Tema 3. Métodos Estadísticos Aplicados.

Sumario:

1. Introducción.
2. Regresión Lineal Múltiple. Supuestos.
3. Problemas Especiales de la Regresión.
4. Selección de Variables.

Bibliografía:

- INTRODUCTORY STATISTICS PREM S. MANN (Capítulo 14)

1. Introducción.

En la regresión Lineal simple tenemos una variable dependiente y una independiente. Pero en la práctica es más común tener una variable dependiente, y dos o más variables independientes. Ese es el caso de la regresión lineal múltiple, aquí no es posible hacer un diagrama de dispersión como se recomienda para la regresión múltiple.

2. Regresión Lineal Múltiple. Supuestos.

El cambio fundamental con respecto al modelo de regresión lineal es la Matriz de Términos Independientes que en vez de ser una matriz de $n \times 2$ sera una matriz de $n \times m$ donde $m-1$ es la cantidad de variables independientes, por consiguiente el vector de parámetros será una vector de tamaño $m \times 1$. Por tanto

$$Y = X\beta + e \tag{1}$$

Donde:

$Y_{n\times 1} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ Vector del Término Dependiente. Variable Aleatoria Observada

$X_{n\times m} = \begin{pmatrix} 1 & \dots & x_{1m} \\ \vdots & \cdots & \vdots \\ 1 & \dots & x_{nm} \end{pmatrix}$ Matriz de Términos Independientes o Matriz del Diseño.

$\beta_{m\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}$ Vector de los Parámetros.

$e_{n\times 1} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$ Vector de los errores. Variable Aleatoria.

En forma algebraica tendríamos que para $i = 1, \dots, n$ tenemos

$$y_i = \beta_0 + x_1\beta_1 + x_2\beta_2 + \cdots + x_m\beta_m + e_i \qquad (2)$$

Por tanto, los supuestos del modelo de regresión lineal múltiple coinciden con los 4 supuestos de la regresión lineal simple más uno en el que intervienen las variables independientes:

1. Existe una relación lineal entre las variables dependientes e independientes.
2. Los errores $(e_1, \dots, e_n)$ son independientes.
3. El valor esperado del error aleatorio e_i es cero $(E(e_i) = 0)$
4. La Varianza del error aleatorio es constante $(V(e_i) = \sigma^2)$. Homocedasticidad.
5. Los errores además de ser independientes son idénticamente distribuidos y siguen distribución normal con media cero y varianza constante $\left(e_i \sim N(0, \sigma^2)\right)$
6. Las variables independientes del modelo no están correlacionadas.

El análisis de la regresión es un procedimiento iterativo, esto es, se establece el modelo y se estiman los parámetros se realiza un análisis de bondad de ajuste que de no ser satisfactoria entonces debe regresarse a los pasos anteriores. Tener un modelo con buen ajuste es muy importante, pues es lo que nos permite realizar predicciones correctas utilizando el modelo. Pero se debe hacer una comprobación minuciosa de las suposiciones, de no cumplirse alguna será necesario tomar decisiones al respecto que pueden significar repetir el procedimiento de análisis con estas nuevas ideas.

Por tanto el procedimiento para relacionar una variable dependiente Y con m regresores o variables independientes seria:

1- Establecer el modelo
2- Estimar los coeficientes
   a. En este caso cada β_i mide el cambio en Y por cada cambio unitario en X_i manteniendo los $X_j \ \forall j \neq i$ contantes.
3- Valorar la calidad del ajuste
4- Chequear las suposiciones del modelo. De no cumplirse regresar al paso 1 y rectificar el modelo tantas veces como sea necesario.

