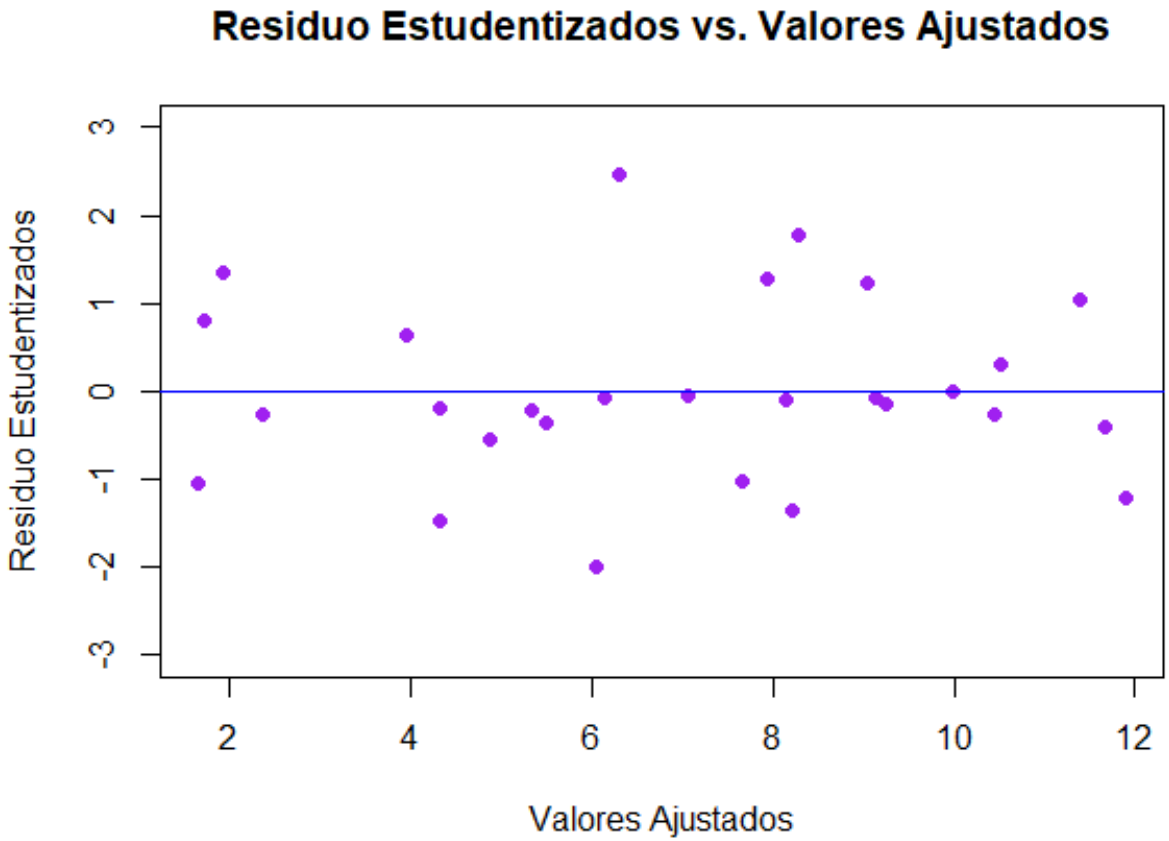


Figura: Gráfica r_i vs. $\hat{y}_i$

De las tres gráficas podemos notar que el punto que mas destaca o más sospechoso en los tres casos es el correspondiente a e_1 y e_{21}, puntos que ya habíamos mencionado. Fuera de esto el modelo parece predecir bastante bien los juegos ganados con los datos que tenemos, lo que quiere decir que no hay defectos aparentes en este.

Gráfica de residuales en función del Regresor

Para el análisis de estas gráficas solo tomaremos en cuenta a los Residuales Estudentizados:

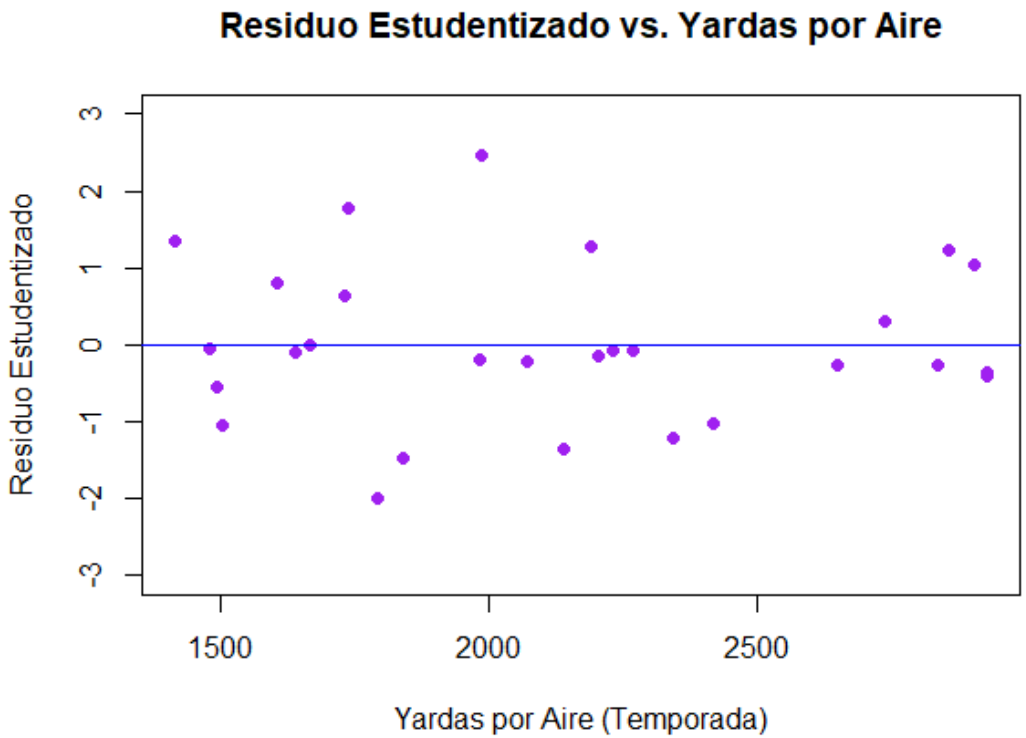


Figura: Gráfica r_i vs. Yardas por aire

Gráfica de residuales en función del Regresor

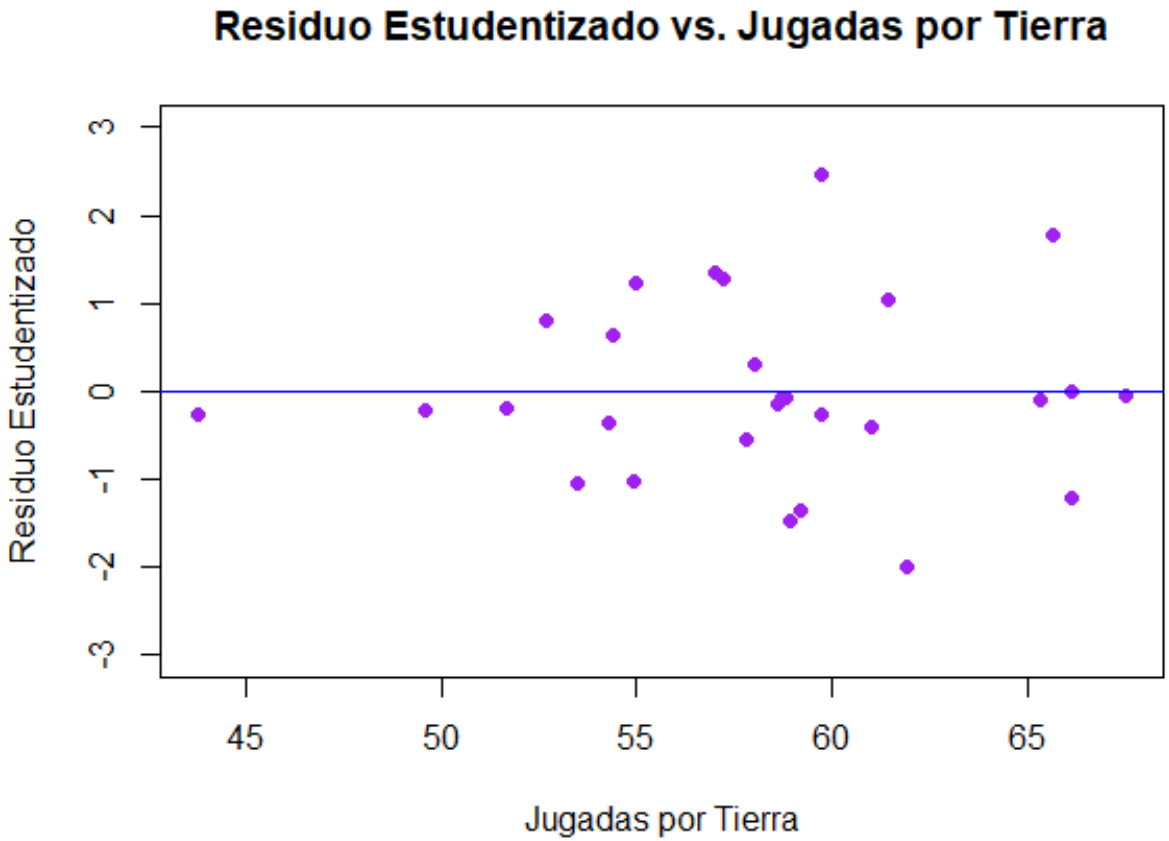


Figura: Gráfica r_i vs. Jugadas por Tierra

Gráfica de residuales en función del Regresor

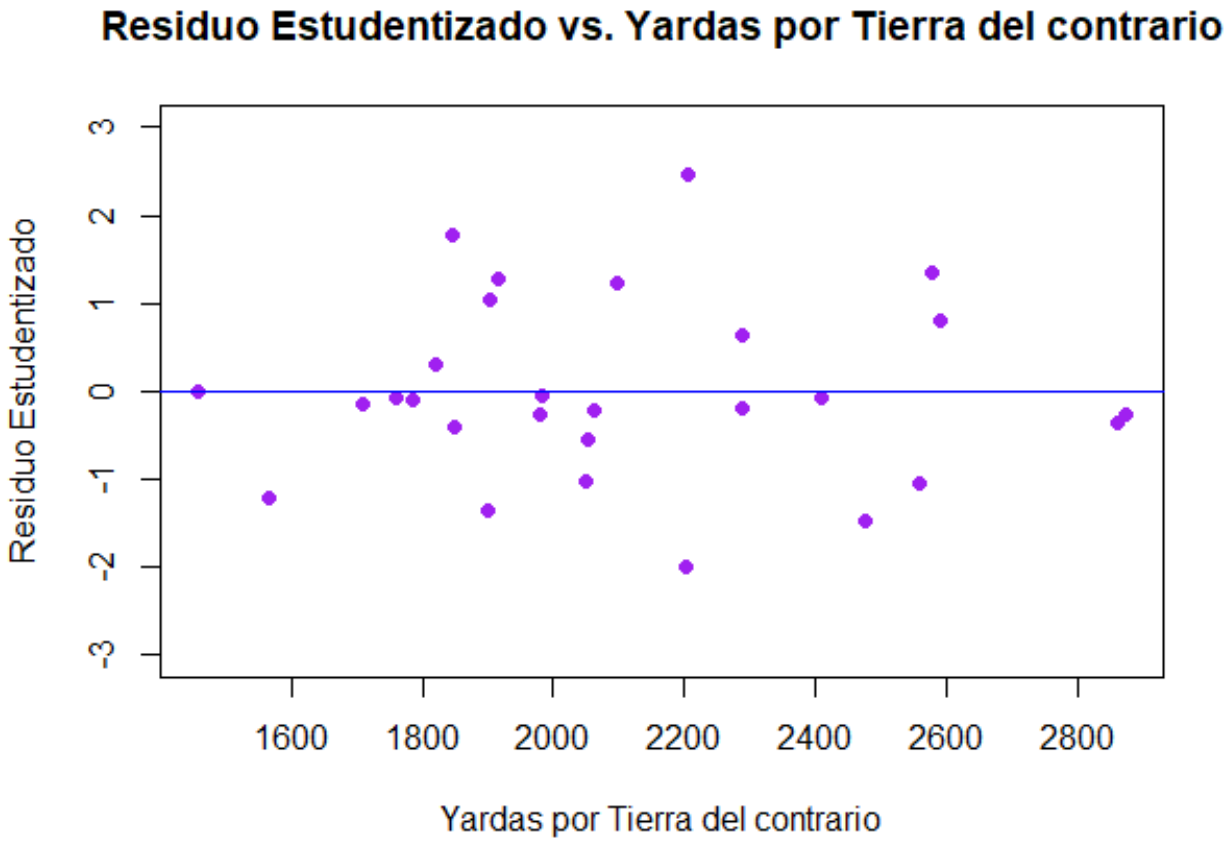


Figura: Gráfica r_i vs. Yardas por tierra del contrario

Como podemos observar las gráficas de los Residuales Estudentizados vs. Yardas por Aire y Yardas por Aire del contrario nos muestran una varianza constante mientras que la gráfica vs. Jugadas por tierra nos muestra una varianza en forma de embudo por lo que se debería considerar aplicarle una transformación a los regresores o a la respuesta.

Gráfica de Residuales en el tiempo

Veamos como se comportan los residuales, residuales estandardizados y estudentizados:

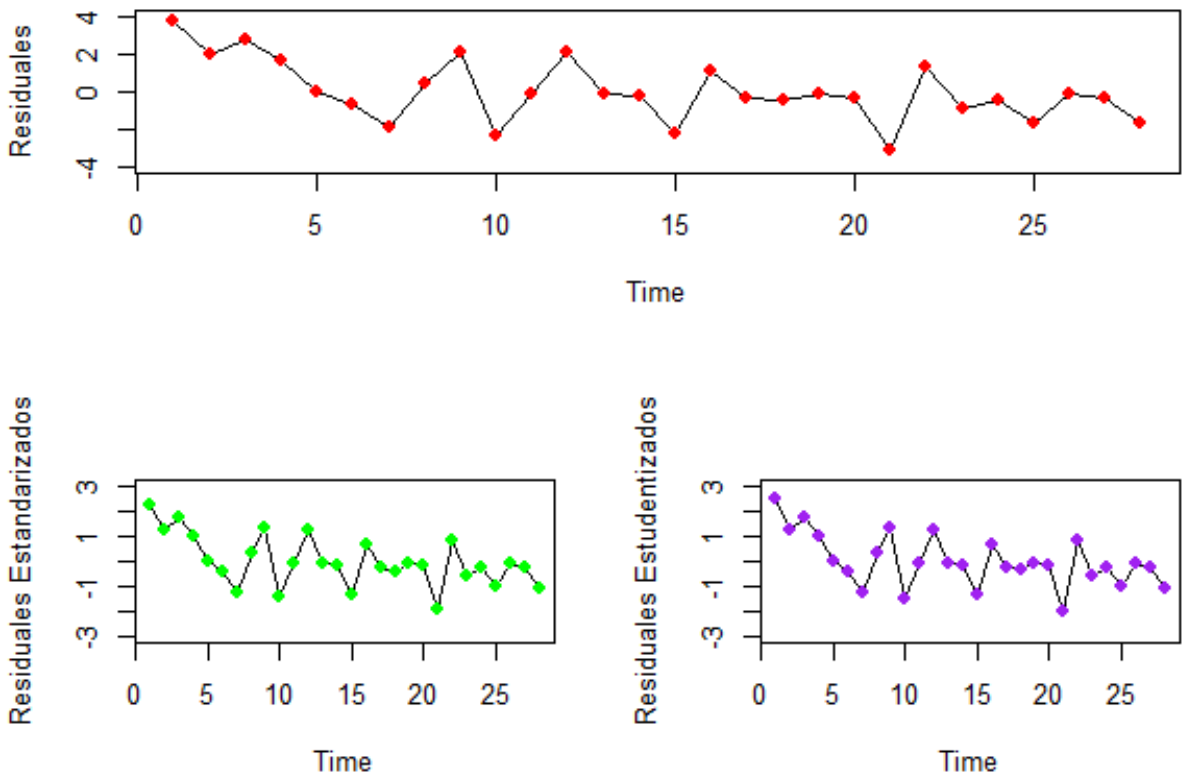


Figura: Gráfica de los Residuales, Estandarizados y Estudentizados vs. Tiempo

De aquí podemos sacar dos conclusiones, la primera es que para la gráfica $e_i vs. t$ se pude considerar que esta es una autocorrelación positiva mientras que para las gráficas $d_i vs. t$ y $r_i vs. t$ tenemos una autocorrelación negativa ya que tenemos picos mas bruscos que en la primera gráfica. Por lo tanto no se podría dar una conclusión al análisis de esta gráfica y es algo que nos confirma la prueba de Durbin-Watson ya que tenemos el caso $d_L < d < d_U$ lo que nos dice que la prueba es inconclusa.

Por otro lado tenemos que aplicando el meto de Cochrane-Orcutt usando la funcion cochrane.orcutt() para ver si se puede arreglar este problema de autocorrelacion obtuvimos:

```
Call:
lm(formula = JueGan ~ YardAir + JugTierr +
JugTierrCon, data = nfl)

 number of interaction: 16
 rho 0.194457

Durbin-Watson statistic
(original):     1.49518 , p-value: 7.919e-02
(transformed): 2.18805 , p-value: 7.139e-01

 coefficients:
(Intercept)      YardAir     JugTierr  JugTierrCon
   1.068639     0.003424     0.153855    -0.004989
```

Donde tenemos el siguiente análisis de la varianza:

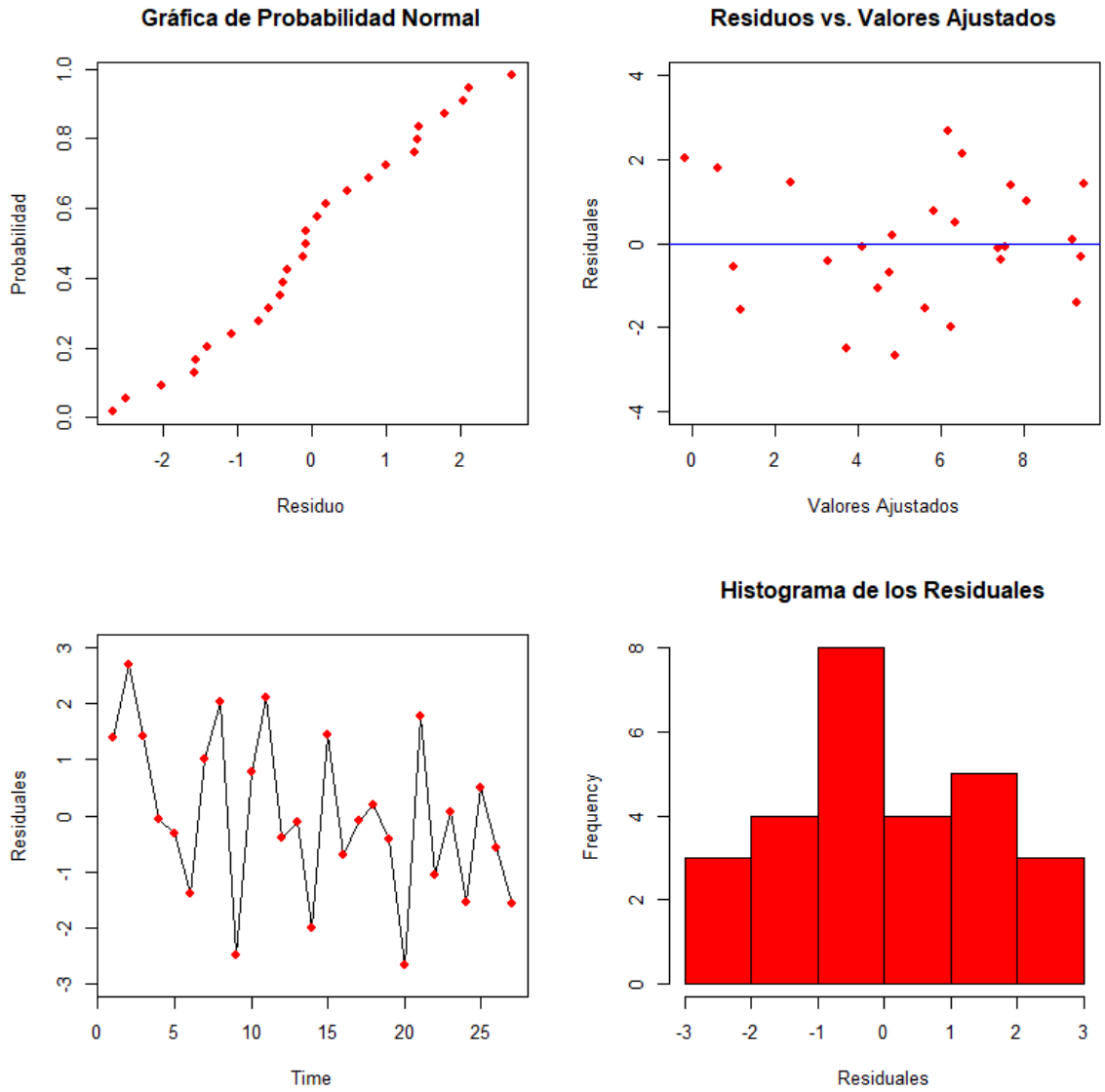


Figura: Varianza con el modelo ajustado por Cochrane-Orcutt

Por ultimo, note que con el modelo ajustado por Cochrane-Orcutt te obtuvo que $D = 2.18805$ y a un nivel de significancia de 5 % tenemos que $D_L = 1.21$ y $D_U = 1.65$, entonce $D_U < D$ por lo tanto no se rechaza la hipótesis nula por lo que la relación entre los residuales es cero.

La Estadística PRESS

Veamos que tan bueno es modelo en términos de predicción, para esto usaremos el estadístico PRESS:

$$PRESS = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}}\right) = 71.13127$$

como este PRESS se podría considerar un valor pequeño pero mayor que $SS_{Res} = 69.897$ por lo que podemos decir que el modelo es bueno prediciendo nuevos datos, pero se puede mejorar.

R^2 para predicción basado en PRESS

Por otro lado, veamos que pasa con $R^2_{predicción}$, calculando este estadístico tenemos:

$$R^2_{predicción} = 1 - \frac{PRESS}{SS_T} = 1 - \frac{71.13127}{326.9643} = 0.7824494$$

por lo que se espera que el modelo explique un 78.24494 % de la variabilidad cuando se predigan nuevas observaciones, que se podría considerar bueno en el contexto del problema.

Detección y Tratamiento de puntos Atípicos

Como dijimos los puntos sospechosos son los correspondientes a e_1 y e_{21} entonces siguiendo con el análisis podríamos ver que pasa si removemos esta observación e implementamos un modelo de regresión lineal entonces

	Con obs. 1 y 21	Sin obs. 1 y 21
$\hat{\beta}_0$	-1.917360632	-2.611008020
$\hat{\beta}_2$	0.003601534	0.003551225
$\hat{\beta}_7$	0.195170486	0.205545517
$\hat{\beta}_8$	-0.004801443	-0.004718313
R^2_{adj}	0.7595017	0.8237333
MS_{Res}	2.912	2.129
$se(\hat{\beta}_2)$	0.0006948656	0.0005953238
$se(\hat{\beta}_7)$	0.0880272786	0.0775157213
$se(\hat{\beta}_8)$	0.0012738517	0.0011155030

De los datos obtenido sin la observación 1 y 21 podemos notar que el modelo no cambia mucho, los coeficientes de los residuales se mantienen muy cercanos a los del modelo original por lo que se podría decir que este punto no controla la pendiente pero si un poco la ordenada al origen, y a su vez podemos observar como el R^2_{adj} mejora sin esta observación.

