Facultad de Ciencias Físico-Matemáticas

Modelos de Regresion

Proyecto Final

19 de diciembre de 2019

Carlos Enrique Ponce Villagran

Análisis del Desempeño de los equipos de la NFL en 1976

Modelo de regresión múltiple

1. Problemas relacionados a la tabla B.1.

   3.1. Para los datos de la Liga Nacional de Fútbol, en la tabla B.1 del apéndice:

   a) Ajustar un modelo de regresión lineal múltiple que relacione la cantidad de juegos ganados con las yardas por aire del equipo (x_2), el porcentaje de jugadas por tierra (x_7) y las yardas por tierra del contrario (x_8).

   Solución: Introduciendo los datos a R $y := Juegos\ ganados$, $x_2 := Yardas\ por\ aire$, $x_7 := Porcentaje\ de\ carreras$ y $x_8 := Yardas\ por\ tierra\ del\ contrario$ y usando la función

   ```
   regres.multp<-lm(JueGan ~ YardAir+JugTierr+JugTierrCon,data = beisbol)
   ```

   se obtuvo que el modelo de regresión lineal múltiple ajustado es

   $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_7 x_7 + \hat{\beta}_8 x_8 = -1.9173606 + 0.0036015 x_2 + 0.1951705 x_7 - 0.0048014 x_8$$

   □

   b) Formar la tabla de análisis de varianza y probar el significado de la regresión.

   Solución: La tabla de análisis de varianza es:

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		

   por lo que la hipótesis nula global del modelo se rechaza con $100(1-\alpha)\,\%$ si $F_0 > F_{\alpha,3,24}$.

   □

   c) Calcular el estadístico t para probar las hipótesis $H_0 : \beta_2 = 0$, $H_0 : \beta_7 = 0$ y $H_0 : \beta_8 = 0$. ¿Qué conclusiones se pueden sacar acerca del papel de las variables x_2, x_7 y x_8 en el modelo?

   Solución: Calculando los errores estándar se obtuvo que:

   $$se(\hat{\beta}_2) = 0.0006948656, \quad se(\hat{\beta}_7) = 0.08802727 \quad \& \quad se(\hat{\beta}_8) = 0.001273852$$

   entonces para la prueba de hipótesis

   $$H_0 : \beta_2 = 0 \quad \text{vs.} \quad H_1 : \beta_2 \neq 0$$

   tenemos que

   $$t_0 = \frac{\hat{\beta}_2}{se(\hat{\beta}_2)} = 5.183017$$

de aquí, si consideramos un nivel de significancia del $5\,\%$ tenemos que $t_{0.025,24} = 2.064$, por lo tanto como $\mid t_0 \mid > t_{0.025,24}$ se rechaza la hipótesis nula por lo que el regresor x_2 contribuye en forma significativa al modelo.
Ahora, para la prueba de hipótesis

$$H_0 : \beta_7 = 0 \quad \text{vs.} \quad H_1 : \beta_7 \neq 0$$

tenemos que

$$t_0 = \frac{\hat{\beta}_7}{se(\hat{\beta}_7)} = 2.21716$$

si consideramos un nivel de significancia del $5\,\%$ tenemos que $t_{0.025,24} = 2.064$, por lo tanto como $\mid t_0 \mid > t_{0.025,24}$ se rechaza la hipótesis nula por lo que el regresor x_7 contribuye en forma significativa al modelo.
Ahora, para la prueba de hipótesis

$$H_0 : \beta_8 = 0 \quad \text{vs.} \quad H_1 : \beta_8 \neq 0$$

tenemos que

$$t_0 = \frac{\hat{\beta}_8}{se(\hat{\beta}_8)} = -3.769056$$

si consideramos un nivel de significancia del $5\,\%$ tenemos que $t_{0.025,24} = 2.064$, por lo tanto como $\mid t_0 \mid > t_{0.025,24}$ se rechaza la hipótesis nula por lo que el regresor x_8 contribuye en forma significativa al modelo.

□

d) Calcular R^2 y R^2_{Adj} para este modelo.

Solución: De R tenemos que el valor de R^2 es $R^2 = 0.7862$ mientras que R^2_{Adj} es

$$R^2_{Adj} = 1 - \frac{SS_{Res}/(n-p)}{SS_T/(n-1)} = 1 - \frac{SS_{Res}/(n-p)}{SS_T/(27)} = 1 - \frac{2.912383}{12.10979} = 1 - 0.2404982 = 0.7595$$

por lo tanto, de estos valor se puede decir que el modelo ajustado es una buena aproximación a los valores observados.