3. Método de Mínimos Cuadrados.

Sea S la suma de cuadrados de los residuos,

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_1 x_{mi}\right)^2 \tag{3}$$

Para aplicar el método de mínimos cuadrados en el modelo de regresión lineal múltiple, calculamos la primera derivada de S con respecto a cada $\hat{\beta}_j$ obteniendo

$$\begin{aligned} \frac{\partial S}{\partial \hat{\beta}_0} &= 2\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_1 x_{mi}\right)(-1) \\ \frac{\partial S}{\partial \hat{\beta}_1} &= 2\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_1 x_{mi}\right)(-x_{1i}) \\ &\vdots \qquad\qquad\qquad \vdots \\ \frac{\partial S}{\partial \hat{\beta}_m} &= 2\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_1 x_{mi}\right)(-x_{mi}) \end{aligned} \tag{4}$$

Los estimadores mínimos cuadráticos se obtienen de igualar a cero las derivadas anteriores:

$$\begin{aligned} &\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_1 x_{mi}\right) = 0 \\ &\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_1 x_{mi}\right) x_{1i} = 0 \\ &\qquad\qquad \vdots \\ &\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_m x_{mi}\right) x_{mi} = 0 \end{aligned} \tag{5}$$

Escrito en notación matricial seria:

$$\boldsymbol{X'X\widehat{\beta}} = \boldsymbol{X'y} \tag{6}$$

Al sistema anterior se le denomina genéricamente sistema de ecuaciones normales del hiperplano. En notación matricial ampliada, el sistema de ecuaciones normales es el siguiente:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{1i} & \cdots & \sum_{i=1}^{n} x_{mi} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \cdots & \sum_{i=1}^{n} x_{mi}x_{1i} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{mi} & \sum_{i=1}^{n} x_{mi}x_{1i} & \cdots & \sum_{i=1}^{n} x_{mi}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{2i}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ki}y_i \end{bmatrix} \tag{7}$$

Obsérvese que:

a) $\boldsymbol{X'X}/n$ es la matriz de momentos muestrales de segundo orden, con respecto al origen, de los regresores, entre los cuales se incluye la variable independiente ficticia x_{0i} asociada al término independiente, que toma el valor $x_{0i} = 1\ \forall i$.

b) $\boldsymbol{X'y}/n$ es el vector de momentos muestrales de segundo orden, con respecto al origen, entre la variable dependiente y las variables independientes.

En este sistema hay m ecuaciones y m incógnitas$(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_m)$. Este sistema puede resolverse fácilmente utilizando álgebra matricial. Con el fin de resolver unívocamente el sistema con respecto a $\widehat{\boldsymbol{\beta}}$, es preciso que el rango de la matriz $\boldsymbol{X'X}$ sea igual a m. Si esto se cumple, ambos miembros de 6 pueden ser pre multiplicados por $[\boldsymbol{X'X}]^{-\mathbf{1}}$:

$$[\boldsymbol{X'X}]^{-\mathbf{1}}\boldsymbol{X'X}\widehat{\boldsymbol{\beta}} = [\boldsymbol{X'X}]^{-\mathbf{1}}\boldsymbol{X'y} \tag{8}$$

Obteniéndose la expresión del vector de estimadores de mínimos cuadrados, o más exactamente, el vector de estimadores de mínimos cuadrados ordinarios (MCO), porque $[\boldsymbol{X'X}]^{-\mathbf{1}}\boldsymbol{X'X} = \mathbf{1}$. Por lo tanto, la solución es la siguiente:

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_m \end{bmatrix} = \widehat{\boldsymbol{\beta}} = [\boldsymbol{X'X}]^{-\mathbf{1}}\boldsymbol{X'y}$$

Como la matriz de segundas derivadas, $2\boldsymbol{X'X}$, es una matriz definida positiva, la conclusión es que S presenta un mínimo en $\widehat{\boldsymbol{\beta}}$.

4. Problemas Especiales de la Regresión.

Estos problemas vienen dados por el incumplimiento de los supuestos del modelo. El requisito de linealidad puede comprobarse a través de gráficos de dispersión o pruebas de linealidad. Para saber que no se cumple el supuesto de homocedasticidad se observan directamente los residuos, los gráficos de los mismos o se realiza una prueba de hipótesis. Mayormente analizaremos los gráficos de los residuos comprobar que se cumplen los supuestos asociados a estos. Sin embargo cuando no se cumple el supuesto de linealidad entre las variables independientes, existe el llamado problema de multicolinealidad.

