Análisis del Desempeño de los equipos de la NFL en 1976

Carlos Enrique Ponce Villagran

Facultad de ciencias Físico-Matemáticas

11 de diciembre de 2019

Tabla de Contenido

El modelo se consideran las siguientes regresoras

$x_1 :=$ *Yardas por tierra (temporada)*

$x_2 :=$ *Yardas por aire (temporada)*

$x_3 :=$ *Promedio de pateo (temporada)*

$x_4 :=$ *Porcentaje de goles de campo (GC hechos/GC intentados, temporada)*

$x_5 :=$ *Diferencia de pérdida de balón (pérdidas ganadas/pérdidas perdidas)*

$x_6 :=$ *Yardas de castigo (temporada)* $x_7 :=$ *Porcentaje de Carreras (jugadas por tierra/jugadas totales)*

$x_8 :=$ *Yardas por tierra del contrario (temporada)*

$x_9 :=$ *Yardas por aire del contrario (temporada)*

y se desea saber si existe una relación con

$Y :=$ *Juegos ganados (por temporada de 14 juegos)*

Team	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
Washington	10	2113	1985	38.9	64.7	+4	868	59.7	2205	1917
Minnesota	11	2003	2855	38.8	61.3	+3	615	55.0	2096	1575
New England	11	2957	1737	40.1	60.0	+14	914	65.6	1847	2175
Oakland	13	2285	2905	41.6	45.3	−4	957	61.4	1903	2476
Pittsburgh	10	2971	1666	39.2	53.8	+15	836	66.1	1457	1866
Baltimore	11	2309	2927	39.7	74.1	+8	786	61.0	1848	2339
Los Angeles	10	2528	2341	38.1	65.4	+12	754	66.1	1564	2092
Dallas	11	2147	2737	37.0	78.3	−1	761	58.0	1821	1909
Atlanta	4	1689	1414	42.1	47.6	−3	714	57.0	2577	2001
Buffalo	2	2566	1838	42.3	54.2	−1	797	58.9	2476	2254
Chicago	7	2363	1480	37.3	48.0	+19	984	67.5	1984	2217
Cincinnati	10	2109	2191	39.5	51.9	+6	700	57.2	1917	1758
Cleveland	9	2295	2229	37.4	53.6	−5	1037	58.8	1761	2032
Denver	9	1932	2204	35.1	71.4	+3	986	58.6	1709	2025
Detroit	6	2213	2140	38.8	58.3	+6	819	59.2	1901	1686
Green Bay	5	1722	1730	36.6	52.6	−19	791	54.4	2288	1835
Houston	5	1498	2072	35.3	59.3	−5	776	49.6	2072	1914
Kansas City	5	1873	2929	41.1	55.3	+10	789	54.3	2861	2496
Miami	6	2118	2268	38.2	69.6	+6	582	58.7	2411	2670
New Orleans	4	1775	1983	39.3	78.3	+7	901	51.7	2289	2202
New York Giants	3	1904	1792	39.7	38.1	−9	734	61.9	2203	1988
New York Jets	3	1929	1606	39.7	68.8	−21	627	52.7	2592	2324
Philadelphia	4	2080	1492	35.5	68.8	−8	722	57.8	2053	2550
St. Louis	10	2301	2835	35.3	74.1	+2	683	59.7	1979	2110
San Diego	6	2040	2416	38.7	50.0	0	576	54.9	2048	2628
San Francisco	8	2447	1638	39.9	57.1	−8	848	65.3	1786	1776
Seattle	2	1416	2649	37.4	56.3	−22	684	43.8	2876	2524
Tampa Bay	0	1503	1503	39.3	47.0	−9	875	53.5	2560	2241

Figura: Datos del Problema

El modelo inicial estaba constituido por las regresoras:
$x_2 :=$ *Yardas por aire (temporada)*
$x_7 :=$ *Porcentaje de Carreras (jugadas por tierra/jugadas totales)*
$x_8 :=$ *Yardas por tierra del contrario (temporada)*
donde la tabla de análisis de varianza arrojo los siguientes resultados.

