Tutorial de Análisis de Regresión Lineal en R

Existen varios tipos de modelos predictivos que son realmente útiles para pronosticar eventos.

Dentro del mundo de los analistas, la mayoría de las veces se cumple el siguiente patrón:

1 – Disponer de la información

2 – Adaptar la información de acuerdo a nuestra conveniencia

3 – Realizar un análisis exploratorio para tener en cuenta de lo que estás tratando

4 – Utilizar un modelo adecuado y validar resultados/Mostrar la información en un formato adecuado

La regresión lineal es uno de los modelos más comunes y más utilizados por los analistas dentro de los **algoritmos de aprendizaje automático supervisado** (O mejor dicho Supervised Machine Learning Algorithms).

Muy probablemente has escuchado al menos una vez sobre el concepto de "Machine Learning". Si es que no sabes qué es, no te preocupes.

A grandes rasgos, Machine Learning significa crear sistemas o técnicas para que las computadoras aprendan automáticamente.

Y cuando hablamos de Supervised Machine Learning, significa que tanto datos de entrada como de salida son proporcionados al modelo.

Cuando hablamos de regresión lineal, tenemos que tener en mente que es un *"método matemático que modeliza la relación entre una variable dependiente Y, las variables independientes Xi y un término aleatorio ε."*

Y dentro de la regresión lineal, se encuentran dos tipos de modelos:

- **Regresión lineal simple:** Sólo una variable independiente se usa para la variable dependiente.
- **Regresión lineal múltiple:** Más de una variable independiente se usa para la variable dependiente.

Para este tutorial se utilizará Regresión Lineal Múltiple, donde la variable respuesta (Dependiente) será los **puntos acumulados** que los equipos obtienen a lo largo de las temporadas. (Las variables independientes serán presentadas más adelante)

A la hora de realizar un estudio, generalmente se ejecutan ciertas hipótesis.

¿Qué es eso una Hipótesis?

En la estadística, se usan las famosas "hipótesis" donde básicamente la persona tiene una teoría de cualquier evento/situación/objetivo, y desea determinar si es que su información valida la teoría.

Todas las **hipótesis tienen una hipótesis nula y una hipótesis alternativa.** En este tutorial podría ser de la siguiente forma:

- **Hipótesis nula (Ho):** No hay relación entre las variables independientes y la dependiente.
- **Hipótesis alternativa (Ha):** Existe una relación entre las variables independientes y la dependiente.

Y nosotros aprovecharemos la regresión lineal para probar la hipótesis.

Por cierto, lo importante de reconocer en las hipótesis es que **el objetivo NO es enseñar que la Hipótesis Alternativa es probablemente verdadera,** ya que el **verdadero objetivo es demostrar que la hipótesis nula es probablemente falsa.**

Otra forma de verlo:

Imagina que una prueba de hipótesis es como si fueras a señalar culpable a un criminal. La hipótesis nula es el acusado, la persona que está haciendo el análisis es el demandante, y la prueba estadística es el juez.

Tal como si fuera un juicio, existe una suposición de que es inocente.

Es decir, la hipótesis nula se considera verdadera a menos que tú, el analista, pueda probar que sea falsa

El **analista es libre de hacer cualquier tipo de experimento con el fin de maximizar las posibilidades de hacer falsa la hipótesis nula,** y tu único obstáculo es que la prueba estadística pone las reglas de la prueba, ya que esas reglas son para **proteger al criminal (La hipótesis nula).**

¿Y por qué proteger tanto al criminal? Simple…Las reglas son específicamente para garantizar que si la hipótesis nula es realmente cierta, las posibilidades de una falsa condena sean bajas.

Después de todo, la hipótesis nula digamos que no tiene un abogado, y porque el analista quiere desesperadamente probar que es falso, alguien tiene que protegerlo.

Y si rechazamos la hipótesis nula cuando realmente es verdadera, se comete el **Error Tipo I**. Para más información, te invito a que leas un poco sobre los tipos de errores en las hipótesis.

Regresión Lineal

Matemáticamente se puede escribir la ecuación de la regresión lineal de la siguiente manera:

Y ≈ β0 + β1X + ε

Donde:

- Y es la variable respuesta
- X es la variable explicativa (Independiente)
- Bo es el coeficiente que representa el intercepto del modelo (O donde cruza en el eje y)
- B1 es el coeficiente que representa la pendiente (Nos da la información necesaria de la inclinación de la línea y su dirección)
- ε es el error que no podemos capturar en el modelo (Lo que X no nos puede decir de Y)
-

La fórmula anterior es para una regresión lineal simple, y para la múltiple en realidad es casi lo mismo, pero con más variables independientes y sus respectivos coeficientes:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

Hablando un poco del término *"Error",* todos sabemos que son una parte inevitable al momento de predecir. No importa qué tan poderoso sea el algoritmo, siempre habrá un "Error" que nos recuerde que el futuro es impredecible.

Aun así, nosotros los humanos tendemos a ser tercos y ser perseverantes, o en otras palabras, conocemos la presencia del error, ¿pero podemos reducirlo?

Precisamente en regresión lo que se aplica es la técnica de Mínimos Cuadrados Ordinarios. Conceptualmente, lo que hace este método es reducir la suma de los errores al cuadrado e intenta encontrar el mejor valor posible para los coeficientes de regresión (Bo, B1, etc.)

Y el cómo se calcula esta técnica la verdad no es nada del otro mundo. Simplemente se hace de la siguiente forma:

$$\sum[Actual(y) - Pronosticado(y')]^2$$

Todo parece maravilloso con la regresión lineal, pero lamento informarte que no todo es de color de rosa debido a que, para aplicar el método, se tienen que **hacer suposiciones**:

1. La **variable dependiente y los errores deben de tener una distribución normal**. Hay una gran cantidad de tipos de distribuciones en la rama de la estadística, y la que se tiene que presentar es la normal.

