Econometría

# Trabajo final

*Estimación de un modelo de regresión lineal múltiple*

Elaborado por:
Roberto González Téllez

Diciembre de 2020

# Modelo de Béisbol - Victorias por Temporada

R. González[a]

[a]*ITAM - Instituto Tecnológico Autónomo de México*
*Calle Río Hondo 1, Progreso Tizapán, Mexico City, Mexico*

**Abstract**

El béisbol es uno de los deportes preferidos por los científicos de datos debido a la enorme cantidad de estadísticas que se generan en él y que, sin duda, siguen aumentando. La mayoría de equipos en años recientes han incorporado en su equipo de trabajo a analistas encargados de maximizar el desempeño de los equipos, ya sea a través de maximizar el costo-beneficio de contratos con jugadores o eligiendo la alineación ideal para cada partido. El objetivo de este trabajo es encontrar las estadísticas principales en las que un equipo de analistas debería fijarse a la hora de buscar jugadores, con el fin de maximizar el número de partidos ganados por temporada. Para ello utilizaremos modelos de regresión lineal por Mínimos Cuadrados Ordinarios (MMCO). Hacemos un análisis incremental para elegir el modelo y concluimos que las principales estadísticas en las que se debería poner atención son: "Earned Run Average (ERA)", "On-Base Percentage (OBP)" y "Slugging percentage (SLG)", que serán detalladas a lo largo del artículo.

*Keywords:* Béisbol, victorias, planeación, on-base, slugging, ERA

## 1. Introducción

Los deportes en años recientes han estado más abiertos a la inclusión de análisis estadístico en su día a día. El bésibol, al contar con una cantidad inmensa de estadísticas para ponderar el valor de los jugadores y evaluar su desempeño, parece ser el deporte ideal para implementar este tipo de análisis. Asimismo, diariamente existe una enormidad de gente tratando de generar estadísticas que sean aún mejores y que tomen en cuenta factores adicionales tales como las dimensiones del campo, el clima del lugar en donde se encuentra un estadio, entre muchas otras.
La idea detrás de este trabajo es tratar de predecir el número de victorias que

puede lograr un equipo en una temporada (Wins) utilizando como variables explicativas ERA, OBP, SLG, "Home-Runs"(HR) y "Walks and Hits per Inning Pitched (WHIP)". También se incluye una variable *dummy* que toma el valor de uno (1) cuando el equipo pertenece a la Liga Americana y cero (0) en otro caso. Esta variable teóricamente es relevante ya que la diferencia entre ligas radica en que en la Americana, en el lugar de bateo del Pitcher se alinea a un jugador designado para batear en esos turnos. Dicho jugador suele ser un jugador con estadísticas ofensivas (sobretodo SLG) mejores a las promedio. Así pues, el modelo teórico es el siguiente:

$$Wins = f(ERA, OBP, SLG, HR, WHIP, American\_League) \qquad (1)$$

Este trabajo agrega valor a la literatura estadística centrada en temas deportivos, dado que la misma es de reciente incorporación a la cotidianeidad. Nuestros resultados prueban estadísticamente que las variables ERA, OBP y SLG suelen ser 3 grandes predictoras de desempeño de los equipos y los equipos deberían centrarse en buscar jugadores que, por un lado, minimicen ERA y, por el otro, maximicen OBP y SLG.

## 2. Datos y Estadística Descriptiva

Los datos fueron obtenidos del sitio web "Baseball Reference" (2015-2020), tomando las estadísticas desagregadas al nivel equipo tanto para bateo como pitcheo desde los años 2015 hasta 2019. Inicialmente el conjunto de datos era conformado por datos panel con 30 observaciones por cada uno de los 5 años estudiados, dando como total 150 observaciones a nivel equipo. Sin embargo, para facilitar el análisis se construyó una nueva base de datos, la cual consiste en solamente 30 observaciones (una observación por equipo) en donde cada dato se calculó tomando el promedio de esa estadística a lo largo de los 5 años de los que se tenía información.
Las estadísticas usadas son:
**ERA**: Calcula el promedio de carreras limpias (ER) que un pitcher permite por cada 9 entradas lanzadas (IP) (i.e. un partido completo). $ERA = \frac{9 \cdot 1ER}{IP}$
**OBP**: calcula el porcentaje de veces que un bateador llega a una base ya sea por hit(H), base por bolas (BB), golpeado por lanzamiento (HBP) o base por bolas intencional (IBB), todo dividido por turnos al bate (AB),BB,HBP y flies de sacrificio (SAC). Es decir: $OBP = \frac{H+BB+HBP+IBB}{AB+BB+HBP+SAC}$
**SLG**: Calcula el número promedio de bases que un jugador obtiene por cada