Primero se procedió a calcular el anova del modelo

Fuente de Variación	**Suma de Cuadrados**	**Grados de Libertad**	**Cuadrado medio**	F_0
Regresión	257.0671	3	85.68903	29.42231
Residuales	69.8972	24	2.912383	
Total	326.9643	27		

donde se concluyo que la hipótesis nula global del modelo se rechaza con $100(1-\alpha)\,\%$ si $F_0 > F_{\alpha,3,24}$, por ejemplo a un nivel de significancia $\alpha = 0.05$ la hipótesis nula se rechaza.

- Además, se obtuvo que cada una de las regresoras (x_2, x_7 y x_8) contribuían de forma significante al modelo haciendo sus respectivas pruebas de hipótesis individuales.
- Por ultimo se procedió a obtener los valores de $R^2 = 0.7862$ y $R^2_{Adj} = 0.7595$.

- Otro problema se abordo fue el de comparar los modelos con las regresoras x_1, x_7 y x_8 y el modelo con las regresoras x_7 y x_8.
- En este punto la herramientas que teníamos para comparar modelos eran limitadas por lo que se procedió a comparar los intervalos de confianza de los estimadores de cada regresor donde en el caso del modelo con dos regresoras se obtuvieron intervalos de confianza más anchos que que no es lo deseado.
- Ahora con las herramientas que tenemos del capitulo de selección de modelos podemos hacer un análisis más completo de cual es el modelo optimo.

De aquí tenemos que considerando todas las regresoras (K=9) obtuvimos la siguiente tabla:

p	Regresoras en el modelo	$SS_{Res}(p)$	R_p^2	$\overline{R}_p^2$	$MS_{Res}(p)$	C_p
3	$x_7 x_8$	147.898	0.548	0.511	13.43311	22.2
4	$x_2 x_7 x_8$	69.897	0.786	0.7595	3.766333	0.859

Primero usando el criterio de coeficientes de determinación múltiple tenemos

$$R_0^2 = 1 - (1 - R_{10}^2)(1 + d_{\alpha,28,9})$$

considerando un nivel de significancia $\alpha = 0.05$ tenemos que $d_{0.05,28,9}) = 9F_{0.05,28,18}/18 = 2.11/2 = 1.055$ y $R_{10}^2 = 0.8156$, entonces

$$R_0^2 = 1 - (1 - 0.8156)(1 + 1.055) = 0.6211$$

De aquí podemos observar que el modelo con 3 regresoras cumple $R_4^2 = 0.786 > 0.6211 = R_0^2$, por lo que este modelo es adecuado $R^2(0.05)$.

- Por otro lado el criterio de $R^2_{Adj,p}$ indica que se debe elegir el modelo que tenga el valor máximo de estos valores el cual es el modelo $R^2_{Adj,4} = 0.7595$ ya que $R^2_{Adj,3} = 0.511$.
- Por otro lado el criterio de los cuadrados medios de los resiudales, indica que se debe elegir el modelo que tenga el menor de estos valores el cual es, una vez más el modelo $MS_{Res}(4) = 3.766$ ya que $MS_{Res}(3) = 13.433$.

C_p de Mallows

Usando el criterio de la estadística C_p tenemos que el mejor modelo sera el que tenga el valor más pequeño de C_p, donde $C_4 = 0.8591$ y $C_3 = 22.1537$, por lo que una vez más el mejor modelo es el que tiene las regresoras x_2, x_7 y x_8.