2. Los **errores no deben de estar correlacionados**. Si es que lo hay, el concepto se le conoce como autocorrelación, y afecta bastante a los coeficientes de regresión.

3. Debe de existir una relación lineal y aditiva entre la variable dependiente y las variables independientes. Cuando hablamos de Lineal, significa que el cambio de la variable dependiente (**Y**) por una unidad de cambio en la variable independiente (X) es constante. Y cuando hablamos de aditiva, significa que al **efecto de la variable independiente (X) en la variable dependiente (Y) es independiente de otras variables**.

4. No debe de **existir correlación ENTRE las variables independientes**. Si hay presencia, se vuelve muy complicado determinar los efectos verdaderos de las X's con la Y. (A eso se le llama multicolinealidad).

5. Los errores deben tener una variabilidad constante, y si es que no lo hay, a eso se le llama heteroscedasticidad.

Estas suposiciones hacen que la regresión lineal no pueda ser aplicada en cualquier situación que nos plazca, por lo que se vuelve algo restrictiva.

Análisis Exploratorio

Para iniciar, primero tenemos que ver si hay una relación entre la variable respuesta y las variables explicativas.

Y por ello vamos a realizar un poco de análisis exploratorio.

La información que vamos a utilizar proviene de la página http://www.football-data.co.uk/

Y los datos recopilados son resultados de partidos de las 5 grandes ligas de Europa desde la temporada 2013/2014 hasta la temporada 2017/2018.

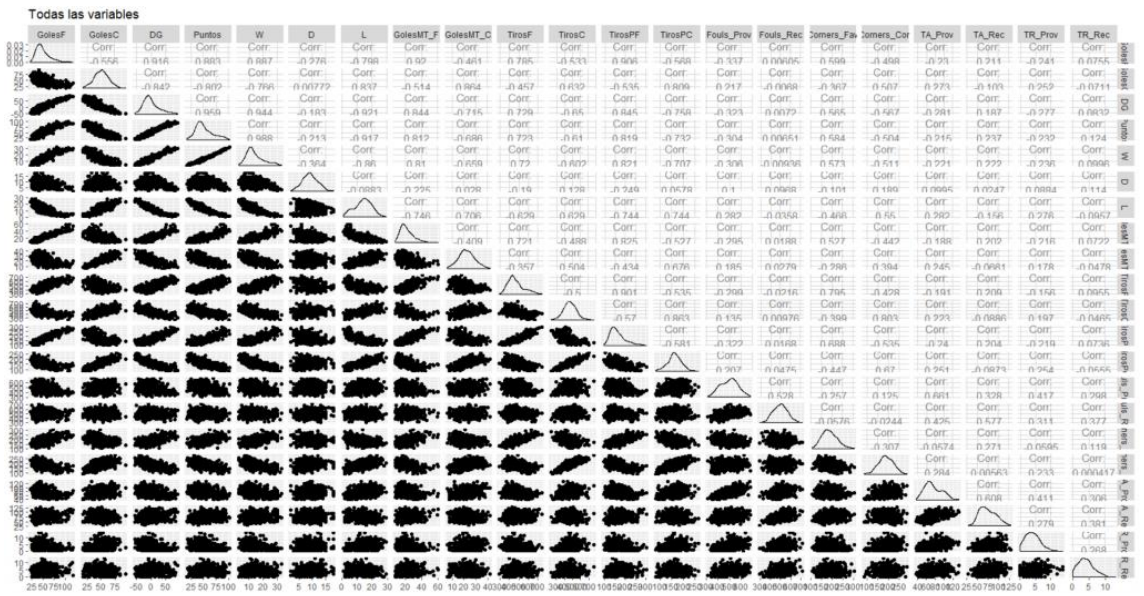
Significado de cada columna.

- **Equipo**: Equipos de las 5 grandes ligas de Europa
- **GolesF:** Goles a Favor
- **GolesC:** Goles en Contra
- **DF**: Diferencia de Goles

- **Puntos:** Puntos acumulados a final de temporada
- **W**: Win (Victorias)
- **D**: Draw (Empates)
- **L**: Lost (Derrotas)
- **GolesMT_F:** Goles 1ª Parte a Favor
- **GolesMT_C:** Goles 1ª Parte en Contra
- **TirosF:** Tiros a Favor
- **TirosC:** Tiros en Contra
- **TirosPF**: Tiros a Puerta a Favor
- **TirosPC**: Tiros a Puerta en Contra
- **Fouls_Prov**: Faltas Provocadas
- **Fouls_Rec**: Faltas Recibidas
- **Corners_Fav**: Corners a Favor
- **Corners_Con**: Corners en Contra
- **TA_Prov**: Tarjetas Amarillas Provocadas
- **TA_Rec:** Tarjetas Amarillas Recibidas
- **TR_Prov**: Tarjetas Rojas Provocadas
- **TR_Rec:** Tarjetas Rojas Recibidas

Vamos a utilizar la función ***ggpairs()*** del paquete ***GGally*** para crear una matriz de correlación.

```
#GGpairs
library(GGally)
ggpairs(data = data, columns = 2:ncol(data), title = "Todas las variables")
```

