VARIABLES CUALITATIVAS E INTERACCIONES

ISADORE NABI

20/10/2021

PREDICTORES CUALITATIVOS

Breve introducción teórica a los predictores cualitativos

Como se señala en (Kutner, Nachtsheim, Neter & Li. Applied Linear Statistical Models, 2005. p. 313), existen diferentes formas de identificar las clases de una variable cualitativa (los diferentes tipos de cualidad que puede adoptar la variable en cuestión, conocidas en el contexto de la bioestadística como “categorías”). Por ejemplo, es posible utilizar variables indicatriz, i.e., aquellas que únicamente pueden adoptar los valores de 0 y 1 de forma excluyente; por supuesto, toda variable indicatriz tiene una función indicatriz -que es una función a trozos- que la modela. Las variables indicatriz son fáciles de usar y ampliamente empleadas, aunque no son bajo ningún escenario la forma exclusiva de cuantificar una variable cualitativa.

```
setwd("C:/Users/User/Desktop/Carpeta de Estudio/Mis Códigos en R")
knitr::include_graphics("FOTO.JPG")
```

$$X_2 = \begin{cases} 1 & \text{if stock company} \\ 0 & \text{otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if mutual company} \\ 0 & \text{otherwise} \end{cases}$$

```
#Figura 1: Ejemplos de Variables Indicatriz
#Fuente:(Kutner, Nachtsheim, Neter & Li. Applied Linear Statistical Models, 2005. p. 313).
```

Como se señala en (Kutner, Nachtsheim, Neter & Li. p.315), una variable cualitativa con c clases puede ser representada por $c-1$ variables indicatriz, cada una tomando el valor de 0 o 1 según se defina en la investigación. Las variables indicatriz a menudo son conocidas como variables dicotómicas, variables binarias o variables “dummy”.

Para el caso de los modelos de regresión que exijan independencia lineal entre las variables explicativas de una determinada respuesta, no es recomendable emplear (al menos no explícitamente) más de una variable indicatriz. Como es sabido, una matriz de diseño X (que es la forma en que en el contexto del diseño de experimentos se llama en ocasiones a la matriz de variables explicativas X) para el modelo de regresión lineal (simple o múltiple) contendrá indefectiblemente en su primera columna únicamente 1’s, puesto que ello se requiere (matemáticamente hablando) para poder estimar el intercepto en Y, es decir, el parámetro β_0. Como se señala en (Kutner, Nachtsheim, Neter & Li. p.314), debido al hecho antes descrito el usar más de una variable indicatriz (que implica cuantitativamente añadir dos columnas de 1’s y 0’s a la matriz de diseño X) se obtiene que la primera columna es igual a la suma de las columnas correspondientes a las variables indicatriz y, por consiguiente, mostrando que existe dependencia lineal entre las variables (lo que viola el supuesto de multicolinealidad en el marco del modelo clásico de regresión lineal). La consecuencia de ello, como se desprende de estudiar el sistema de ecuaciones planteado por (Greene. Econometric Analysis, International Edition, 2012. p. 66-68), es que el vector columna β que satisface la ecuación de mínimos cuadrados normales $X'X\beta = X'y$ no existiría de forma única porque la matriz $X'X$ no tendría inversa (que por su definición formal -matemática- la inversa de X es siempre única) y, por consiguiente, el sistema de ecuaciones no podría ser resuelto de forma analítica y, por consiguiente, los resultados obtenidos (encontrados mediante

aproximaciones empíricas por métodos numéricos) no reflejarían una forma analítica que garantice que la estimación estadística realizada cumple con todas las propiedades deseables bien-definidas en la literatura científica correspondiente.

Caso de Aplicación: Innovaciones Financieras en el Mercado de Seguros (Kutner, Nachtsheim, Neter & Li. p.313-318)

Sea un determinado mercado de seguros dentro del cual se aplican con cierta periodicidad innovaciones financieras para maximizar la tasa de ganancia de las firmas o empresas que innovan y en el que ocurre un determinado efecto “bola de nieve” para que la innovación sea adoptada por otras firmas. Se desea determinar la relación existente de la velocidad a la cual una innovación financiera es adoptada por una firma (la respuesta Y) con el tamaño de la empresa aseguradora (medida por el monto total de activos de la misma) y con el tipo financiero de la empresa aseguradora (si es accionaria o mutual).

Así, retomando lo visto en la figura 1 y lo planteado respecto a la multicolinealidadm, sea ahora X_2 una variable indicatriz o dicotómica que adopta el valor de 1 si la empresa analizada es de tipo mutual y 0 si es de tipo accionario. Con lo anterior se solventa el problema de colinealidad y únicamente se requirió restringir el ámbito de aplicación de las funciones indicatriz, es decir, “accionario o mutual” es una versión restringida de “Accionario u otro; Mutual u otro”; en la matriz de datos aquí empleada, la variable X_2 se denota como $Tipo.f$, mientras que X_1 se denota como $Tamano$.