4.1 Multicolinealidad.

Las principales causas de la multicolinealidad son:

a) Método de recolección de datos inadecuado.
b) Restricciones en el modelo o en la población.
c) Selección de un modelo no adecuado.
d) Un modelo sobre definido, o sea se tienen más variables independientes que observaciones.

Para diagnosticar la multicolinealidad existen se puede proceder de diferentes formas, pero la más utilizada e intuitivita es la primera:

1. Examinar la matriz de correlaciones, si las variables independientes X_i, X_j son linealmente dependientes entonces $|r_{ij}|$ será cercano a uno.
2. Buscar los factores de inflamación de la varianza VIF
3. Analizar los Valores Propios de la matriz $\boldsymbol{X^T X}$

Para tratar datos con multicolinealidad lo ideal es recoger datos adicionales, pero como esto casi nunca será posible, este método no es bueno. Otro procedimiento seria re especificar el modelo, podría trabajarse en este sentido con dos enfoques, uno redefinir los regresores, por ejemplo tomar $X_* = X_1 * X_2$ o alguna otra expresión o eliminar variables independientes.

Otra opción sería buscar otros métodos de estimación de los coeficientes β_j. El método de mínimos cuadrados hace que $\hat{\beta}$ sea un estimador insesgado de β. Una solución sería buscar estimadores sesgados.

Con los avances tecnológicos del mundo de hoy por lo general las regresiones se hacen utilizando algún sistema computacional, por tanto el trabajo del investigador se reduce a saber determinar cuándo un modelo cumple los supuestos y que variables seleccionar para incluir o eliminar de dicho modelo.

5. Selección de Variables.

A veces se quiere elegir entre un número de posibles variables independientes, algún subconjunto de ellas que brinde un "mejor" modelo de regresión. Aunque se van a exponer algunos métodos de como hallar este "mejor" modelo, es imprescindible señalar que en la selección de un modelo, el criterio más importante es el conocimiento del investigador de la situación de estudio, que es lo que determina al final las variables a incluir. Sin este conocimiento, la regresión resultante, aunque tenga una elevada capacidad de estimación y de predicción, es un modelo sin relevancia teórica ni práctica.

5.1 Regresión Paso a Paso (Stepwise)

Este método de selección de variables es quizás el más popular. Permite examinar la contribución de cada variable independiente al modelo de regresión. Se considera la inclusión de cada variable antes de desarrollar la ecuación, añadiéndose primero la variable independiente con la contribución más alta. Posteriormente, se seleccionan las variables independientes para su inclusión, basándose en su contribución incremental sobre las variables ya existentes en la ecuación.

Primero, se calcula el modelo de regresión simple que incluye la variable independiente más altamente correlacionada con la variable respuesta. Después, se examinan los coeficientes de correlación parcial para encontrar una variable independiente adicional que explique la mayor parte del error que queda de la primera ecuación de regresión. Se vuelve a calcular la ecuación de regresión utilizando las dos variables independientes y se examina el valor de la prueba F parcial de la variable original del modelo para ver si todavía realiza una contribución significativa, dada la presencia de la nueva variable independiente. Si no lo hace, se elimina la variable.

5.2 Regresión hacia delante ("forward")

Estos procedimientos son procesos de ensayo y error para buscar los mejores estimadores de la regresión.

El modelo hacia delante es similar al procedimiento paso a paso, pero se diferencia en que una vez que se añade una variable no existe la posibilidad de revertir la acción posteriormente.

5.3 Regresión hacia atrás ("backward")

Quizás el método más utilizado se la regresión hacia atrás, que implica calcular una regresión de todas las variables independientes, para a continuación, ir eliminando las variables que no contribuyan significativamente. Al igual que en la regresión hacia adelante no se puede revertir el proceso.

6. Ejemplo de Regresión en R

Uno de los problemas que se tratan en disciplinas como la Ecología o la Biología de la Conservación es el de identificar factores que influyen en variables como la riqueza de una especie (medida como el número de individuos de la especie en un área dada). En un estudio sobre la población de un parásito se hizo un recuento de parásitos en 15 localizaciones con diversas condiciones ambientales.