turno al bate. A diferncia de OBP, esta estadística solo utiliza los hits y pondera según el número de bases. Así pues: $SLG = \frac{1B+2\cdot 2B+3\cdot 3B+4\cdot 1HR}{AB}$
**HR**: Un HR se hace cuando el batazo del jugador rebasa las paredes del outfield saliendo por la zona delimitada como "en juego" o cuando el batazo conecta con uno de los postes de foul que delimitan la zona de "en juego"
**WHIP**: Contabiliza el número promedio de jugadores que alcanzan base con un pitcher determinado por cada entrada que lanza ese pitcher. Solo toma en cuenta hits y Bases por Bola (no intencionales). Así: $WHIP = \frac{H+BB}{IP}$
En la sección de Anexos podrá encontrar todas las visualizaciones de los métodos empleados en el presente trabajo.
Al inicio de dicha sección (p.7) hallará la estadística descriptiva y en las Figuras 1(a)-1(c) los diagramas de caja y brazos de las variables del modelo.
Asimismo, se presentan los diagramas de dispersión de cada regresora contra la variable de respuesta para confirmar que la forma funcional que mejor explica la relación es de tipo lineal (1(d)-1(h)).

## 3. Métodos y Verificación de Supuestos

El objetivo del trabajo es predecir el número de victorias que un equipo de MLB tendrá en una temporada dadas ciertas estadísticas consideradas relevantes en el ámbito del bésibol.
Para hacer este estudio, realizamos un análisis incremental y corrimos regresiones lineales comenzando por la regresora más correlacionada con la variable de respuesta y posteriormente fuimos agregando el resto de independientes según la correlación que guardaran con la variable explicada. Tras observar la significancia de las regresoras en cada uno de los modelos y su $\overline{R}^2$. Posteriormente se verificó que el mejor modelo fuera de hecho el elegido a través de los Criterios de Información de Akaike (AIC) y Schwartz (BIC). Por último, se verificó que el modelo elegido cumpliera con los supuestos del modelo de regresión lineal múltiple por MMCO.
En la página 8 (Figura 1(i)) se presenta la matriz de correlaciones utilizada para el análisis incremental.
En la siguiente página (p.9) se presentan los resultados de los modelos de regresión estimados.
Es sencillo notar que el mejor modelo para continuar con el análisis es el modelo número 3 de la primera tabla de resultados de las regresiones, que tiene como variables explicativas a ERA, OBP y SLG.

Es decir, el modelo econométrico que se usará queda expresado como:

$$Wins_i = \beta_0 + \beta_1 * ERA_i + \beta_2 * OBP_i + \beta_3 * SLG_i + \epsilon_i \tag{2}$$

Para confirmar, se calcularon los criterios de información (CI) de todos los modelos y observamos que en efecto aquél modelo con CI mínimos fue el tercero. El resultado con el CI de Akaike es de 135.17 y el del CI de Schwartz es de 142.176.
Sin embargo, cabe notar que la teoría nos dice que pertenecer a la Liga Americana debería aumentar significativamente SLG, por lo que haremos una prueba más de significancia entre un modelo restringido (Modelo 3) y uno no restringido (Modelo 3 + interacción entre SLG y pertenencia a Liga Americana) Los resultados de la regresión se presentan en la página 10. Al calcular el estadístico F, se obtiene que $F = 0.0126 < 4.31 = F_{2,24}$, por lo que no podemos rechazar la hipótesis de que los coeficientes extra del modelo no restringido sean 0.
El intervalo de confianza para los coeficientes de nuestros modelo también se presenta en la página 10.
La bondad de ajuste del modelo se refleja en: $R^2 = .949$ & $\overline{R}^2 = .943$
Visualmente podemos observar esta bondad de ajuste en la sección de Anexos (p.11)(Figura 1(n)), donde la línea roja representa los valores observados y la línea azul los valores predichos por el modelo. Se presenta como complemento la tabla de Análisis de Varianza (Figura 1(o)).
Finalmente, comprobaremos que nuestros estimadores de hecho sean MELI, es decir, los mejores estimadores lineales insesgados y eficientes. En cada una de las verificaciones de los supuestos para asegurar que los estimadores sean MELI iniciaremos con una aproximación gráfica y posteriormente realizaremos las pruebas estadísticas relevantes para comprobar nuestras hipótesis.