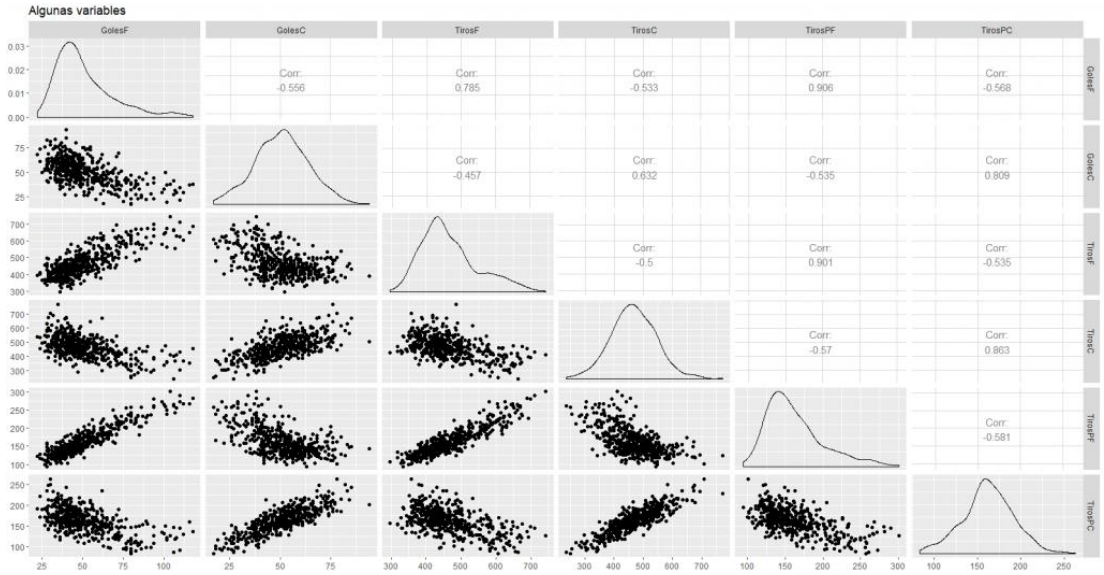


Vemos que esta función no es la mejor debido a que son demasiadas variables. Lo que podemos hacer por el momento es excluir algunas y ver por divisiones esta gráfica.

Usamos los ***Puntos, Goles a Favor (GolesF), Goles en Contra (GolesC), Tiros a Favor (TirosF), Tiros en Contra (TirosC), Tiros a Puerta a Favor (TirosPF), y Tiros a Puerta en Contra (TirosPC)***, obtenemos lo siguiente:

```r
#Seleccionar columnas
columnas_seleccionadas <- data %>%
  select(Puntos, GolesF, GolesC, TirosF, TirosC, TirosPF, TirosPC)

#GGpairs
ggpairs(data = columnas_seleccionadas,
        columns = 2:ncol(For_ggpairs),
        title = "Algunas variables")
```



Para las correlaciones, si se acerca a 0 es que no están correlacionadas, si se acerca a 1 es que están muy correlacionadas. (Aplica tanto para positivo como negativo, relación directa 1 inversa -1)

Observando esta gráfica, podemos ver que la correlación que existe entre **TirosF** y **GolesF** es de **0.785**, ¿por lo que tiene sentido no crees? A más tiros, más goles.

TirosC y **GolesC** tienen una correlación de 0.632, por lo que no están tan correlacionadas. ¿Qué podría estar afectando? ¿Será que los defensas afectan? ¿Qué tal el portero?

En realidad, es algo tedioso estar viendo la correlación de esta forma.

Para nuestra fortuna, existe otra forma para ver la correlación de las variables de forma efectiva.

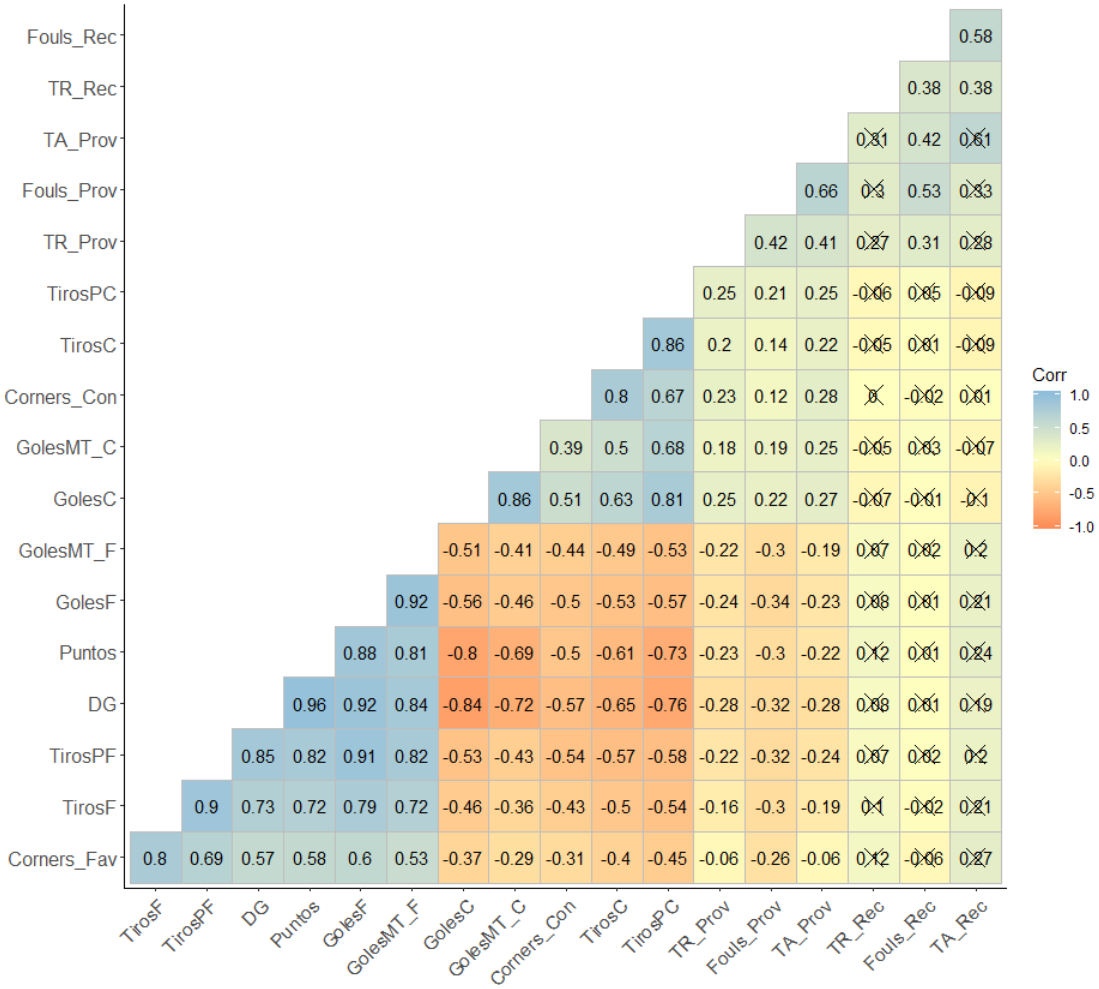
Vamos a manejar dos paquetes, **corrr** y **ggcorrplot** para realizar lo siguiente:

```r
#Utilizando corrr, quitar variables y prepararlo para una matriz
correlate(data[,-1], diagonal = 1) %>%
  focus(-W, -D, -L, mirror = TRUE) %>%
  remove_rownames() %>%
  column_to_rownames(var = "rowname") -> M

#Crear matriz y valores P
cor_matrix <- as.matrix(M)
p_mat <- cor_pmat(cor_matrix)

#ggcorrplot
ggcorrplot(cor_matrix, p.mat = p_mat, hc.order = TRUE,
           type = "lower",
           lab = TRUE,
           lab_size = 4,
           ggtheme = ggplot2::theme_classic,
           colors = brewer.pal(n = 3, name = "RdYlBu"))
```