Los datos obtenidos son los siguientes:

Temperatura	15 16 24 13 21 16 22 18 20 16 28 27 13 22 23
Humedad	70 65 71 64 84 86 72 84 71 75 84 79 80 76 88
Recuento	156 157 177 145 197 184 172 187 157 169 200 193 167 170 192

Tabla 1. Datos de parasitos.csv

Lo que se quiere averiguar es si la Riqueza de la Especie o sea el Recuento tiene alguna relación con las variables Temperatura $(Temp)$ y Humedad (Hum). Utilizaremos el método backward. Cuando la regresión es múltiple tenemos más de una variable, por lo tanto necesitamos ver cómo se comportan todas las variables. Lo primero sería realizar un gráfico de dispersión para estar seguros de que tiene sentido hacer una regresión lineal. Además podemos calcular la matriz de correlación de los datos, en caso que el diagrama no sea suficiente.

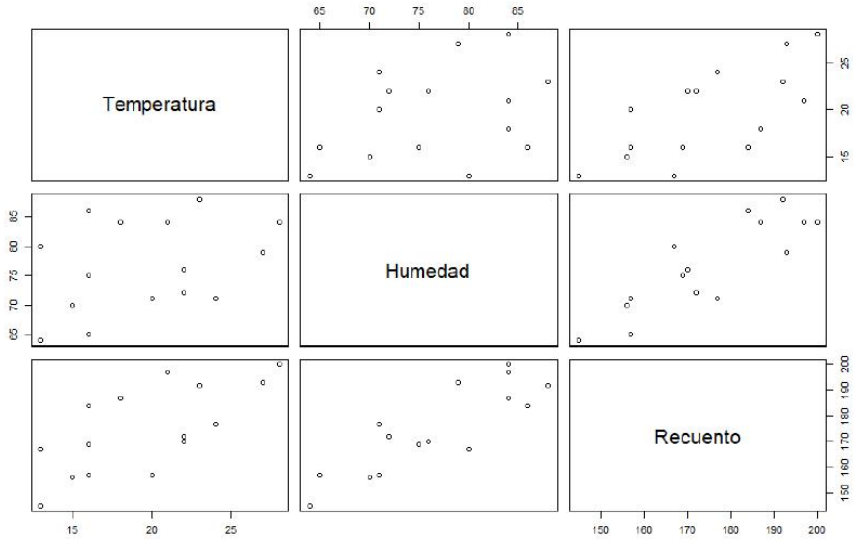

Figura 1. Cargar Datos y Calcular Regresión lineal.

Como se puede observar las variables Temperatura y Humedad tiene una relación lineal, esto lo apoya también la matriz de correlación.

```
> cor(data)
            Temperatura   Humedad  Recuento
Temperatura   1.0000000 0.3559092 0.6967744
Humedad       0.3559092 1.0000000 0.8597143
Recuento      0.6967744 0.8597143 1.0000000
```

Figura 2. Matriz de Correlación.

Después planteamos el modelo de regresión lineal múltiple de la siguiente forma:

$$Rec = \beta_0 + Temp * \beta_1 + Hum * \beta_2 + e$$

Donde la variable dependiente seria Recuento (Rec) y las independientes serian Humedad y Temperatura. Utilizando el comando los comandos que se muestran en el listado de código 1 podemos calcular la regresión múltiple. Primero cargamos los datos, luego como los datos tienen en la primera fila el nombre de la variable utilizamos el comando ***attach*** para cargar estos como variables y poder usarlos sin necesidad de referenciar a los datos. La regresión lineal se realiza en la línea 3 y se guarda en la variable *multi.fit*.

```
1 data <- read.csv("parasito.csv")
2 attach(data)
3 multi.fit = lm(Recuento~Temperatura+Humedad, data=data)
4 summary(multi.fit)
```

Listado de Código 1. Cargar Datos y Calcular Regresión lineal.