□

e) Con la prueba F parcial, determinar la contribución de x_7 al modelo. ¿Cómo se relaciona el estadístico F parcial con la prueba t para β_7 calculada en la parte c anterior?

Solución: Solo no interesa la contribución de x_7 al modelo por lo que queremos probar la hipótesis

$$H_0 : \beta_7 = 0 \quad \text{vs.} \quad H_1 : \beta_7 \neq 0$$

entonces para probar esta hipótesis nula usaremos el estadístico

$$F_0 = \frac{SS_R(\beta_2 \mid \beta_1)/r}{MS_{Res}} = \frac{(SS_R(\beta) - SS_R(\beta_1))/r}{MS_{Res}}$$

donde $SS_R(\beta_1)$ es la suma de cuadrados de la regresión del modelo reducido $\hat{y} = \beta_0 + \beta_2 x_2 + \beta_8 x_8 + \varepsilon$ y SS_R es la suma de cuadrados del modelo completo al igual que MS_{Res}.
entonces calculando la $SS_R(\beta_1)$ del modelo reducido se obtuvo que $SS_R(\beta_1) = 242.7504$, entonces de la tabla de análisis de varianza que se obtuvo en (b) tenemos que

$$F_0 = \frac{(257.0671 - 242.7504)/1}{2.912383} = 4.915803$$

entonces, con un nivel de significancia del $5\,\%$, $F_{0.05,1,24} = 4.26$ por lo tanto $F_0 > 4.26$ y se concluye que x_7 contribuye al modelo de forma significante.
Como esta prueba solo se hizo para una sola de las variables es equivalente a la prueba t que se hizo en el inciso (c).

□

3.2. Con los resultados del Problema 3.1, demostrar en forma numérica que el cuadrado del coeficiente de correlación simple entre los valores observados y_i y los valores ajustados $\hat{y}_i$ es igual a R^2.

Solución: El coeficiente de correlación lo calcularemos con

$$r = \frac{S_{y\hat{y}}}{[S_{yy}SS_T]^{1/2}}$$

donde $S_{y\hat{y}} = 257.0652$ y $S_{yy} = 326.9643 = SS_t$, por lo que

$$r = \frac{257.0652}{326.9643} = 0.7862 = R^2$$

□

3.3. Vease el problema 3.1. Calcular:

a) Un intervalo de confianza del $95\,\%$ para β_7

Solución: Sabemos que un intervalo de confianza para β_7 es de la forma

$$\hat{\beta}_7 - t_{\alpha/2,n-p}se(\hat{\beta}_7) \leq \beta_j \leq \hat{\beta}_7 + t_{\alpha/2,n-p}se(\hat{\beta}_7)$$

del ejercicio (c) del problema 3.1. tenemos que $\hat{\beta}_7 = 0.1951705$, $se(\hat{\beta}_7) = 0.08802727$ y $t_{0.025,24} = 2.064$ por lo que un intervalo de confianza para β_7 es

$$0.01348221472 \leq \beta_7 \leq 0.3768587853$$

□

b) Un intervalo de confianza del $95\,\%$ para la cantidad media de juegos ganados por un equipo cuando $x_2 = 2300$, $x_7 = 56$ y $x_8 = 2100$.

Solución: Sea $x_0 = [1, 2300, 56, 2100]^t$ un intervalo de confianza para la respuesta media en x_0 tiene la forma

$$\hat{y_0} - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 x_0^t (X^t X)^{-1} x_0} \leq E(y \mid x_0) \leq \hat{y_0} + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 x_0^t (X^t X)^{-1} x_0}$$

donde $\hat{y_0} = x_0^t \hat{\beta} = 7.212697$, $t_{0.025,24} = 2.064$ y $\sqrt{\hat{\sigma}^2 x_0^t (X^t X)^{-1} x_0} = \sqrt{(2.912383)(0.04900961)} = 0.3778025$, por lo que un intervalo de confianza para la cantidad media de juegos ganados es

$$6.43267264 \leq E(y \mid x_0) \leq 7.99248136$$

□

3.4. Para los datos de la Liga Nacional de Fútbol del problema 3.1, ajustar un modelo a esos datos, usando solo x_7 y x_8 como regresores.

a) Probar la significancia de la regresión.

Solución: Para realizar la prueba de significancia haremos la siguiente prueba de hipotesis

$$H_0 : \beta_7 = \beta_8 = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0 \text{ al menos para una j}$$

para saber si se rechaza o no H_0 usaremos el estadistico F_0 donde

$$F_0 = \frac{MS_R}{MS_{Res}}$$

entonces como $SS_R = 178.8283$, $SS_{Res} = 148.1359$, $k = 2$ y $n - p = 25$ tenemos que

$$F_0 = \frac{178.8283/2}{148.1359/25} = 15.08989$$

de aquí si consideramos un nivel $\alpha = 0.05$ tenemos que $F_{0.05,2,25} = 3.39$, por lo tanto como $F_0 > 3.39$ se rechaza H_0, entonces se concluye que al menos uno de los regresores contribuye al modelo en forma significativa.