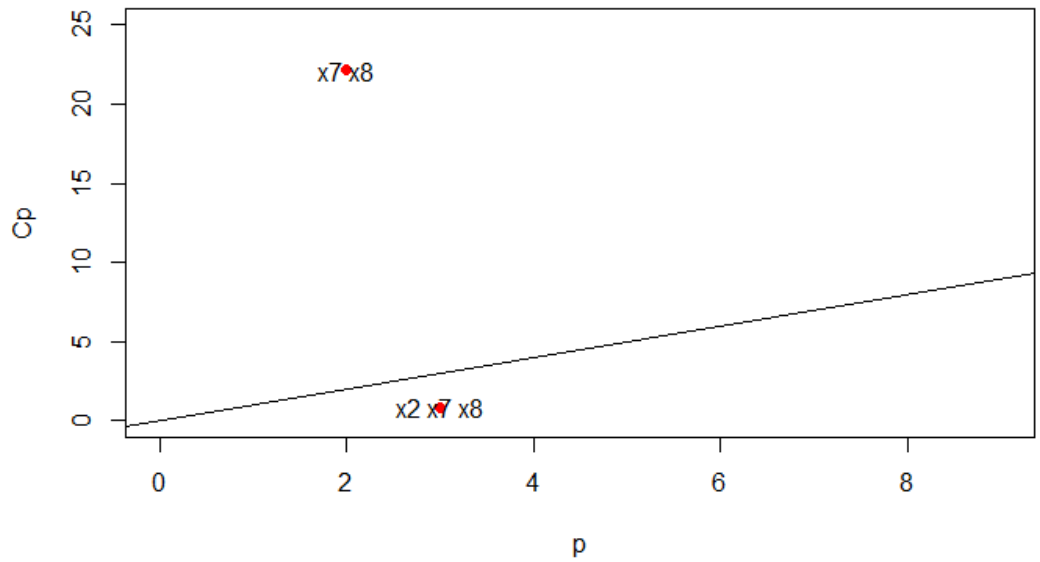

Figura: Gráfica de C_P

Como podemos observar, según los nuevos criterios que tenemos para comparar modelos la conclusión a la que se llego en la Tarea 2 era correcta, de estos dos modelos el mejor era el que tenia las regresoras x_2, x_7 y x_8, y el modelo perdía información y precisión.

Como recordamos el problema del modelo inicial era que los datos no seguían el supuesto de normalidad y además la gráfica de los Residuales vs. x_7 mostraba la forma de un embudo y por ende no era constante.