Se eliminó W, D, L que son Win (Victoria), Draw (Empate) y Lost (Derrota) porque son variables que consideramos que no son necesarias.



De esta forma se puede ver claramente las correlaciones entre las variables.

Se puede observar que la **Diferencia de Goles (DG) tiene 0.96 de correlación con los puntos**, mientras que **Puntos y GolesC tiene -0.8 de correlación**.

Los cuadros que están en cruz son prácticamente correlaciones nulas de acuerdo a los valores p calculados anteriormente.

En fin, podemos ver claramente que hay varias variables que tienen demasiada correlación, y de hecho, usualmente una correlación arriba de 80% (Subjetivo) es considerado demasiado alto.

Por ello, para efectos de este tutorial vamos a dejar algunas variables y también vamos a eliminar otras.

Creación del Modelo:

Ahora, vamos efectuando la regresión lineal múltiple:

```
#Correr modelo
Modelo <- lm(Puntos ~ GolesF + GolesC + TirosF + TirosC + TirosPF + TirosPC
             + Fouls_Prov + Fouls_Rec + Corners_Fav + Corners_Con, data = data)

summary(Modelo)
```

```
> summary(Modelo)

Call:
lm(formula = Puntos ~ GolesF + GolesC + TirosF + TirosC + TirosPF +
    TirosPC + Fouls_Prov + Fouls_Rec + Corners_Fav + Corners_Con,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-11.346  -3.564  -0.021   2.800  16.179

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.3809842  3.3780039  11.954  < 2e-16 ***
GolesF       0.6073782  0.0300228  20.231  < 2e-16 ***
GolesC      -0.5585848  0.0298247 -18.729  < 2e-16 ***
TirosF       0.0007079  0.0075199   0.094 0.925045
TirosC       0.0046219  0.0069950   0.661 0.509094
TirosPF     -0.0008715  0.0205197  -0.042 0.966141
TirosPC     -0.0470568  0.0199693  -2.356 0.018857 *
Fouls_Prov   0.0041093  0.0038278   1.074 0.283579
Fouls_Rec   -0.0001875  0.0037474  -0.050 0.960109
Corners_Fav  0.0263467  0.0103662   2.542 0.011353 *
Corners_Con  0.0365340  0.0108629   3.363 0.000833 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.712 on 473 degrees of freedom
Multiple R-squared:  0.9266,    Adjusted R-squared:  0.925
F-statistic:   597 on 10 and 473 DF,  p-value: < 2.2e-16
```

Si en la fórmula del modelo escribes *“Puntos ~ .”*, la fórmula usará todas las variables independientes disponibles.

Interpretación del resultado de la regresión:

- **Intercept**: Este valor es la interceptación, o el Bo explicado anteriormente. El coeficiente es una predicción hecha por el modelo cuando todas las variables restantes son igual a 0.
- **Estimate**: Estos valores representan los **coeficientes de regresión para cada respectiva variable**, y además representa el valor de la pendiente. Por ejemplo, interpretando el coeficiente de **GolesC**, **cuando esta variable es incrementada por 1** (manteniendo las otras variables constantes) **Puntos disminuye por -0.5585848**.
- **Std. Error**: Este valor representa la variabilidad asociada con la estimación. **Si es menor este valor, más efectiva será la predicción**.
- **t value**: El estadístico t se usa para determinar lo importante que es una variable, o en otras palabras, **si una variable está agregando significativamente información al modelo**.
- **Pr(>|t|):** Este valor es el **famoso valor p**, definido como la probabilidad de observar cualquier valor igual o mayor que t si es que Ho es verdadera. Entre más grande sea el

valor t, menor será el valor p. Generalmente se usa 0.05 como límite para darle significado, y **si los valores p son menores a 0.05, se rechaza Ho**.

Si nos fijamos en el valor p, entonces de acuerdo con el modelo, los coeficientes importantes son **GolesF, GolesC y Corners_Con (Corners en contra)**.