### *3.1. Normalidad de $\epsilon_i$*

Se presenta el Histograma de residuales con una curva normal sobrepuesta en la Figura 1(p). Parece haber indicios de que en efecto los residuales se distribuyen normales.
Probaremos esta hipótesis formalmente con la prueba de Jarque-Bera dado que el tamaño de muestra es de 30.
Notar que como el p-value del estadístico es mayor (p. 12, Figura 1(q)) a 0.05 entonces no se rechaza al 95% de confianza la hipótesis de que los residuales se distribuyan de forma normal.

*3.2. No multicolinealidad*

Se presenta la matriz de correlaciones entre variables, misma que se usó para el análisis incremental en un inicio (Figura 1(i)). La matriz de correlaciones sugiere que puede existir multicolinealidad en el modelo, específicamente entre OBP y SLG. Para verificar que este no sea el caso realizamos tres regresiones auxiliares, dadas por:

$$OBP_i = \delta_0 + \delta_1 * ERA_i + \delta_2 * SLG_i + \mu_i \tag{3}$$

$$ERA_i = \delta_0 + \delta_1 * OBP_i + \delta_2 * SLG_i + \mu_i \tag{4}$$

$$SLG_i = \delta_0 + \delta_1 * ERA_i + \delta_2 * OBP_i + \mu_i \tag{5}$$

Ningún $R^2$ (Figura 1(r)) supera aquel de la regresión original, por lo que podemos descartar que exista multicolinealidad en el modelo. Asimismo, al calcular el Factor Inflador de la Varianza (Figura 1(s)) de las 3 regresoras del modelo elegido se observa que ninguno es mayor a 10, confirmando lo dicho anteriormente.

*3.3. Homoscedasticidad*

Se presenta el diagrama de dispersión de los residuales contra cada observación, así como el de los residuales al cuadrado (SR) contra los valores predichos (Figuras 1(t) & 1(u)).
Los gráficos parecen tener algunas observaciones atípicas, sin embargo, parecen indicar que existe homoscedasticidad en el modelo.Para confirmar lo anterior, realizamos tanto la Prueba de Breusch, Pagan y Godfrey (Figura 1(v)) como la de White. Al ser el p-value de ambas pruebas mayor a 0.05, podemos concluir al 95% de confianza que existe homoscedasticidad en el modelo.

*3.4. No autocorrelación*

Gráficamente se puede analizar tanto la dispersión de los residuales contra la observación (Figura 1(t)), así como la dispersión de los mismos contra la observacion inmediata anterior (Figura 1(w)).
Al no ser identificable algún patrón concreto parece ser que no existe autocorrelación de los residuales. Para confirmarlo realizamos las Pruebas de Durbin-Watson y de Breusch y Godfrey (Figuras 1(x) & 1(y)).
De igual forma, no se rechaza la hipótesis de que no exista autocorrelación de los residuales en el modelo.

*3.5. No endogeneidad*

De nueva cuenta se exhorta al lector a revisar la dispersión de los residuales (Figura 1(t)).
Como habíamos observado anteriormente, no parece existir un problema con los residuales. Con la Prueba RESET (Figuras 1(z)-1(ab)) nos aseguramos de que el modelo esté bien especificado, ya que no se rechaza nuesta hipótesis de que el modelo está bien especificado y por lo tanto concluimos que nuestros estimadores son MELI.

## 4. Interpretación de los resultados

Habiendo reslizado nuestras estimaciones podemos concluir que,, en promedio, un aumento de 0.20 en ERA reduce los juegos ganados por temporada de un equipo en casi 3 unidades. Por otro lado, un aumento de 0.02 en SLG en promedio hará que un equipo gane aproximadamente 3.4 partidos más. Por último, pero no por ello menos importante (aunque sí menos significativo estadísticamente), un aumento de 0.02 en OBP aumenta en promedio las victorias de un equipo en cerca de 5 unidades.
Todos nuestros resutados hacen sentido, pues es de esperarse que aumentos en ERA se traduzcan en derrotas y, por el contrario, aumentos en OBP y SLG tengan efectos positivos. Asimismo, es de esperarse que OBP tenga el mayor efecto en las victorias, pues entre más tiempo pasen los jugadores de un equipo en las bases es más factible que anoten carreras y por ende, ganen.