Por lo tanto, para investigar los resultados de la regresión solo tenemos que correr el comando ***summary*** como se observa en la figura 3.

```
Call:
lm(formula = Recuento ~ Temperatura + Humedad, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8617 -2.0406  0.4319  2.9881  8.5047

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.7115    14.3725   1.789 0.098876 .
Temperatura   1.5818     0.3203   4.939 0.000343 ***
Humedad       1.5424     0.1995   7.731 5.32e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.351 on 12 degrees of freedom
Multiple R-squared:  0.914,     Adjusted R-squared:  0.8996
F-statistic: 63.75 on 2 and 12 DF,  p-value: 4.051e-07
```

Figura 3. Salida del Modelo de Regresión Múltiple.

En este caso los valores de los β_j están en la columna de Estimados, por lo que el modelo con los coeficientes sustituidos seria el siguiente:

$$\widehat{Rec} = 25.7115 + Temp * 1.5818 + Hum * 1.54.24$$

6.1 Interpretando la salida de Regresión de R

- **Residuals**: Aquí se observa el sumario de los residuales, el error entre la predicción del modelo y los resultados reales. Mientras más pequeños sean los residuales mejor.
- **Coefficients**: Para cada variable independiente y el intercepto se tiene:
    - **Estimate**: Estimado. Este es el valor de los β_j
    - **Std. Error**: Error Estándar. Nos da la precisión del estimador. Útil para calcular el t-value.
    - **t-value** y **Pr(>[t])**: Es realmente importante ya que es una forma de medir si la variable en cuestión o el intercepto aportan algo significativo al modelo. El t-value es calculado dividiendo el coeficiente entre el error estándar y luego es utilizado para plantear una prueba de hipótesis donde mide si el coeficiente es diferente de 0. Si no es significativa, entonces el coeficiente no está aportando nada al modelo por lo que la variable podría ser eliminada, y esta es una de las formas de eliminar variables del modelo. Para que sea significativa el **Pr(>[t])** tiene que ser menor que 0.05.

- **Performance Measures**: Como su nombre lo indica muestran que tan buena es la recta de regresión.
    - **Residual Standard Error**: Error estándar de los residuos Mientras más pequeña mejor.
    - **Multiple / Adjusted R-Square**: Cuando trabajamos con una sola variable independiente no importa la distinción entre el R-cuadrado o R-square. Esta medida nos dice la cantidad de variación explicada en el modelo. El R-cuadrado ajustado toma en cuenta el número de variables independientes, por tanto es el más usado en regresión múltiple. Mientras más cercano a uno sea este valor, será mejor. Si está por debajo de 0.70 entonces el modelo es muy malo.
    - **F-Statistic**: La prueba F dice si al menos uno de los β_j es significativamente diferente a cero. Esta es una prueba de hipótesis global para poder valorar el modelo. Si el p-valor no es significativo (o sea es mayor que 0.05) entonces nuestro modelo no está haciendo nada.

Comencemos analizando los estimados de los coeficientes β_j en nuestro problema, lo primero que observamos es que el coeficiente del intercepto es mucho más grande que los coeficientes del resto de las variables.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.7115    14.3725   1.789 0.098876 .
Temperatura   1.5818     0.3203   4.939 0.000343 ***
Humedad       1.5424     0.1995   7.731 5.32e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 4. Salida de los Coeficientes del Modelo de Regresión Múltiple.

Lo que quiere decir que hay una gran parte del Recuento de la población de parásitos que no está explicada a partir de las variables independientes. Como en este caso no tenemos más variables para adicionar al modelo podemos ver si realizando una transformación lineal de todos las variables y aplicando la regresión con estas variables transformadas podemos obtener un valor más pequeño del intercepto. Una de las transformaciones más comunes seria la estandarización de los datos, o sea calcular la media y la desviación estándar de cada variable y tipificarlas, así obtendrían tres variables nuevas $RecT, HumT\ y\ TempT$ y con estas volverían a repetir la regresión.

Lo siguiente a ver en los coeficientes seria ver si son o no significativos los coeficientes y a que porciento, se puede observar que los coeficientes de la temperatura y la humedad son significativos al 0% mientras que el intercepto es significativo al 10%, lo que no es muy bueno.