¿Por qué los córneres en contra afectan los puntos? ¿Acaso es mejor estar defendiendo córneres para contraatacar?

¿Por qué los tiros a favor no son importantes? Hace poco vimos que los tiros a favor y los goles a favor se correlacionaban y por ende son importantes para obtener puntos.

Más adelante explicaremos esto. Por el momento sigamos con la interpretación del modelo.

Hay **otros resultados que son interesantes al momento de evaluar la regresión**:

- **Residual Standard Error:** Este valor representa la cantidad promedio que nuestra variable respuesta se desvía de nuestro modelo lineal.
- **Coeficiente de Determinación (R-squared):** Este valor representa como de cerca nuestra información se encuentra al modelo de regresión lineal. También se puede interpretar como el **porcentaje de variabilidad en la variable respuesta que es explicado por las variables independientes**. Su rango varía entre 0-1.
- **Degrees of Freedom:** Este valor representa el número de valores independientes de información que fueron usados para calcular un estimado. Si quieres verlo de otra forma, digamos que “a + b = 50”, y tomando en cuenta esta ecuación, la definición podría ser los elementos que puedes cambiar para que 50 permanezca igual, y por ello sólo “a” puedes cambiar.
- **Adjusted R-squared:** Este valor sirve para ver si una nueva variable agregada al modelo realmente afecta a la ecuación de regresión. Con R-squared, el valor sigue aumentando al igual que vas agregando nuevas variables. (Independientemente si es que afectan o no)
- **F-statistic:** Este valor sirve para informarnos si es que hay relación entre las variables independientes y dependiente. Técnicamente evalúa el significado global del modelo, ya que es la proporción explicado por la variabilidad entre la variabilidad no explicada. (Modelo completo vs sólo intercepto)
- **p-value:** Este valor está relacionado con F-statistic, y es el que se utiliza para interpretar todo el modelo.

Por cierto, hay otras métricas que son cruciales a la hora de la evaluación pero que no se encuentran en resultado, y son las métricas de errores (Como son errores, a menor valor, mejor el modelo)

- **MSE (Mean Squared Error):** Este valor tiende a amplificar los valores atípicos del modelo. Por ejemplo, si el valor actual es 40 y el valor pronosticado es 80, el resultado sería (80-40) ² = 1,600
- **MAE (Mean Absolute Error):** Este es lo opuesto al MSE. Tomando el ejemplo anterior, el resultado sería (80-40) = 40
- **RMSE (Root Mean Square Error):** Este valor te informa qué tan lejos en promedio los residuales están del 0. El hecho de que tenga raíz cuadrada hace que proporcione el resultado en unidades originales.

Tanto **R-squared** como **RMSE** son números interesantes al momento de interpretar el modelo, aunque RMSE explícitamente sabe en que medida las predicciones se desvían, en promedio, de los valores actuales en la información analizada.

Entonces…

Tomando en cuenta el resultado de la regresión, ¿podemos concluir que es buen modelo para explicar los puntos obtenidos?

Tiene una **R-squared de 0.9266,** el valor **p es menor a 0.05**, el **Residual Standard Error es de 4.712**…¿Será que está bien?

¿Eso quiere decir que los puntos pueden ser explicados de la siguiente forma?

Puntos = 40.39 + 0.608 * GolesF – 0.559 * GolesC + 0.00071 * TirosF + 0.004 * TirosC – 0.0008 * TirosPF – 0.048 * TirosPC + 0.004 * Fouls_Prov – 0.0001 * Fouls_Rec + 0.027 * Corners_Fav + 0.037 * Corners_Con

Para empezar, a simple vista…¿Cómo puede ser que entre menos tiros a puerta a favor (TirosPF) te genera más puntos?

¿Serán necesarios todas las variables para la variable dependiente Puntos?

De acuerdo al resultado de regresión, hay más de una variable que rechaza la hipótesis alternativa. (Hay relación entre la variable respuesta y la variable explicativa)

¿Podrá mejorar el modelo al eliminar las variables indicadas? Eso se mencionara más adelante pero por el momento quiero que pongas atención a lo siguiente:

¿Recuerdas que te mencioné que se tienen que cumplir ciertas suposiciones en el modelo para poder afirmar con certeza de que la regresión lineal sí aplica?

Te tengo una buena noticia…Hay una forma para evaluar si las suposiciones fueron desobedecidas

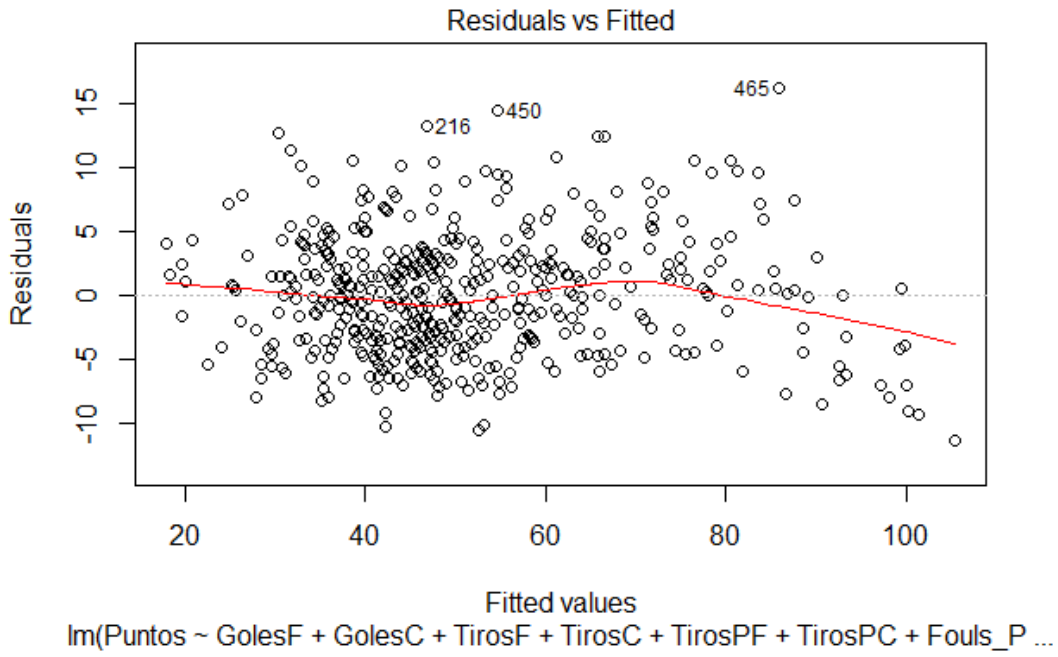
¿Cómo saber si se cumplieron esos supuestos?