## 5. Conclusiones

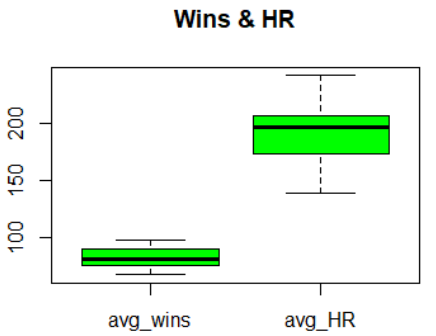
La inclusión del análisis estadístico en deportes en años recientes ha vuelto de éstos un todavía mejor espectáculo cuando se usa correctamentee. En este caso, podemos defender la idea de que aquellos equipos que centran sus objetivos en jugadores que logran estar en base del modo que sea y en aquellos con poderío a la hora de batear, seguramente tendrán éxito. Asimismo, se deben buscar lanzadores que no permitan muchas carreras por partido.

## 6. Anexos

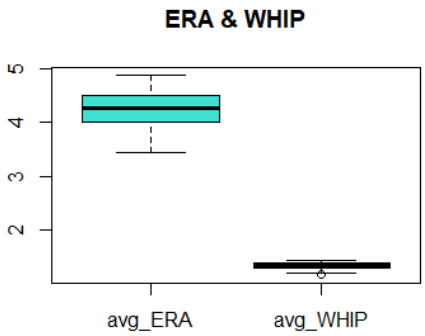
## Estadística Descriptiva

| avg_wins | avg_ERA | avg_HR | avg_OBP | avg_SLG | avg_WHIP |
|---|---|---|---|---|---|
| Min. :67.00 | Min. :3.450 | Min. :139.0 | Min. :0.3010 | Min. :0.3880 | Min. :1.160 |
| 1st Qu.:75.50 | 1st Qu.:4.010 | 1st Qu.:174.5 | 1st Qu.:0.3152 | 1st Qu.:0.4055 | 1st Qu.:1.290 |
| Median :80.50 | Median :4.270 | Median :197.0 | Median :0.3205 | Median :0.4185 | Median :1.330 |
| Mean :80.90 | Mean :4.227 | Mean :193.2 | Mean :0.3205 | Mean :0.4183 | Mean :1.321 |
| 3rd Qu.:88.25 | 3rd Qu.:4.503 | 3rd Qu.:207.0 | 3rd Qu.:0.3265 | 3rd Qu.:0.4298 | 3rd Qu.:1.370 |
| Max. :97.00 | Max. :4.870 | Max. :242.0 | Max. :0.3360 | Max. :0.4500 | Max. :1.430 |

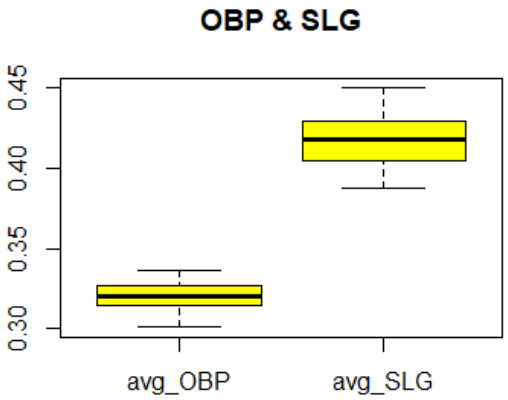
## Diagramas de Caja y Brazos




(a)

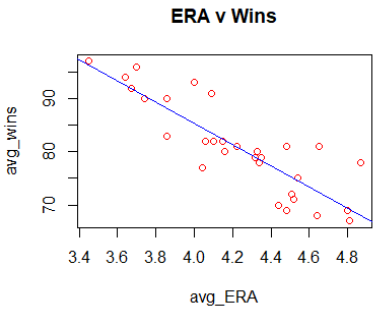



(b)