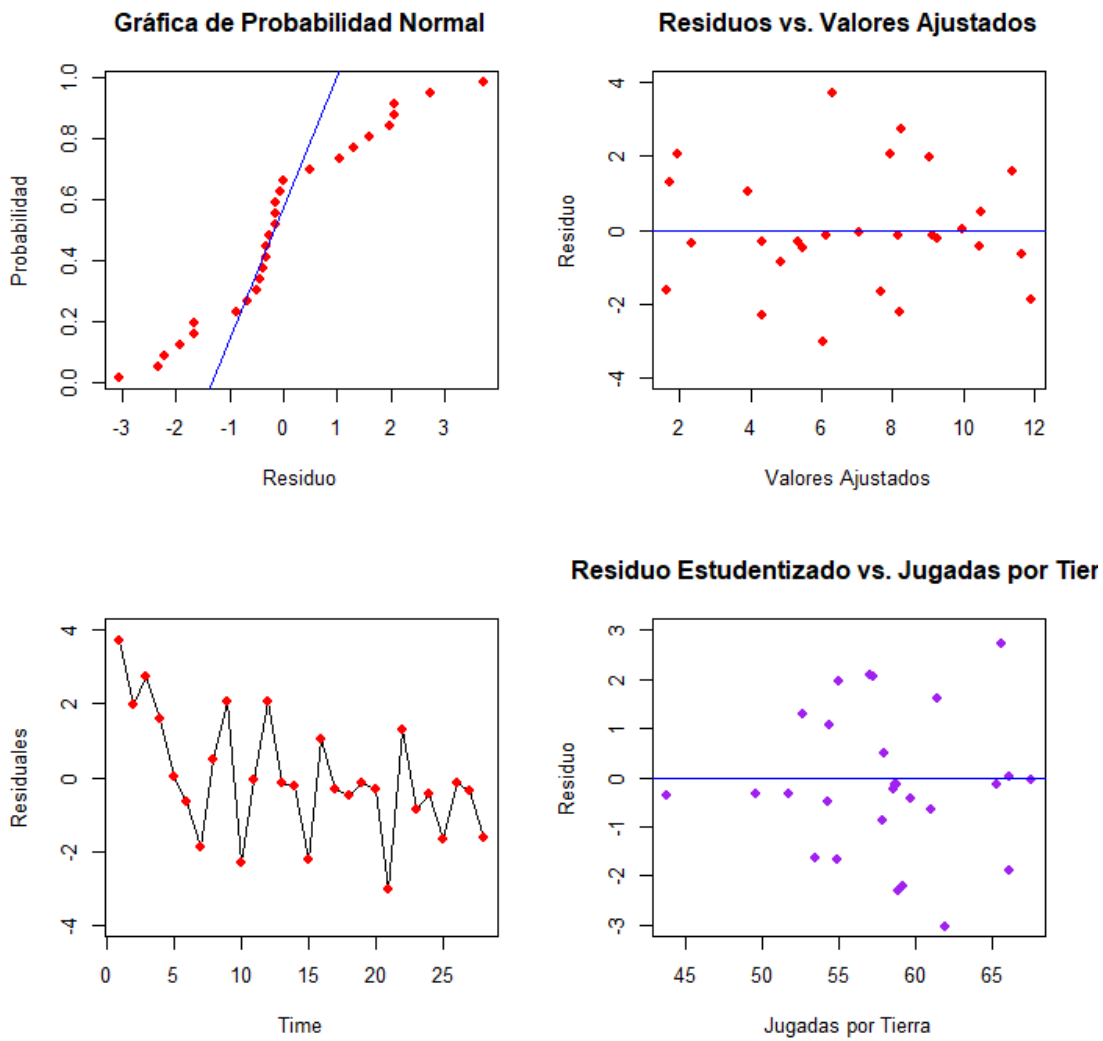

Figura: Analisis de la Varianza

Otro aspecto a considerad es que la prueba de Durbin-Watson era inconclusa para nuestro modelo, y al aplicar el método de Cochrane-Orcutt el problema de autocorrelación quedo solucionado en el nuevo modelo, además de otros resultados en los análisis de la varianza que dieron solución a dos de nuestros problemas.