Por ultimo analicemos los valores de los coeficiente, por cada grado de incremento de temperatura debemos esperar que el recuento de la población de parásitos se incremente en 1.6 individuos. De forma análoga por cada unidad de incremento en la humedad relativa se espera esperar que el recuento de la población de parásitos se incremente en 1.5 individuos

Por ultimo pasamos al análisis de los residuos, lo primero que podemos observar es que el R-cuadrado ajustado es 0.90 por lo que el modelo podemos decir es bastante bueno, pero no es lo ideal. Sumado esto a que el nivel de significación del coeficiente del intercepto es malo podría indicarnos que es posible sea necesario considerar otros factores además de la temperatura y la humedad en el recuento de la población de parásitos, quizás otras condiciones climáticas o la presencia de un depredador. Porque lo ideal es que el R-cuadrado sea lo más cercano a 1 posible.

```
Residual standard error: 5.351 on 12 degrees of freedom
Multiple R-squared:  0.914,     Adjusted R-squared:  0.8996
F-statistic: 63.75 on 2 and 12 DF,  p-value: 4.051e-07
```

Figura 5. Salida de los Residuos del Modelo de Regresión Múltiple.

El valor del Error estándar de los residuos nos permitirá comparar el modelo con otros que tengan similares valores de R-cuadrado ajustado, para así escoger el de menor error. El p-valor del estadígrafo F es menor que 0.05 por lo que podemos afirmar que existe al menos una variable significativamente diferente a cero en el modelo.

6.2 Analizando los Residuos

Como hemos visto cualquiera puede ajustar un modelo de regresión en R. La prueba real viene a la hora de analizar los residuos para comprobar que se cumplen los supuestos del modelo. En general tenemos que analizar cuatro cuestiones con respecto a los residuos.

- La media de los errores es cero y la suma de los errores es cero.
- Los errores tiene distribución normal
- Los errores son independientes
- La varianza de los errores es constante (Homocedasticidad)

Lo primero sería acceder a los residuos y analizarlos gráficamente. En la variable resultado de la regresión podemos acceder a los residuos mediante el comando ***multi.fit$residuals***. Luego comprobamos los supuestos

1- La media de los errores es cero y la suma de los errores es cero.

Como se puede observar en el listado de código 2, este supuesto se cumple.

```
> mean(multi.fit$residuals)
[1] 1.480297e-16
> sum(multi.fit$residuals)
[1] 2.220446e-15
```

Listado de Código 2. Supuesto 1.

2- Errores normalmente distribuidos

El histograma de residuos y el grafico QQ-plot son formas de evaluar visualmente si los residuos siguen una distribución normal. Por tanto buscamos que el histograma tenga forma de campana y en el QQ-plot que la mayoría de los puntos de los residuos se encuentren sobre la recta o muy cercana a ella. Como se puede observar el histograma de residuos no sigue exactamente el patrón de una distribución normal, sin embargo el QQ-plot muestra solo una pequeña desviación con respecto a la línea de la normal por tato podemos asumir la normalidad de los errores.

Test de Shapiro–Wilk se usa para contrastar la normalidad de un conjunto de datos. Se plantea como hipótesis nula que una muestra proviene de una población normalmente distribuida. Fue publicado en 1965 por Samuel Shapiro y Martin Wilk y es considerado uno de los test más potentes para el contraste de normalidad, sobre todo para muestras de tamaño menor que 50 (n<50). Para muestras más grandes

se puede utilizar la prueba no paramétrica de Kolmogórov-Smirnov para determinar la bondad de ajuste de dos distribuciones de probabilidad entre sí.

```
> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.96319, p-value = 0.7476
```

Listado de Código 3. Prueba Shapiro-Wilk.

En el ejemplo el p-valor del test de Shapiro-wilk es $0.75 \gg 0.05$ no se podemos rechazar la hipótesis nula por lo que los errores siguen una distribución normal.