Una vez que esos supuestos no se cumplan, la regresión hace predicciones erradas y sesgadas.

La forma rápida es ver un diagnóstico de gráficas llamado *"gráficas de residuales"*, y las más importantes son las siguientes:

1 – Residual vs Fitted: Esta gráfica no debe de mostrar ningún patrón, pero si es que vez uno *tipo U o curvo*, muy probablemente presenta *no linealidad.* Si es que ves un patrón de embudo, quiere decir que los datos sufren de heteroscedasticidad.

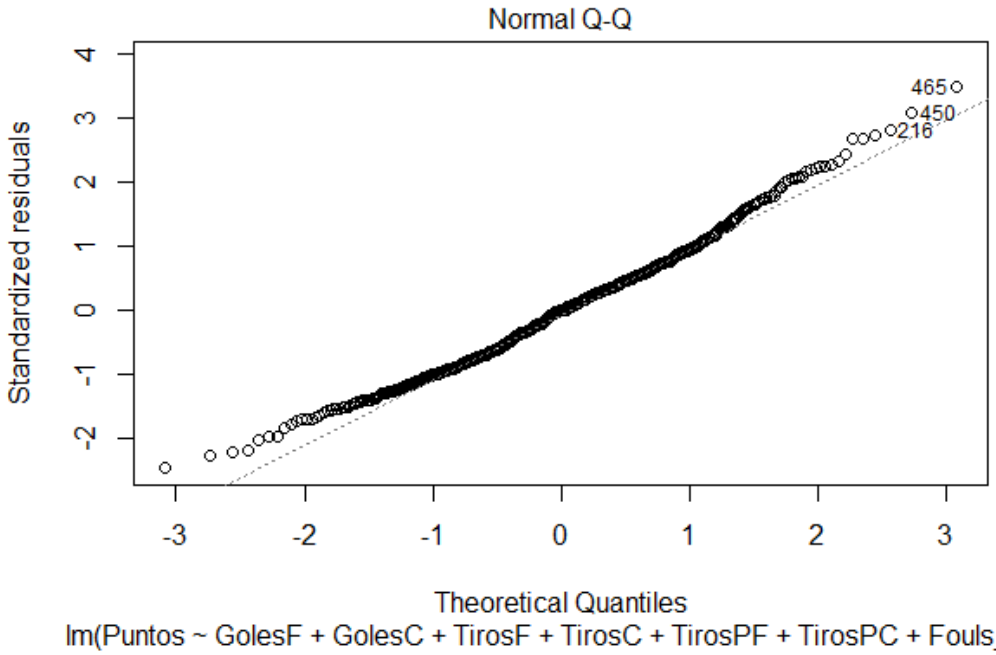
```
#Residuals vs Fitted
plot(Modelo, which=1)
```



La línea roja en la gráfica debe ser recta y horizontal, no curva, si es que la suposición de linealidad se cumple. Al final se percibe que la línea tiende a ir hacia abajo, y los valores se concentran más al inicio que al final. Se supone que debe de ser constante la variabilidad en toda la imagen, así que para estar seguros usaremos otra prueba.

2 – Normality Q-Q Plot: Esta gráfica es para determinar la distribución normal de los errores. Se debería de ver una línea derecha, y si es que observas una línea curva o distorsionada, entonces tus residuales no siguen una distribución normal.

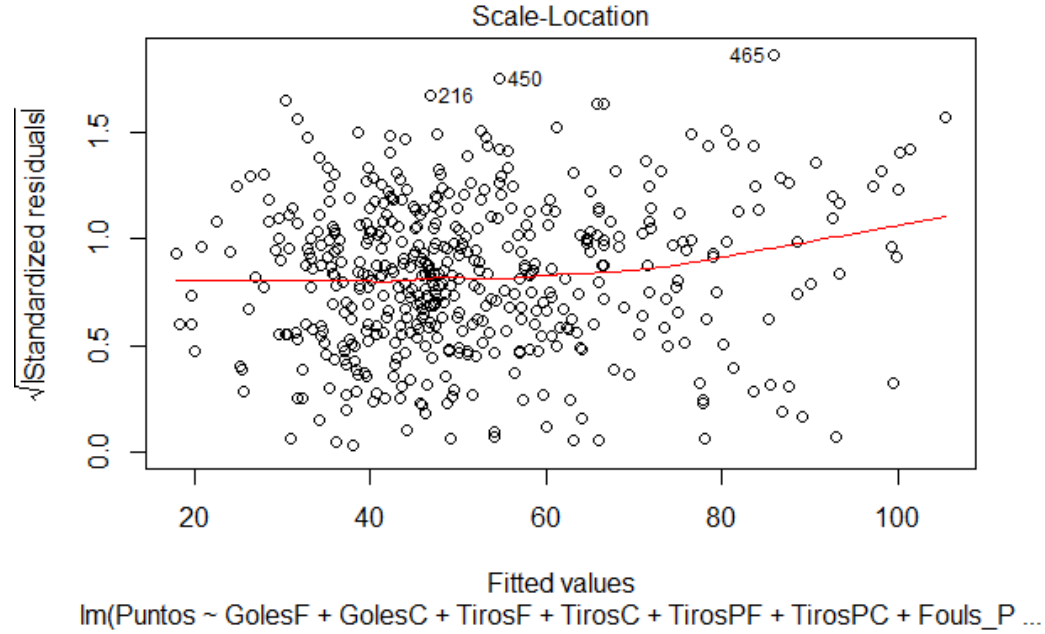
```
#Normality Q-Q plot
plot(Modelo, which = 2)
```