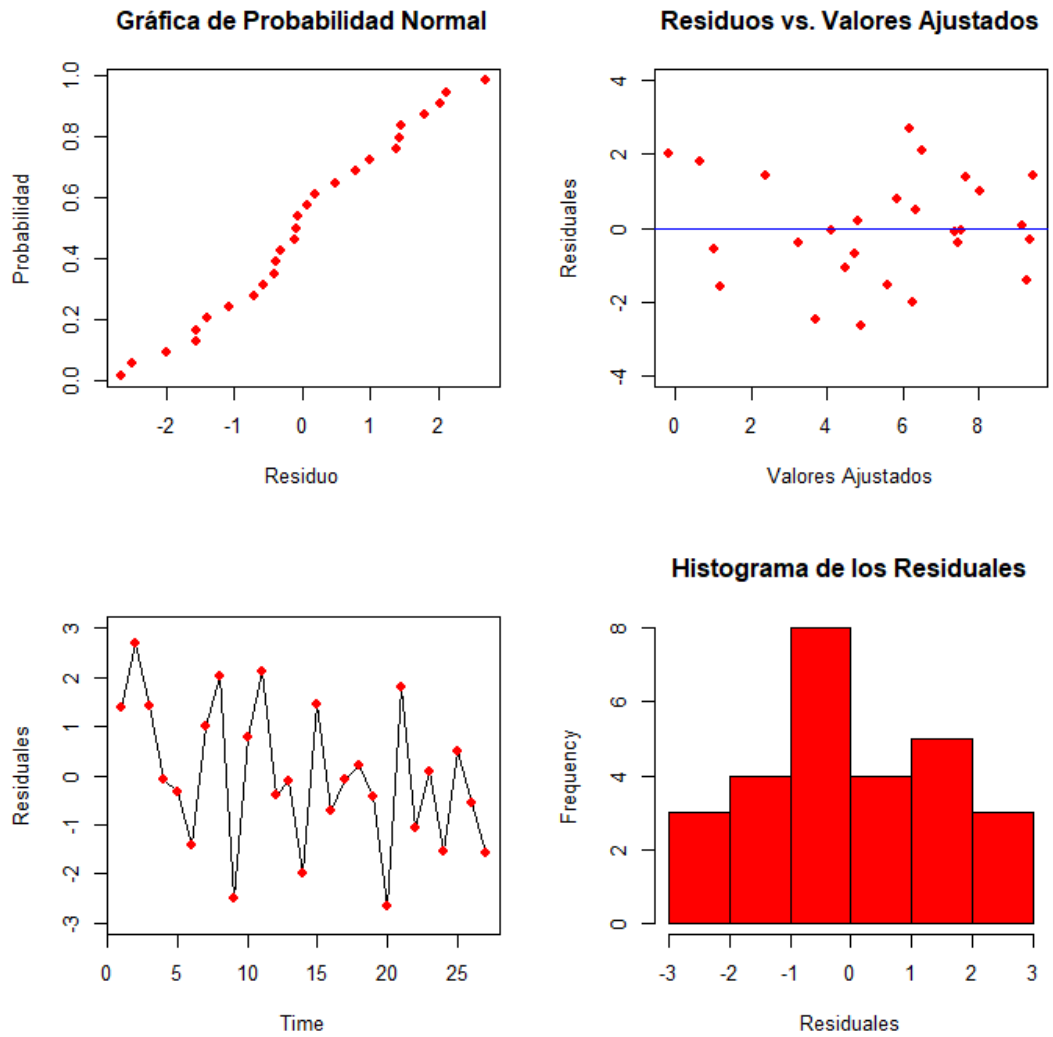

Figura: Modelo después de usar Cochrane-Orcutt

- Como podemos notar la gráfica de probabilidad normal cambia bastante y cumple lo ideal que es que los residuales estén sobre una linea recta, la Varianza es constante (todo los residuos se puede contener sobre dos bandas) y el problema de autocorrelación quedó solucionado.
- También se trato el problemas de puntos atípicos, pero al remover los puntos sospechosos del modelo este no aparentaba mucha mejoría por lo que se decidió no eliminar ninguno de estos del modelo para que no se perdiera más información.

Nuestros modelo completo consta de 9 regresoras en total lo que quiere decir que tenemos $2^9 = 512$ modelos para revisar, como obviamente no es viable comparar y revisar cada uno de los modelos usaremos el método de selección por pasos para ver que modelos obtenemos.

Usando la función de R stepAIC() para el método de selección por pasos se obtuvo el siguiente modelo final

```
y ~ x2 + x7 + x8 + x9

       Df Sum of Sq     RSS     AIC
<none>               65.004  33.583
- x9    1     4.866  69.870  33.604
- x7    1    16.908  81.913  38.057
- x8    1    23.299  88.303  40.160
- x2    1    82.892 147.897  54.601

Call:
lm(formula = y ~ x2 + x7 + x8 + x9, data = NFLTabla)

Coefficients:
(Intercept)           x2           x7           x8           x9
  -1.821703     0.003819     0.216894    -0.004015    -0.001635
```

Una vez más usando la misma función stepAIC(, direction = "backward") para usar el método de selección hacia atrás se obtuvo el siguiente modelo final

```
y ~ x2 + x7 + x8 + x9

       Df Sum of Sq     RSS    AIC
<none>              65.004 33.583
- x9    1     4.866  69.870 33.604
- x7    1    16.908  81.913 38.057
- x8    1    23.299  88.303 40.160
- x2    1    82.892 147.897 54.601

Call:
lm(formula = y ~ x2 + x7 + x8 + x9, data = NFLTabla)

Coefficients:
(Intercept)           x2           x7           x8           x9
  -1.821703     0.003819     0.216894    -0.004015    -0.001635
```