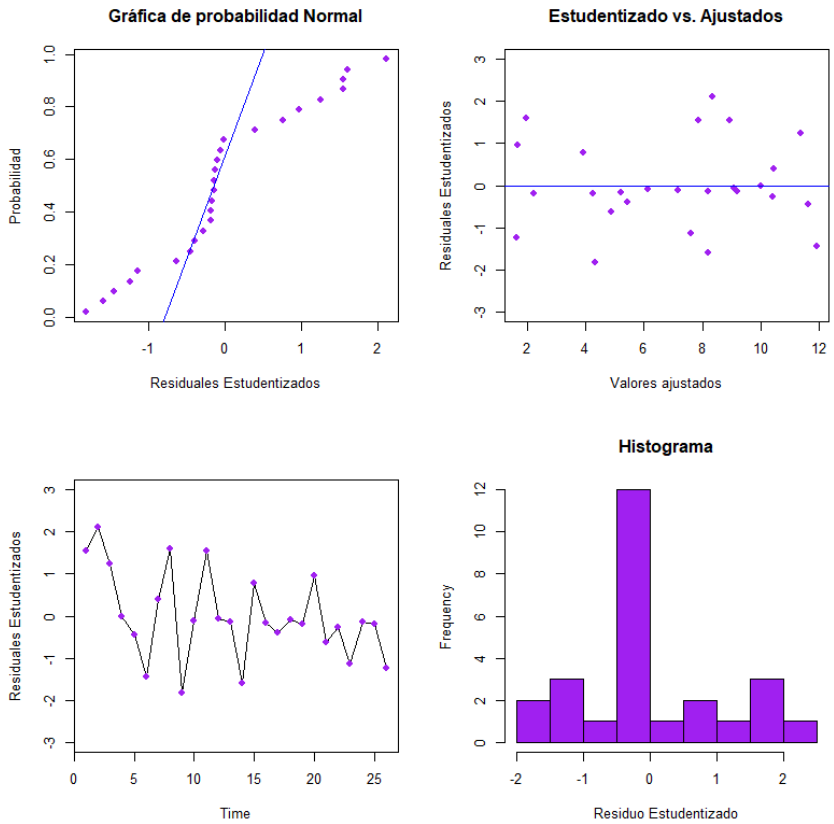


Figura: Análisis sin la observación 1 y 21.

Balanceo

A continuación presentamos la tabla de los h_{ii}:

n	h_{ii}	n	h_{ii}
1	0.05340136	15	0.05080683
2	0.12029037	16	0.08937553
3	0.11755846	17	0.26133019
4	0.15958409	18	0.39190019
5	0.19205088	19	0.10695559
6	0.15824274	20	0.11334766
7	0.14501355	21	0.0938162
8	0.11735977	22	0.14441799
9	0.17427869	23	0.11455174
10	0.12105929	24	0.12363034
11	0.21425913	25	0.07550686
12	0.06684418	26	0.12174679
13	0.08936696	27	0.31951924
14	0.11263885	28	0.15114654

Sabemos que $2p/n = 0.2857143$, por lo que valores en rojo de la tabla anterior son puntos de balanceo:

Corrida	$\hat{\beta}_0$	$\hat{\beta}_2$	$\hat{\beta}_7$	$\hat{\beta}_8$
Con 18 y 27	-1.917360632	0.003601534	0.195170486	-0.004801443
Sin 18 y 27	-3.594664120	0.003820771	0.200319979	-0.004337422

De estos dos modelos podemos notar que el valor de los coeficientes no cambia mucho, se podría decir que sin las observaciones 18 y 27 el modelo es casi el mismo sin embargo la $R^2_{adjSin18\&27} = 0.7353722$ menor que la R^2_{adj} del modelo original, por lo que podríamos decir que el modelo pierde un poco la capacidad de predicción, por muy poco pero la pierde.