Tanto en la gráfica anterior como en esta, se registraron tres valores que tienen residuales grandes (observación **216, 450 y 465**), y parece dar la sensación de que tanto al inicio como al final de los puntos no están alineados. Por el momento supongamos que si se cumple.

3 – Scale-Location Plot: Esta gráfica es útil para la heteroscedasticidad. No debería de tener ningún patrón, y si es que lo tiene, existe la misma.

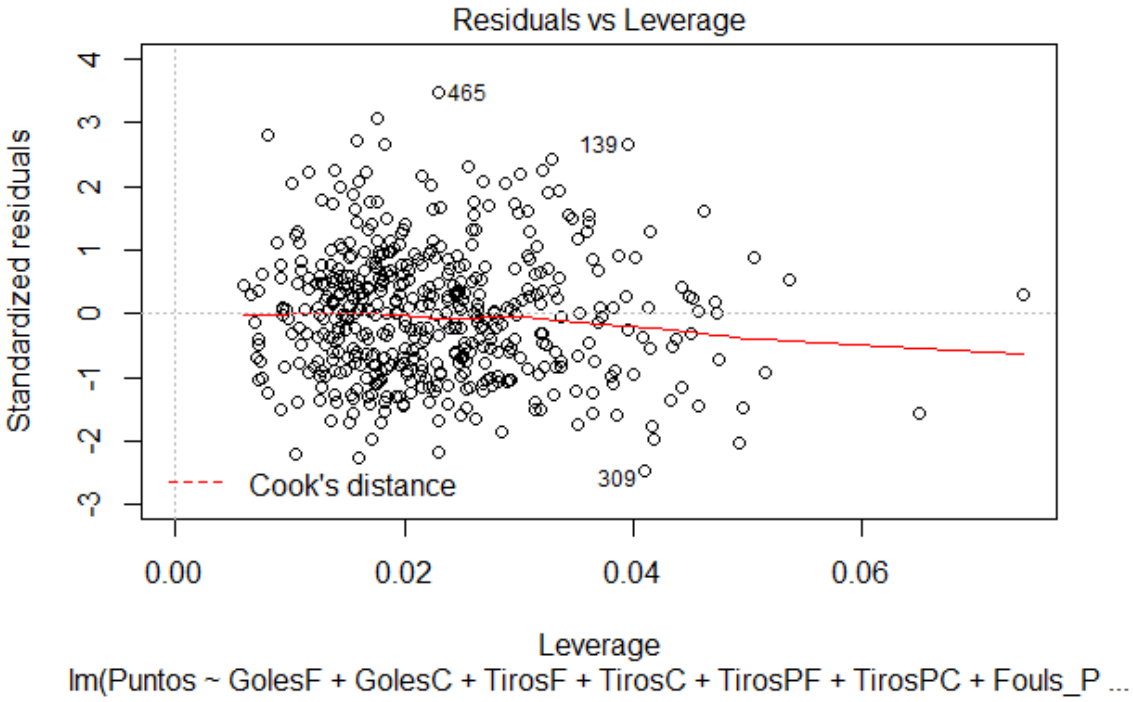
```r
#Scale-Location
plot(Modelo, which = 3)
```



Parece que hay un patrón en donde la mayoría de los datos se condensan entre 40-60. Además, la línea roja NO es plana ni horizontal, de hecho tiende a ir hacia arriba y por ende tendremos que ver de otra forma si es que se cumple dicha suposición (No es del todo claro).

4 – Residuals vs Leverage Plot: Esta gráfica es útil para ver datos influyentes. Hay que tener en cuenta que los datos atípicos pueden o no pueden afectar, por lo que hay que tener mucho cuidado al momento de interpretarlos (Podrían alterar el resultado).

```
#Residual vs Leverage
plot(Modelo, which = 5)
```



No se presenta la línea "Cook's distance" en la gráfica (Línea roja discontinua), por lo que podemos concluir que NO hay datos atípicos.

También se pueden hacer algunas pruebas sin gráficas para determinar si se cumplieron las suposiciones:

1 – Variance Inflation Factor: O mejor conocido como VIF, esta métrica sirve para ver la multicolinealidad. Si es menor a 4, no presenta multicolinealidad, y si es mayor a 10 entonces presenta muy alta multicolinealidad.

```r
#VIF
library(car)
vif(Modelo)  #Multicolinearity
```

```
> vif(Modelo)  #Multicolinearity
     GolesF      GolesC      TirosF      TirosC     TirosPF     TirosPC  Fouls_Prov   Fouls_Rec Corners_Fav
   6.093132    3.317710    8.150668    6.531727   12.966762    7.594731    1.693339    1.504711    2.796048
Corners_Con
   3.028763
```

¿Recuerdas que anteriormente mencioné sobre los tiros a puerta si eran importantes para los puntos? Bueno, una de las razones por la cual el modelo está arrojando estos resultados de importancia de variables es porque hay multicolinealidad en más de una variable, y por ello se rompe la suposición de NO correlación entre variables!