□

b) Calcular R^2 y R^2_{Adj}. ¿Cómo se comparan esas cantidades con las calculadas para el modelo del problema 3.1, que tenia un regresor más (x_2)?

Solución: De R se obtuvo que

$$R^2 = 0.5469 \quad \& \quad R^2_{Adj} = 0.5107$$

En ambos casos los valores de R^2 y R^2_{Adj} decrecieron por lo que el regresor x_2 si aporta al sistema de manera significativa, por lo que se esperaría que este permaneciera en el modelo.

□

c) Calcular un intervalo de confianza de $95\,\%$ para β_7. También, un intervalo de confianza de $95\,\%$ para la cantidad media de juegos ganados por un equipo cuando $x_7 = 56$ y $x_8 = 2100$. Comparar la longitud de esos intervalos de confianza con las longitudes de los correspondientes en el problema 3.3.

Solución: Sabemos que un intervalo de confianza para β_7 es de la forma

$$\hat{\beta}_7 - t_{\alpha/2,n-p} se(\hat{\beta}_7) \leq \beta_j \leq \hat{\beta}_7 + t_{\alpha/2,n-p} se(\hat{\beta}_7)$$

de aquí tenemos que $\hat{\beta}_7 = 0.050218$, $se(\hat{\beta}_7) = 0.119055$ y $t_{0.025,25} = 2.06$ por lo que un intervalo de confianza para β_7 es

$$-0.1950353 \leq \beta_7 \leq 0.2954713$$

ahora, sea $x_0 = [1, 56, 2100]^t$ un intervalo de confianza para la respuesta media en x_0 tiene la forma

$$\hat{y_0} - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 x_0^t (X^t X)^{-1} x_0} \leq E(y \mid x_0) \leq \hat{y_0} + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 x_0^t (X^t X)^{-1} x_0}$$

donde $\hat{y_0} = x_0^t \hat{\beta} = 6.919985$, $t_{0.025,25} = 2.06$ y $\sqrt{\hat{\sigma}^2 x_0^t (X^t X)^{-1} x_0} = \sqrt{(5.925438)(0.04791235)} = 0.5328242$, por lo que un intervalo de confianza para la cantidad media de juegos ganados es

$$5.822367148 \leq E(y \mid x_0) \leq 8.017602852$$

a comparación con los intervalos obtenidos en 3.3, tenemos que el intervalo de confianza para β_7 de este problema es más grande que el obtenido en 3.3, mientras que el intervalo de confianza para la respuesta media de este problema también es más grande que el obtenido en 3.3.

□

d) ¿Qué conclusiones se pueden sacar de este problema, acerca de las consecuencias de omitir un regresor importante de un modelo?

Solución: Si se pierde un regresor importante del sistema que aporte de manera significante al modelo, este mismo no sera un "buen"modelo o bien hará falta información para considerar este un "buen"modelo.

□

Comprobación de la Adecuación del Modelo

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)*

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:

i	e_i	d_i	r_i	i	e_i	d_i	r_i
1	3.70382102	2.23070825	2.45276788	15	-2.21647025	-1.33309262	-1.35620152
2	1.9644307	1.22727746	1.24101022	16	1.05513548	0.64790941	0.63988855
3	2.72657876	1.70079039	1.7754083	17	-0.3248969	-0.22151142	-0.21706951
4	1.60858433	1.02818774	1.02946799	18	-0.49235925	-0.3699733	-0.36322079
5	0.01213948	0.00791378	0.00774716	19	-0.13114518	-0.0813189	-0.0796177
6	-0.6566606	-0.419395	-0.41207744	20	-0.32068925	-0.19956457	-0.19552503
7	-1.90514129	-1.2073225	-1.21951571	21	-3.04006094	-1.87132777	-1.98224182
8	0.48350328	0.30156663	0.29577807	22	1.29315427	0.81920998	0.8134148
9	2.07339399	1.33702894	1.36052785	23	-0.87961837	-0.5477576	-0.53960816
10	-2.30942598	-1.44344412	-1.47869499	24	-0.44260879	-0.27704613	-0.27164765
11	-0.06085327	-0.04022722	-0.03938156	25	-1.66544826	-1.01497424	-1.01564064
12	2.0670156	1.25384078	1.26973229	26	-0.15120631	-0.09454439	-0.092571
13	-0.13114062	-0.08052695	-0.07884211	27	-0.36261807	-0.25758302	-0.25250889
14	-0.25174324	-0.15659696	-0.15337819	28	-1.64567034	-1.04665174	-1.04883101