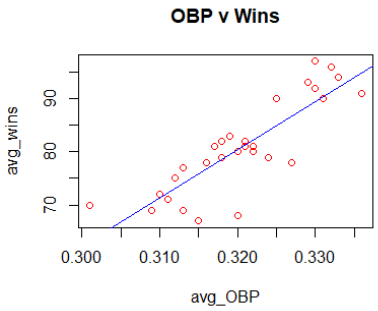



(c)

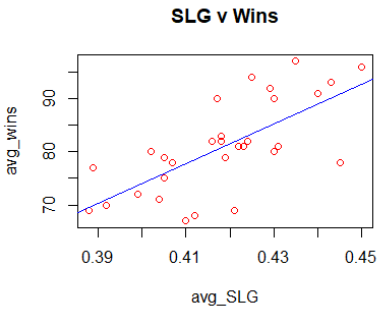
## Diagramas de Dispersión



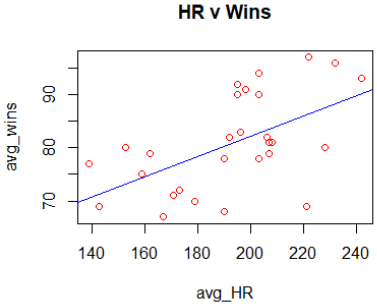

(d)



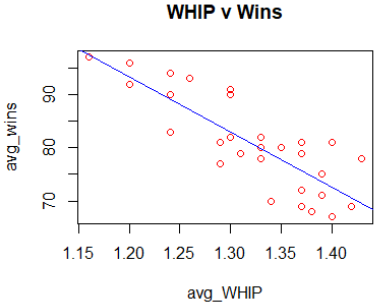

(e)




(f)

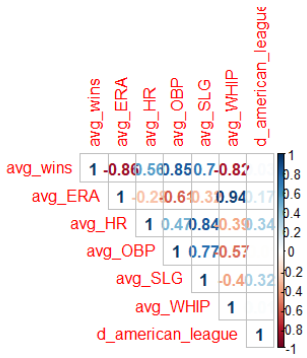



(g)




(h)

## Matriz de correlaciones




(i)

## Modelos de Regresión

| | Dependent variable: | | | |
|---|---|---|---|---|
| | avg_wins | | | |
| | (1) | (2) | (3) | (4) |
| avg_ERA | -19.938*** | -12.621*** | -14.271*** | -18.833*** |
| | (2.246) | (1.688) | (1.348) | (3.727) |
| avg_OBP | | 555.753*** | 247.842** | 198.214* |
| | | (77.604) | (91.358) | (97.757) |
| avg_SLG | | | 172.349*** | 200.082*** |
| | | | (38.756) | (43.698) |
| avg_WHIP | | | | 24.696 |
| | | | | (18.846) |
| Constant | 165.184*** | -43.883 | -10.306 | -19.338 |
| | (9.531) | (29.745) | (24.060) | (24.715) |
| Observations | 30 | 30 | 30 | 30 |
| R2 | 0.738 | 0.910 | 0.949 | 0.952 |
| Adjusted R2 | 0.728 | 0.903 | 0.943 | 0.944 |
| Residual Std. Error | 4.558 (df = 28) | 2.726 (df = 27) | 2.093 (df = 26) | 2.065 (df = 25) |
| F Statistic | 78.808*** (df = 1; 28) | 135.812*** (df = 2; 27) | 160.098*** (df = 3; 26) | 123.815*** (df = 4; 25) |
| Note: | | | | *p<0.1; **p<0.05; ***p<0.01 |

(j)

| | Dependent variable: | | |
|---|---|---|---|
| | avg_wins | | |
| | (1) | (2) | (3) |
| avg_ERA | -18.596*** | -19.938*** | -20.089*** |
| | (3.877) | (3.981) | (4.064) |
| avg_OBP | 222.081* | 254.167* | 247.463* |
| | (127.061) | (128.231) | (131.282) |
| avg_SLG | 177.437* | 150.701 | 146.609 |
| | (87.109) | (88.752) | (90.741) |
| avg_WHIP | 24.557 | 31.392 | 32.648 |
| | (19.203) | (19.762) | (20.291) |
| avg_HR | 0.011 | 0.015 | 0.013 |
| | (0.037) | (0.037) | (0.038) |
| d_american_league | | 1.155 | -9.392 |
| | | (0.927) | (22.821) |
| I(d_american_league * avg_SLG) | | | 25.107 |
| | | | (54.275) |
| Constant | -20.503 | -24.173 | -21.080 |
| | (25.470) | (25.353) | (26.650) |
| Observations | 30 | 30 | 30 |
| R2 | 0.952 | 0.955 | 0.956 |
| Adjusted R2 | 0.942 | 0.943 | 0.941 |
| Residual Std. Error | 2.104 (df = 24) | 2.080 (df = 23) | 2.116 (df = 22) |
| F Statistic | 95.471*** (df = 5; 24) | 81.653*** (df = 6; 23) | 67.627*** (df = 7; 22) |