2 – Breush-Pagan test: Esta prueba es para ver la presencia de la heteroscedasticidad. Si el valor p < 0.05, rechazamos la hipótesis nula y puedes suponer que si hay heteroscedasticidad.

```r
# Breush-Pagan test
library(car)
ncvTest(Modelo)
```

```
> ncvTest(Modelo)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 12.57378, Df = 1, p = 0.0003912
```

Si es que tenías duda de heteroscedasticidad con las gráficas residuales, entonces con este método podemos comprobar que p < 0.05, rechazamos la hipótesis nula, y concluimos que si lo hay!

3 – Durbin Watson Statistic: Esta prueba es para ver si hay autocorrelación en los residuales, su rango es entre 0 – 4. Si el valor es 2 (O se aproxima mucho a 2), entonces no hay autocorrelación, y si el valor p está más cerca de 0, significa que se puede rechazar la hipótesis nula (No hay autocorrelación en los residuales, o son independientes)

```
#Durbin-Watson
library(car)
durbinWatsonTest(Modelo)  #Autocorrelation residuals
```

```
> durbinWatsonTest(Modelo)  #Autocorrelation residuals
 lag Autocorrelation D-W Statistic p-value
   1       0.1760339      1.643918       0
 Alternative hypothesis: rho != 0
```

El valor p es 0, y el valor D-W no se acerca a 2, así que se rechaza la hipótesis nula, y se infiere que hay autocorrelación.

¿Si ya hicimos esas pruebas, ahora qué sigue?

Puedes mejorar tu modelo si eso deseas.

La primera opción ya fue mencionada anteriormente como pregunta, y fue el seleccionar variables de acuerdo a los valores p. (Si valor p > 0.05, podemos quitar la variable)

Sin embargo, tienes que volver a verificar las suposiciones con las técnicas anteriores, ya que seguirá arrojando resultados sesgados y errados.

También hay otras maneras para mejorar la efectividad de tu modelo. Lo más recomendable es usar algoritmos diferentes (Como Decision Trees), pero si quieres saber cómo mejorar dentro de la regresión lineal:

- Si hay multicolinealidad, puedes usar una matriz de correlación (Ya se hizo al principio!) para ver las variables que sufren demasiada correlación.
- Si hay heteroscedasticidad, puedes transformar tu variable dependiente usando logaritmo, raíz cuadrada, etc.
- Si hay no linealidad, transforma tus variables independientes usando logaritmo, raíz cuadrada, etc.
- Si hay datos atípicos, es recomendable quitarlos siempre y cuando sean considerados de es forma. Supongamos que en un partido hubo más de 30 goles. ¿Será que ese valor se repite varias veces, o fue una ocasión especial? Si es que están esos datos, probablemente sesgen tus coeficientes y también afecten tu suma de cuadrados.

Y por último, si quieres checar varios supuestos por medio de un comando, existe un paquete llamado ***gvlma*** *package:*

```r
#gvlma package
library(gvlma)
gvmodel <- gvlma(Modelo)
summary(gvmodel)
```

```
ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma::gvlma(x = Modelo)

                     Value   p-value                   Decision
Global Stat        18.74388 0.0008824 Assumptions NOT satisfied!
Skewness           12.57398 0.0003912 Assumptions NOT satisfied!
Kurtosis            0.04685 0.8286328    Assumptions acceptable.
Link Function       5.78074 0.0162027 Assumptions NOT satisfied!
Heteroscedasticity  0.34231 0.5585000    Assumptions acceptable.
```

De acuerdo a la explicación de este enlace:

- **Global Stat:** El rechazo de la hipótesis nula ($p < 0.05$) indica que hay una no linealidad entre una o más de tus variables independientes con la variable dependiente.
- **Skewness:** La distribución está sesgada positiva o negativamente, necesitando una transformación para la normalidad. Si es que no se cumple, se debería de transformar la data.
- **Kurtosis:** Si es que hay picos muy altos o superficiales, entonces se rechazaría Ho y tendrías que transformar.
- **Link Function:** ¿Tu variable dependiente es continua o categórica? Si rechazas Ho, se debería de usar otra alternativa para hacer este análisis.
- **Heteroscedasticity:** Como ya antes visto, la varianza no es constante en el rango de las X's.

¿Entonces existe o no existe la heteroscedasticidad?

Una prueba me arroja algo mientras que otra me arroja otro resultado, ¿qué es lo que está pasando?

Probablemente lo que está sucediendo aquí es que son muy pocas observaciones para el modelo.

Como la ley de los grandes números, entre más información tengas, más se acerca el dato a la realidad.

También probablemente lo que esté pasando es que la información analizada no está en la forma más correcta para usarlo como modelo. En este caso se utilizaron acumulados de puntos, tiros, goles, etc. de varias temporadas en vez de usar datos de partido a partido.

Conclusión

Implementar regresión lineal en R puede ser fácil de ejecutar si es que se cumplen todas las suposiciones mencionadas anteriormente. Es necesario tomar en cuenta todos los factores posibles para ejecutar lo mejor posible esta técnica.

Tal vez sea tedioso al principio (Me refiero a ver todas las pruebas, evaluar variables, correlación, etc.), pero con el tiempo y si es que lo sigues practicando, será muy sencillo detectar si es que te es útil o no la regresión lineal.

Hay que recordar que en el caso de las hipótesis no siempre es intentar rechazar la hipótesis nula. De hecho, la mayoría de las investigaciones y reportes siempre hay validación de hipótesis alternativa y NO debería de ser así. (Tal vez en otro artículo lo escriba)

Es fundamental tener presente que NO toda la información está preparada para realizar un análisis. Para agregar, no siempre los resultados serán como lo esperabas, y tendrías que evaluar diferentes opciones para seguir estudiando o para concluir dicha investigación.