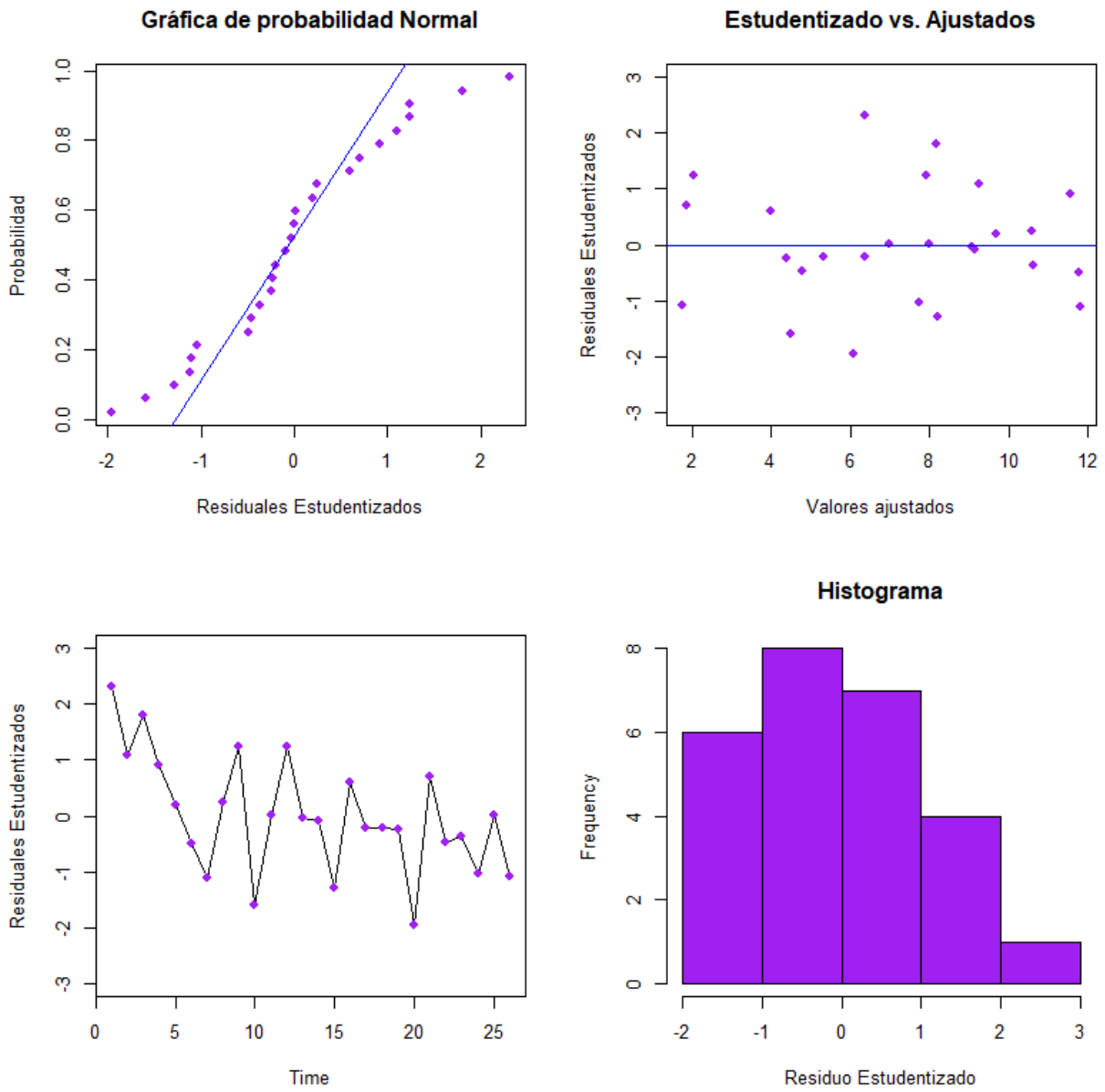


Figura: Análisis de Varianza del Modelo sin 18 y 27

La D de Cook

Siguiendo con el análisis tenemos que las distancias de Cook son

n	Cook(D_i)	n	Cook(D_i)
1	0.07017978	15	0.02378089
2	0.0514893	16	0.01030025
3	0.09634064	17	0.00433982
4	0.05018572	18	0.0220537
5	3.72E-06	19	0.00019799
6	0.00826653	20	0.00127282
7	0.06180676	21	0.09063611
8	0.00302303	22	0.0283198
9	0.09432621	23	0.0097041
10	0.07174283	24	0.00270696
11	0.00011032	25	0.02103453
12	0.02815362	26	0.00030978
13	0.00015909	27	0.00778853
14	0.00077821	28	0.04876519

De aquí podemos ver que ninguno de los puntos D_1, D_{18}, D_{21} y D_{27} es influyente ya que todo son menores que 1, de hecho ninguno de los puntos es influyente según el método de la distancia de Cook.