(k)

## Regresión Auxiliar para Prueba de Significancia Parcial

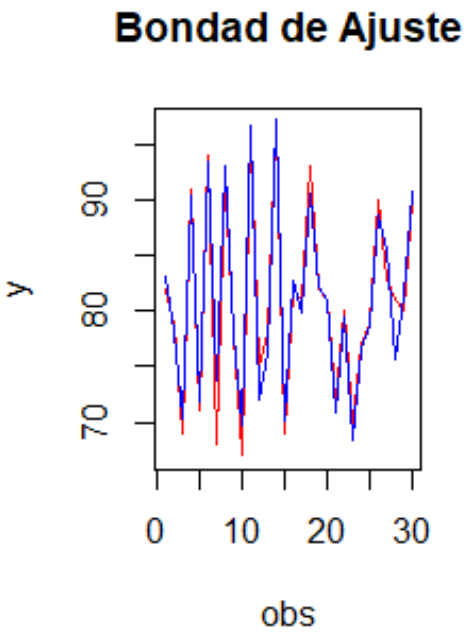
| | Dependent variable: | |
|---|---|---|
| | avg_wins | |
| | (1) | (2) |
| avg_ERA | -14.271*** | -14.353*** |
| | (1.348) | (1.361) |
| avg_OBP | 247.842** | 272.277*** |
| | (91.358) | (96.778) |
| avg_SLG | 172.349*** | 154.242*** |
| | (38.756) | (44.953) |
| I(d_american_league * avg_SLG) | | 1.733 |
| | | (2.137) |
| Constant | -10.306 | -10.584 |
| | (24.060) | (24.222) |
| Observations | 30 | 30 |
| R2 | 0.949 | 0.950 |
| Adjusted R2 | 0.943 | 0.942 |

(l)

## Intervalos de Confianza

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | -59.76113 | 39.14933 |
| avg_ERA | -17.04237 | -11.50011 |
| avg_OBP | 60.05371 | 435.63042 |
| avg_SLG | 92.68568 | 252.01214 |

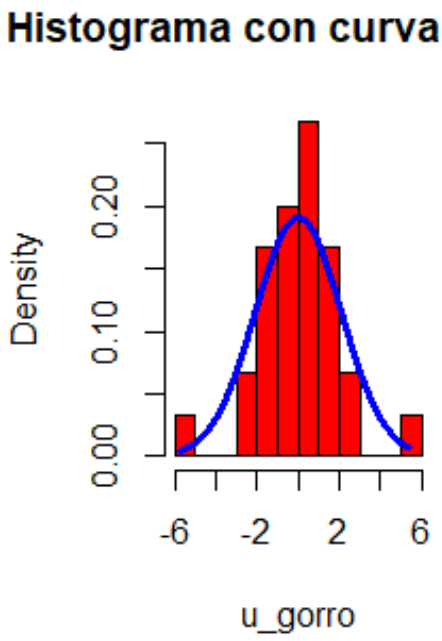
(m)




(n)

```
            Df Sum Sq Mean Sq F value   Pr(>F)
avg_ERA      1 1637.1  1637.1  373.57  < 2e-16 ***
avg_OBP      1  381.0   381.0   86.95 8.91e-10 ***
avg_SLG      1   86.7    86.7   19.78 0.000145 ***
Residuals   26  113.9     4.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(o)




(p)