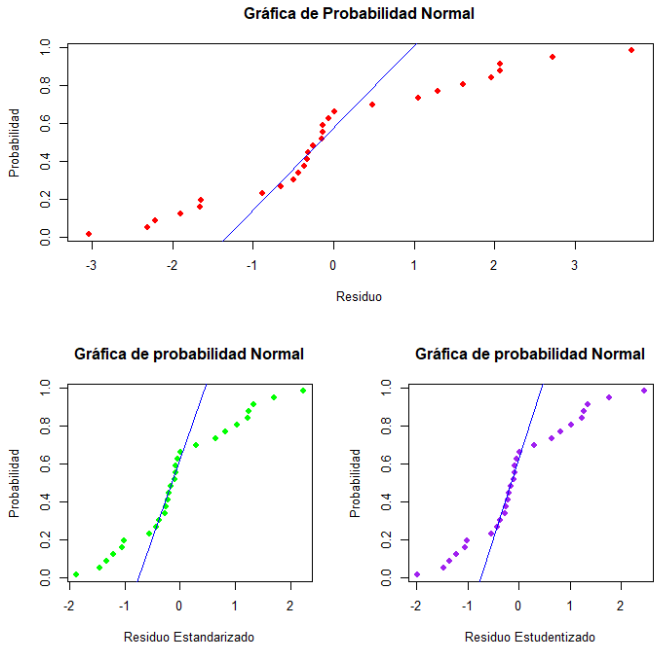

Figura 1: Graficas de Probabilidad Normal

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}.

- Note que usando la función shapiro.test() a los residuales obtenemos un p-valor igual a $p = 0.4568$ por lo que a un nivel de significancia igual a $\alpha = 0.05$ tenemos que $p > \alpha$ por lo que la hipótesis nula no se rechaza, sin embargo veremos que esto se puede mejorar aún más.

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

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

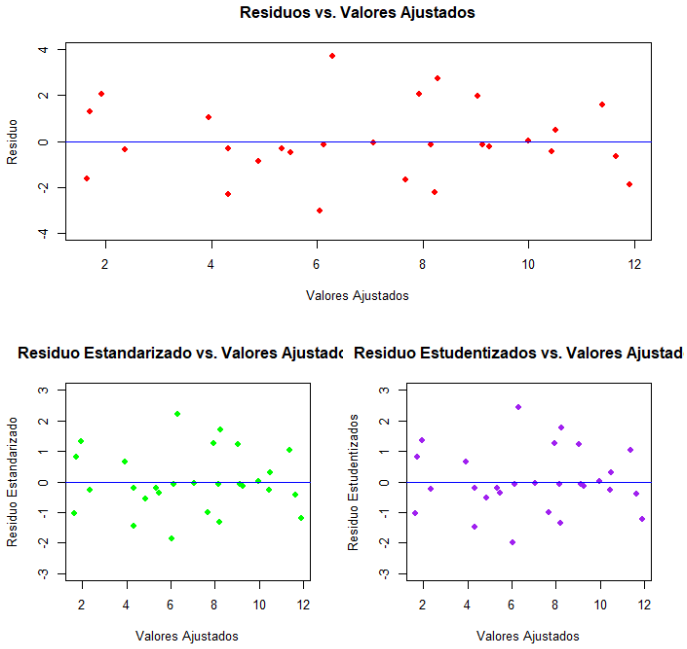

Figura 2: Graficas de los Residuales vs.$\hat{y}_i$

De las tres gráficas podemos notar que los puntos que más destacan o son más sospechosos, 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:

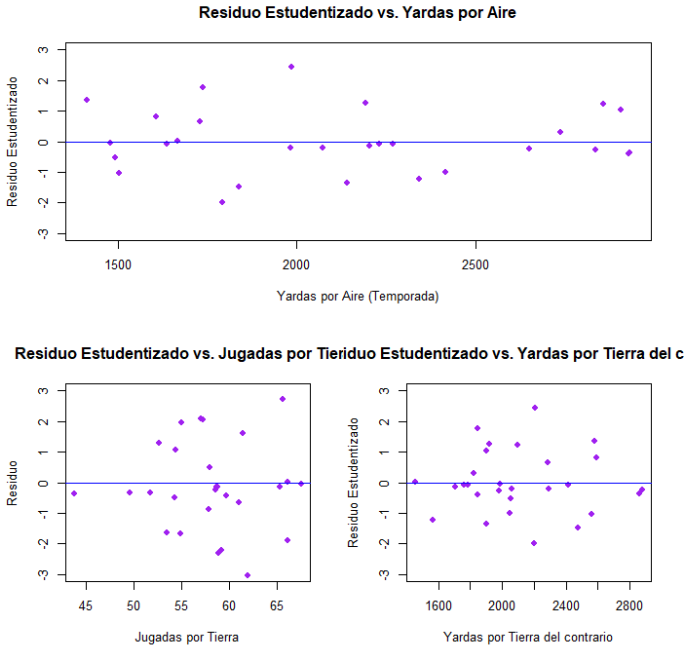

Figura 3: Grafica r_i vs. Regresor

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:

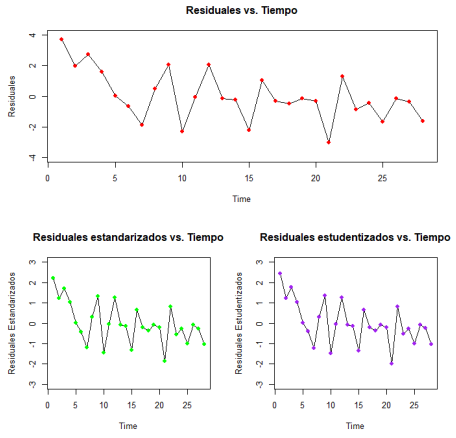

Figura 4: Gráfica de los Residuales vs. Tiempo

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.