Entonces usando la función ols_step_all_possible() para que nos muestre todos los modelos posibles tenemos que los que tienen la C_p más pequeña son:

n	Regresoras	R^2_p	$R^2_{Adj,p}$	C_p	AIC
3	$x_2 x_7 x_8$	0.7863069	0.7595953	0.8590659	115.0647
4	$x_2 x_7 x_8 x_9$	0.8011882	0.7666123	1.4064672	115.0435
3	$x_1 x_2 x_8$	0.7775056	0.7496938	1.7181792	116.1948
⋮	⋮	⋮	⋮	⋮	⋮

Como podemos notar el mejor modelo según los métodos de selección hacia adelante y atrás muestras que el modelo es el que tiene las regresoras (x_2, x_7, x_8, x_9) es el mejor en ambos casos, por lo que compararemos este modelo con el que tiene las regresoras (x_2, x_7, x_8) que ha sido el modelo con el que hemos estado trabajando.

(x_2, x_7, x_8, x_9) vs. (x_2, x_7, x_8)

Compararemos los modelos con los criterios ya antes mencionados los cuales son usando R_p^2, $R_{Adj,p}^2$, $MS_{Res}(p)$ y C_p.

p	Regresoras en el modelo	$SS_{Res}(p)$	R_p^2	$R_{Adj,p}^2$	$MS_{Res}(p)$	C_p
5	$x_2x_7x_8x_9$	65.004	0.8012	0.7666	2.826	1.4065
4	$x_2x_7x_8$	69.897	0.786	0.7595	2.912	0.859

Recordemos que ya calculamos el valor $R_0^2 = 0.6211$.

- Como podemos observar de la tabla tanto como R_4^2 y R_5^2 son adecuados $R^2(0.05)$ por lo que la elección de cual es el "mejor modelo" no es clara con este criterio.
- Por otro lado si consideramos las $R_{Adj,p}^2$ tenemos que el modelo que maximiza es $R_{Adj,5}^2$ pero note que esto es por una diferencia muy pequeña.
- De nuevo tenemos que el modelo que minimiza los $MS_{Res}(p)$ es $MS_{Res}(5)$.
- Por ultimo, el criterio de la C_p de Mellow indica que el modelo ideal es el modelo con C_4.

Ahora, para poder decidir cual es el "Mejor modelo" entre estos dos primero hay que saber para que queremos nuestro modelo, en este caso lo que queremos es estimar los juegos ganados de un equipo usando sus estadísticas a lo largo de la temporada, por lo que lo ideal sería comparar las $PRESS_p$ de los modelos.

De aquí tenemos que $PRESS_5 = 65.00435$ y $PRESS_4 = 69.8972$, como podemos observar el modelo (x_2, x_7, x_8) tiene mejor capacidad de predicción por lo que para nuestros propósitos del problema este es el mejor modelo.

Para determinar si nuestro modelo con las regresoras (x_2, x_7, x_8) y aplicado el método de Cochrane-Orrcut tiene multicolinealidad usaremos el factor de incremento de la varianza

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

usando la funciona vif() se obtuvieron los siguientes valores

x0.2	x0.7	x0.8
1.147736	2.032427	1.843541

De aquí como los tres valor VIF_i son menor que 10 se concluye que el modelo (reducido) no tiene multicolinealidad por lo que no hay que hacerle ningún otro cambio.

Conclusión

El modelo final (x_2, x_7, x_8), resulto ser el mejor modelo para nuestras necesidades, note que no hay mucha diferencia con el modelo que tenia las regresoras (x_2, x_7, x_8, x_9) respecto a los valores R^2, R^2_{Adj}, MS_{Res} y C_p pero la capacidad de predicción era peor que el modelo con 3 regresoras.

De aquí, lo que podemos concluir es que para estimar los juegos ganados de un equipo en la temporada, la mejor forma de hacerlo solo implica saber cuantas Yardas por aire hicieron (x_2), Porcentaje de carreras (x_7) y las yardas por tierra del contrario (x_8).