Gráficamente tenemos que los *DFFITS* son:

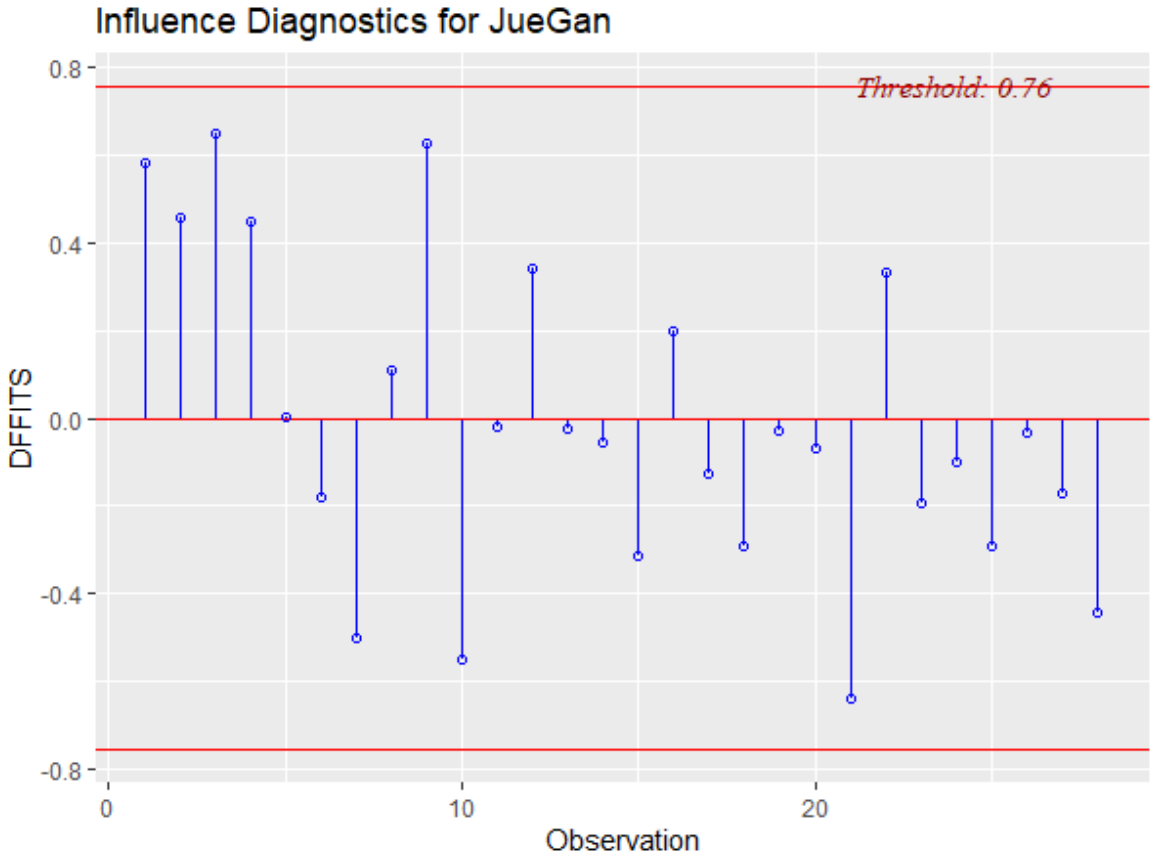
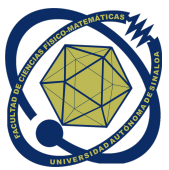


Figura: Gráfica de los *DFFITS*

La tabla de los datos es:

n	$DFFITS$	n	$DFFITS$
1	0.5825721	15	-0.31376751
2	0.45890323	16	0.20046751
3	0.64801075	17	-0.12911255
4	0.44860129	18	-0.29158891
5	0.0037771	19	-0.02755338
6	-0.17866822	20	-0.06990882
7	-0.5022409	21	-0.63780424
8	0.10785337	22	0.33418881
9	0.62504722	23	-0.19408758
10	-0.54877948	24	-0.10202928
11	-0.02056473	25	-0.29025615
12	0.33983404	26	-0.03446623
13	-0.02469875	27	-0.17302835
14	-0.05464589	28	-0.44257642



De la tabla anterior tenemos que los valores de $DFFITS_1, DFFITS_{18}, DFFITS_{21}$ y $DFFITS_{27}$ cumplen que sean mayor que $2\sqrt{4/28} = 0.7559289$ por lo que tenemos que ninguna de estas observaciones es influyentes, y ademas este método concuerda con el resultado que obtuvimos en la Distancia de Cook acerca de que ninguna de las observaciones es influyente.

Conclusión

El modelo original no aparenta "mejora" en el sentido de que la gráfica de probabilidad normal adopte la forma de una recta, por lo que lo ideal seria agregar más regresoras al modelo, o como vimos en la gráfica de los Residuales Estudentizados vs. Jugadas por Tierra aplicar una transformación a esta regresora para logar una forma constante en la gráfica y ver si esto tiene un repercusión en el modelo en general, aunque como pudimos observar al aplicar el meto de Cochrane-Orcutt lo mejor es seguir con este modelo transformado ya que la varianza se hace constante y no hay indicios de ningún punto influyente en el modelo.