- Esto se con confirma con la prueba de Durbin-Watson. Utilizando la funcion durbinWatsonTest() tenemos que $d = 1.495182$ y como $d_L = 1.21$ y $d_U = 1.65$ por lo tanto 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
```

Entonces comprobando la adecuación de este nuevo modelo obtuvimos los siguientes resultados:

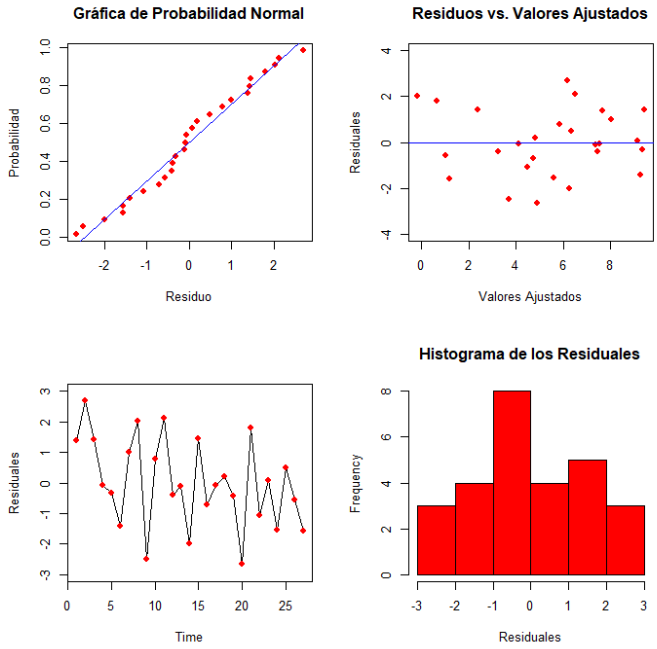

Figura 5: Modelo con Cochrane-Orcutt

- 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.
- Aplicando la prueba Shapiro-Wilk a este nuevo modelo se obtuvo un $p - valor$ igual a $p = 0.8174$ el cual es mucho más grande que el del modelo anterior, y por lo tanto a un nivel de significancia igual a $\alpha = 0.05$ tenemos que $p > \alpha$ por lo que la hipótesis nula no se rechaza.

A partir de ahora continuaremos el análisis con este modelo, ya que soluciono los problemas de normalidad que teníamos.

La estadistica PRESS

Como nuestro modelo se basa en estimar los juegos ganas en la temporada de los equipos, es importante saber el valor del estadístico PRESS para saber que tan bueno es el modelo a la hora de hacer predicciones.

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

$$PRESS = \sum_{i=1}^{n} (\frac{e_i}{1 - h_{ii}}) = 70.447$$

como este PRESS se podría considerar un valor pequeño pero mayor que $SS_{Res} = 53.582$ por lo que podemos decir que el modelo es bueno prediciendo nuevos datos.

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

$$R^2_{predicción} = 1 - \frac{PRESS}{SS_T} = 1 - \frac{70.477}{254.019} = 0.7225523$$

por lo que se espera que el modelo explique un $72.26\,\%$ de la variabilidad cuando se predigan nuevas observaciones, que se podría considerar bueno en el contexto del problema.

Blanceo

Veamos si el modelo cuenta con puntos de balanceo para esto vemos la siguiente tabla:

i	h_{ii}	i	h_{ii}
1	0.12080186	15	0.09507569
2	0.13301924	16	0.22200644
3	0.16331008	17	0.44231351
4	0.1934601	18	0.06623145
5	0.16692495	19	0.1286421
6	0.13215392	20	0.11051123
7	0.10293176	21	0.15639193
8	0.23939764	22	0.09183477
9	0.09520727	23	0.15579325
10	0.17332406	24	0.07016943
11	0.08953722	25	0.12779636
12	0.07936516	26	0.36441391
13	0.09280188	27	0.14400717
14	0.0425776		

sabemos que $2p/n = 0.2962963$, por lo que los elementos en rojo de la tabla anterior son puntos de balanceo.

Corrida	$\hat{\beta}_0$	$\hat{\beta}_2$	$\hat{\beta}_7$	$\hat{\beta}_8$	MS_{Res}	R^2
Con 17 y 26	0.860836452	0.003424222	0.153855001	-0.004988947	2.330	0.7891
Sin 17	0.590867721	0.003445906	0.156710271	-0.004927371	2.435	0.7874
Sin 26	1.264032164	0.003451853	0.141421084	-0.004892546	2.413	0.7672
Sin 17 y 26	0.457526716	0.003525133	0.148793478	-0.004681906	2.522	0.7649

De estos modelos podemos notar que el valor de los coeficientes no cambia mucho, se podría decir que sin las observaciones 17 y 26 el modelo es casi el mismo sin embargo las $R^2_{Sin17\&26}$ son menores que la R^2 del modelo original, por lo que podríamos decir que el modelo pierde un poco la capacidad de predicción, por lo que lo ideal es seguir con las observaciones 17 y 26 para no perder información que no nos afecta de forma considerable.

La D de Cook

Siguiendo con el análisis tenemos que las distancias de Cook son:

i	$Cook(D_i)$	i	$Cook(D_i)$
1	0.03202861	15	0.02602966
2	0.13712403	16	0.01996979
3	0.05100934	17	0.00098715
4	0.00015331	18	0.00031674
5	0.00264721	19	0.00323195
6	0.03700603	20	0.10621457
7	0.013805	21	0.0753605
8	0.18411371	22	0.01373622
9	0.0776888	23	0.00012398
10	0.01652663	24	0.02102349
11	0.0520469	25	0.00435442
12	0.0015179	26	0.03103553
13	0.00015379	27	0.05210058
14	0.01994226		

De aquí podemos notar que ninguna de las observaciones D_{17} y D_{26} son influyentes ya que ambas son menores que 1, de hecho ninguna de las observaciones es influyente según el método de la distancia de Cook.

DFFITS

Gráficamente tenemos que los DFFITS son:

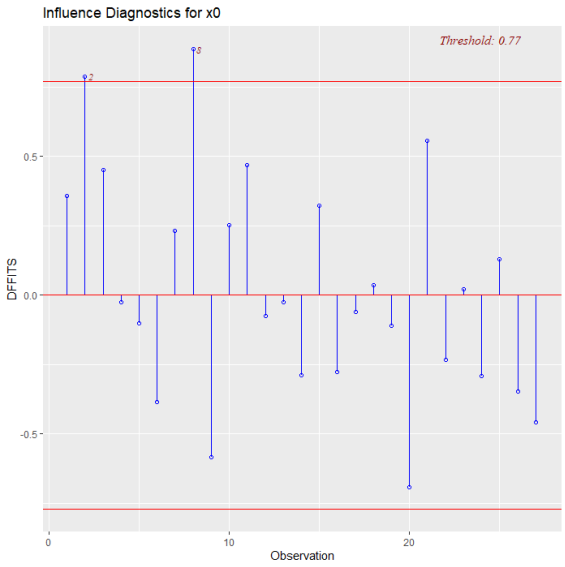

Figura 6: Grafica de los $DFFITS$

mientas que la tabla de los datos es:

i	$DFFITS_I$	i	$DFFITS_I$
1	0.3573823	15	0.32260815
2	0.78816457	16	-0.27811422
3	0.45217123	17	-0.06146346
4	-0.02422047	18	0.03482535
5	-0.10075612	19	-0.11141361
6	-0.38449501	20	-0.69091218
7	0.23226696	21	0.55701986
8	0.88555887	22	-0.23200773
9	-0.58398052	23	0.02178139
10	0.25320191	24	-0.29074614
11	0.46831919	25	0.12940995
12	-0.07632459	26	-0.34622695
13	-0.02426062	27	-0.45900861
14	-0.28767115		

donde las observaciones en rojo son puntos atípicos ya que cumplen que $\mid DFFITS_i \mid > 0.77 = 2\sqrt{p/n}$, entonces veamos que tanto mejor el modelo sin estas observaciones.

Corrida	$\hat{\beta}_0$	$\hat{\beta}_2$	$\hat{\beta}_7$	$\hat{\beta}_8$	MS_{Res}	R^2
Con 1 y 8	0.860836452	0.003424222	0.153855001	-0.004988947	2.330	0.7891
Sin 1	0.572359336	0.003282855	0.160320605	-0.004888730	2.337	0.7862
Sin 8	1.956024696	0.003706396	0.140237821	-0.005601790	2.188	0.8
Sin 1 y 8	1.653671310	0.003569757	0.146519432	-0.005486228	2.213	0.7963

Como podemos observar de la tabla los coeficientes no cambian mucho en los modelos, sin embargo el error cuadrático medio disminuye y el R^2 aumenta en la corrida sin la observación 8 por lo que es preferible usar el modelo sin esta observación.

Selección de Variables

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() donde se busca minimizar el AIC para el método de selección por pasos se obtuvo el siguiente modelo final