3- Independencia de los residuos

La prueba Durbin-Watosn se usa para probar si los residuos son independientes. Para esto necesitan utilizar el paquete *lmtest*. La hipótesis nula de esta prueba es que los errores son independientes.

```
> dwtest(multi.fit) #Test for independence of residuals

        Durbin-Watson test

data:  multi.fit
DW = 1.7548, p-value = 0.3247
alternative hypothesis: true autocorrelation is greater than 0
```

Listado de Código 4. Prueba Dubin-Watson.

Como el p-valor de esta prueba es $0.32 \gg 0.05$ no podemos rechazar la hipótesis nula por lo que podemos afirmar que los errores son independientes.

4- Supuesto de Homocedasticidad.

Para probar este supuesto podemos graficar los residuos (figura 6, gráficos superiores), si estos gráficos siguen un patrón como el explicado en la conferencia 6, como es el caso, entonces tenemos homocedasticidad, si no podemos recurrir a la prueba de Breusch-Pagan se utiliza para determinar la heterocedasticidad en un modelo de regresión lineal.

```
> bptest(multi.fit)

        studentized Breusch-Pagan test

data:  multi.fit
BP = 0.40552, df = 2, p-value = 0.8165
```

Listado de Código 5. Prueba Shapiro-Wilk.

Como el p-valor de esta prueba es $0.82 \gg 0.05$ no podemos rechazar la hipótesis nula por lo que podemos afirmar que se cumpla la heterocedasticidad. Por lo que el supuesto de Homocedasticidad se mantiene.

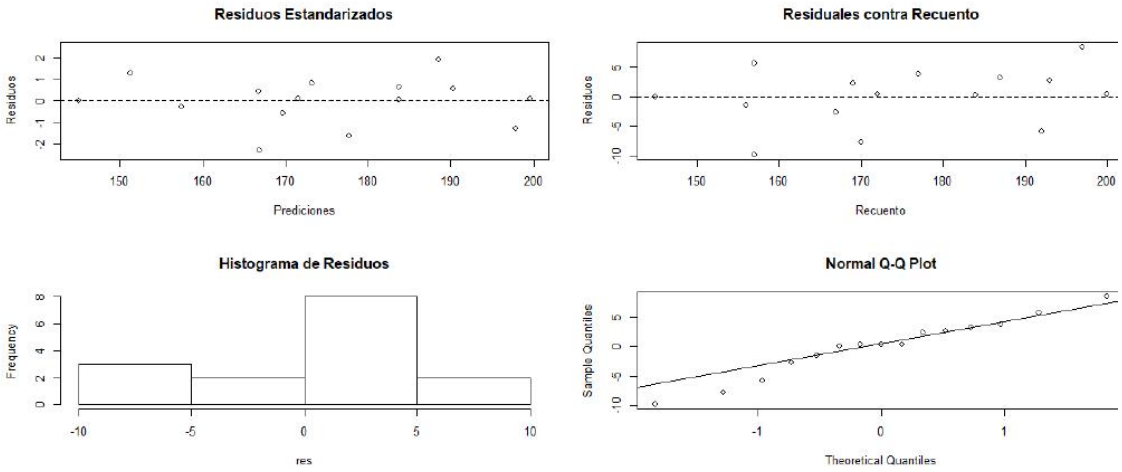

Figura 6. Salida de los Residuos del Modelo de Regresión Múltiple.

Además de las gráficas para comprobar la homocedasticidad también podemos apoyarnos en gráficas para los supuestos de normalidad. Fundamentalmente en los histogramas y QQ-Plots, el código para generar los cuatro gráficos se presenta en el listado de código 6.

```
#Multiple Regression Residual Plots
layout(matrix(c(1,2,3,4),2,2,byrow=T))

plot(multi.fit$fitted.values, rstandard(multi.fit),
     main="Multi Fit Standarized Residuals",
     xlab="Predictions",ylab="Standarized Resid",
     ylim=c(-2.5,2.5))
abline(h=0, lty=2)

plot(Recuento, res,
     main="Residuales contra Recuento",
     xlab="Recuento",ylab="Residuos")
abline(h=0,lty=2)

hist(res,main="Histograma de Residuos")

qqnorm(res)
qqline(res)
```

Listado de Código 6. Graficando los residuos.