```
        Jarque Bera Test

data:  u_gorro
X-squared = 5.8901, df = 2, p-value = 0.0526
```

(q)

## Regresiones Auxiliares y FIV

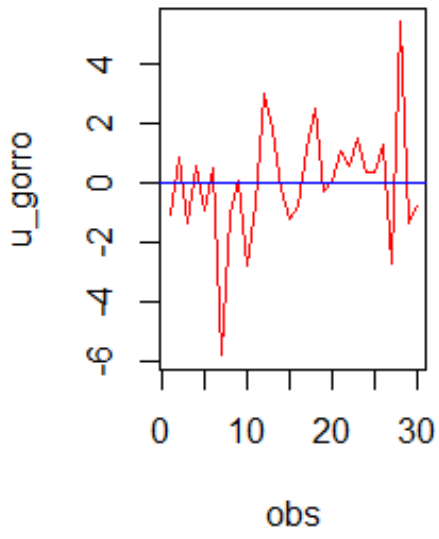
| | Dependent variable: | | |
|---|---|---|---|
| | avg_ERA (1) | avg_OBP (2) | avg_SLG (3) |
| avg_OBP | -39.866*** (10.546) | | 1.787*** (0.296) |
| avg_ERA | | -0.009*** (0.002) | 0.010 (0.006) |
| avg_SLG | 7.912 (5.319) | 0.322*** (0.053) | |
| Constant | 13.696*** (2.202) | 0.223*** (0.027) | -0.195* (0.113) |
| Observations | 30 | 30 | 30 |
| R2 | 0.415 | 0.730 | 0.619 |
| Adjusted R2 | 0.371 | 0.710 | 0.591 |
| Residual Std. Error (df = 27) | 0.299 | 0.004 | 0.010 |
| F Statistic (df = 2; 27) | 9.557*** | 36.572*** | 21.926*** |
| Note: | | *p<0.1; **p<0.05; ***p<0.01 | |

(r)

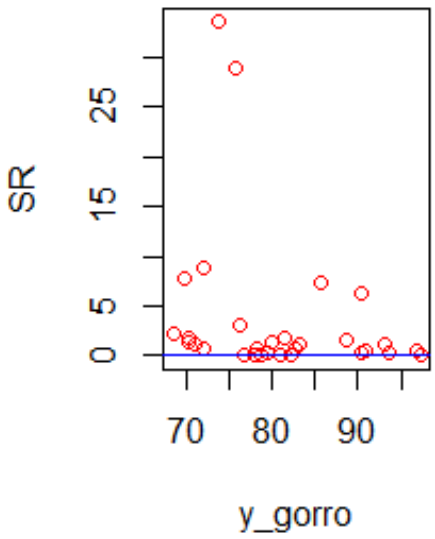
```
avg_ERA  avg_OBP  avg_SLG
1.707960 3.709067 2.624170
```

(s)

**Homoscedasticidad**




(t)



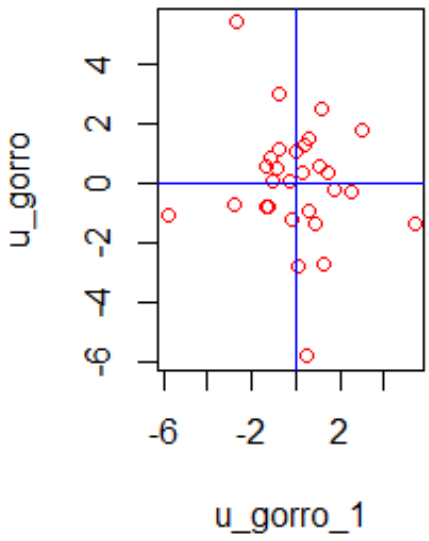

(u)

```
        studentized Breusch-Pagan test

data:  best_mod
BP = 4.945, df = 3, p-value = 0.1759
```

(v)

## Pruebas de Autocorrelación




(w)

```
        Durbin-Watson test

data:  best_mod
DW = 2.2457, p-value = 0.7948
alternative hypothesis: true autocorrelation is greater than 0
```

(x)

```
        Breusch-Godfrey test for serial correlation of order up to 1

data:  best_mod
LM test = 0.61417, df = 1, p-value = 0.4332
```

(y)

## Pruebas RESET (Power=2,3,2:3)

```
        RESET test

data:  best_mod
RESET = 0.17967, df1 = 1, df2 = 25, p-value = 0.6753
```

(z)

```
        RESET test

data:  best_mod
RESET = 0.00010036, df1 = 1, df2 = 25, p-value = 0.9921
```

(aa)

```
        RESET test

data:  best_mod
RESET = 0.12931, df1 = 2, df2 = 24, p-value = 0.8793
```

(ab)