```
Start:  AIC=41.48
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9

       Df Sum of Sq     RSS    AIC
- x5    1     0.000  60.293 39.476
- x1    1     0.549  60.842 39.730
- x3    1     0.746  61.039 39.821
- x6    1     0.803  61.096 39.847
- x4    1     1.968  62.261 40.376
- x7    1     3.451  63.744 41.035
<none>              60.293 41.476
- x9    1     5.348  65.642 41.856
- x8    1    12.072  72.365 44.587
- x2    1    62.448 122.741 59.380

Step:  AIC=39.48
y ~ x1 + x2 + x3 + x4 + x6 + x7 + x8 + x9

       Df Sum of Sq     RSS    AIC
- x1    1     0.553  60.846 37.732
- x3    1     0.750  61.043 37.822
- x6    1     0.818  61.111 37.854
- x4    1     2.053  62.346 38.414
- x7    1     3.859  64.152 39.213
<none>              60.293 39.476
- x9    1     5.351  65.644 39.857
- x8    1    12.086  72.379 42.592
- x2    1    66.979 127.272 58.395

Step:  AIC=37.73
y ~ x2 + x3 + x4 + x6 + x7 + x8 + x9

       Df Sum of Sq     RSS    AIC
- x6    1     0.690  61.536 36.048
- x3    1     1.715  62.561 36.510
- x4    1     3.051  63.897 37.102
<none>              60.846 37.732
- x9    1     4.852  65.698 37.880
- x7    1     8.961  69.807 39.579
- x8    1    16.599  77.445 42.486
- x2    1    67.010 127.856 56.524

Step:  AIC=36.05
y ~ x2 + x3 + x4 + x7 + x8 + x9

       Df Sum of Sq     RSS    AIC
```

```
- x3     1      1.726   63.262 34.822
- x4     1      2.767   64.303 35.279
<none>                   61.536 36.048
- x9     1      4.831   66.367 36.164
- x7     1      9.390   70.926 38.024
- x8     1     18.314   79.851 41.343
- x2     1     66.447  127.984 54.552

Step:  AIC=34.82
y ~ x2 + x4 + x7 + x8 + x9

       Df Sum of Sq     RSS    AIC
- x4    1      1.743   65.004 33.583
<none>                 63.262 34.822
- x9    1      5.629   68.891 35.209
- x8    1     17.701   80.962 39.730
- x7    1     18.583   81.845 40.033
- x2    1     75.598  138.860 54.835

Step:  AIC=33.58
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
```

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

```
Start:  AIC=41.48
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9

       Df Sum of Sq     RSS    AIC
- x5    1      0.000   60.293 39.476
- x1    1      0.549   60.842 39.730
- x3    1      0.746   61.039 39.821
- x6    1      0.803   61.096 39.847
- x4    1      1.968   62.261 40.376
- x7    1      3.451   63.744 41.035
<none>                 60.293 41.476
- x9    1      5.348   65.642 41.856
- x8    1     12.072   72.365 44.587
- x2    1     62.448  122.741 59.380

Step:  AIC=39.48
```

```
y ~ x1 + x2 + x3 + x4 + x6 + x7 + x8 + x9

       Df Sum of Sq     RSS    AIC
- x1    1     0.553  60.846 37.732
- x3    1     0.750  61.043 37.822
- x6    1     0.818  61.111 37.854
- x4    1     2.053  62.346 38.414
- x7    1     3.859  64.152 39.213
<none>              60.293 39.476
- x9    1     5.351  65.644 39.857
- x8    1    12.086  72.379 42.592
- x2    1    66.979 127.272 58.395

Step:  AIC=37.73
y ~ x2 + x3 + x4 + x6 + x7 + x8 + x9

       Df Sum of Sq     RSS    AIC
- x6    1     0.690  61.536 36.048
- x3    1     1.715  62.561 36.510
- x4    1     3.051  63.897 37.102
<none>              60.846 37.732
- x9    1     4.852  65.698 37.880
- x7    1     8.961  69.807 39.579
- x8    1    16.599  77.445 42.486
- x2    1    67.010 127.856 56.524

Step:  AIC=36.05
y ~ x2 + x3 + x4 + x7 + x8 + x9

       Df Sum of Sq     RSS    AIC
- x3    1     1.726  63.262 34.822
- x4    1     2.767  64.303 35.279
<none>              61.536 36.048
- x9    1     4.831  66.367 36.164
- x7    1     9.390  70.926 38.024
- x8    1    18.314  79.851 41.343
- x2    1    66.447 127.984 54.552

Step:  AIC=34.82
y ~ x2 + x4 + x7 + x8 + x9

       Df Sum of Sq     RSS    AIC
- x4    1     1.743  65.004 33.583
<none>              63.262 34.822
- x9    1     5.629  68.891 35.209
- x8    1    17.701  80.962 39.730
- x7    1    18.583  81.845 40.033
- x2    1    75.598 138.860 54.835

Step:  AIC=33.58
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
```

Por ultimo usando la función ols_step_forward_aic() para usar el método de selección hacia adelante se obtuvo el siguiente modelo final

```
ols_step_forward_aic(regres.multp)
Forward Selection Method
---------------------------

Candidate Terms:

1 . x1
2 . x2
3 . x3
4 . x4
5 . x5
6 . x6
7 . x7
8 . x8
9 . x9


Variables Entered:

 x8
 x2
 x7
 x9

No more variables to be added.
                        Selection Summary
--------------------------------------------------------------------------
Variable        AIC        Sum Sq       RSS        R-Sq      Adj. R-Sq
--------------------------------------------------------------------------
x8           132.245       178.092   148.872     0.54468       0.52717
x2           118.201       243.026    83.938     0.74328       0.72274
x7           115.065       257.094    69.870     0.78631       0.75960
x9           115.044       261.960    65.004     0.80119       0.76661
--------------------------------------------------------------------------
```

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, donde note que estos modelos son los que tienen mejor valores de R^2_p, $R^2_{Adj,p}$, C_p y AIC.

n	Regresoras	R_p^2	$R_{Adj,p}^2$	C_p	AIC
3	$x_2x_7x_8$	0.7863069	0.7595953	0.8590659	115.0647
4	$x_2x_7x_8x_9$	0.8011882	0.7666123	1.4064672	115.0435
3	$x_1x_2x_8$	0.7775056	0.7496938	1.7181792	116.1948
⋮	⋮	⋮	⋮	⋮	⋮

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

calcularemos R_0^2

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

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

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

- 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".

Multicolinealidad

Primero veamos si el modelo completo cuenta con multicolinealidad. Para esto usaremos el factor de incremento de la varianza

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

usando la función vif() se obtuvieron los siguientes valores de VIF del modelo completo:

```
      x1       x2       x3       x4       x5       x6       x7       x8       x9
4.827645 1.420161 2.126597 1.566107 1.924035 1.275979 5.414572 4.535643 1.423390
```

por lo que no tenemos problema de multicolinealidad alta ya que ninguno de estos valores es mayor que 10.

Entonces considerando el modelo con las regresoras (x_2, x_7, x_8) la función vif() arrojo los siguientes resultados:

```
    x2       x7       x8
1.118038 2.045056 1.905125
```

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 otra modificación a nuestro modelo.

- Note que este modelo es el modelo sin aplicarle la transformación de Cochrane-Orcutt, pero esto no cambia la conclusión sobre la multicolinealidad ya que esta transformación es lineal por lo que se conserva el hecho de que el modelo no tiene multicolinealidad.

Conclusión

El modelo original no aparenta "mejora" en el sentido de que la gráfica de probabilidad normal adopte la forma de una recta, aunque como pudimos observar al aplicar el método de Cochrane-Orcutt mejoro la gráfica de probabilidad normal bastante por lo que se procedió a seguir con este modelo transformado ya que la varianza se hace constante y también eliminar la observación influyente (8) que hizo que bajara el error cuadrático medio y el R^2.

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) aplicándole la transformación de Cochrane-Orcutt.