La muestra disponible es la correspondiente a 10 firmas accionarias (de stock o simplemente stock para este ejercicio) y 10 firmas mutuales, es decir, 20 observaciones para cada una de las tres variables disponibles ($Y = velocidaddeadopciónde la innovación$, $X_1 = Tamaño$, $X_2 = Tipo.f$) que impllica un tamaño de muestra global de 60 observaciones.

La forma general del modelo de regresión lineal clásico es $E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. El modelo anterior adoptará, para el caso en que se considere en el análisis únicamente a las compañías mutuales, la forma $E[Y] = \beta_0 + \beta_1 X_1$, puesto que para dicho caso se cumple rigurosamente que $X_2 = 0$; mientras que adoptará la forma $E[Y] = (\beta_0 + \beta_2) + \beta_1 X_1$ cuando se analice únicamente a las compañías accionarias puesto que $X_2 = 1$.

Una base de datos como la anterior presentaría la siguiente forma:

```
knitr::include_graphics("BASE_DE_DATOS.JPG")
```

TABLE 8.2 Data and Indicator Coding—Insurance Innovation Example.

Firm i	(1) Number of Months Elapsed Y_i	(2) Size of Firm (million dollars) X_{i1}	(3) Type of Firm	(4) Indicator Code X_{i2}	(5) $X_{i1}X_{i2}$
1	17	151	Mutual	0	0
2	26	92	Mutual	0	0
3	21	175	Mutual	0	0
4	30	31	Mutual	0	0
5	22	104	Mutual	0	0
6	0	277	Mutual	0	0
7	12	210	Mutual	0	0
8	19	120	Mutual	0	0
9	4	290	Mutual	0	0
10	16	238	Mutual	0	0
11	28	164	Stock	1	164
12	15	272	Stock	1	272
13	11	295	Stock	1	295
14	38	68	Stock	1	68
15	31	85	Stock	1	85
16	21	224	Stock	1	224
17	20	166	Stock	1	166
18	13	305	Stock	1	305
19	30	124	Stock	1	124
20	14	246	Stock	1	246

```
#Figura 2: Base de Datos con Variable Indicatriz
#Fuente: (Kutner, Nachtsheim, Neter & Li. p.317).
```

Esta base de datos puede transformarse de tal forma que si se trata de una compañía mutual la variable X_2 adoptará el valor de 1, mientras que si es una compañía accionaria adoptará el valor de 2. Esto se realizará en este documento con dos fines:

1. Poder representar gráficamente en dos dimensiones el modelo de tres dimensiones, aprovechando el hecho de que una respuesta binaria puede representarse mediante una dualidad de colores.
2. Manipular el ejercicio del libro de tal forma que de su realización en R se obtenga el máximo aprovechamiento.

Comparación Gráfica de los Modelos Explicativos

Con la finalidad de realizar lo establecido antes en 1), se construirá la función personalizada "(kol=2(Tipo.f=="Mutual")+4(Tipo.f=="Stock"))", con la finalidad que sea igual a la casilla "col" de la sintaxis "plot" y se seleccione automáticamente el color para cada observación al elaborar la gráfica con las rectas de regresión de ambas ecuaciones de regresión (por eso las codificaciones de "Mutual" y "Stock" se cambiaron a "1" y "2", respectivamente). Adicionalmente, $E[Y] = \beta_0 + \beta_1 X_1$ se graficará mediante la sintaxis "mod1=lm(Tiempo~Tamano,Seguros[Tipo.f=="Mutual",])", representado el escenario en que únicamente se contempla estudiar compañañías mutuales, por lo que $X_2 = Tipo.f = 1$ (lo que descarta a las accionarias) aquí planteado es equivalente al escenario localizado en la fuente citada (p. 315) en el que se aborda una compañía mutual. Por otro lado, $E[Y] = (\beta_0 + \beta_2) + \beta_1 X_1$ se graficará mediante la sintaxis "mod2=lm(Tiempo~Tamano,Seguros[Tipo.f=="Stock",])", representando el escenario en que únicamente se contempla estudiar compañías accionarias o de stock, por lo que $X_2 = Tipo.f = 2$ (lo que descarta a las mutuales) aquí planteado es equivalente al escenario localizado en la fuente citada (p. 315) en el que se aborda una compañía accionaria o de stock.

```
load("Seguros.Rdata")
str(Seguros)
```

```
## 'data.frame':    20 obs. of  3 variables:
##  $ Tiempo: num  17 26 21 30 22 0 12 19 4 16 ...
##  $ Tamano: num  151 92 175 31 104 277 210 120 290 238 ...
##  $ Tipo  : num  1 1 1 1 1 1 1 1 1 1 ...
```

```
#Configuración de la Variable Indicatriz X_2="Tipo.f" (inicio)
Seguros$Tipo.f= as.factor(Seguros$Tipo)
Seguros$Tipo.f
```

```
##  [1] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2
## Levels: 1 2
```

```
levels(Seguros$Tipo.f)=c("Mutual","Stock")
Seguros$Tipo.f
```

```
##  [1] Mutual Mutual Mutual Mutual Mutual Mutual Mutual Mutual Mutual Mutual
## [11] Stock  Stock  Stock  Stock  Stock  Stock  Stock  Stock  Stock  Stock
## Levels: Mutual Stock
```

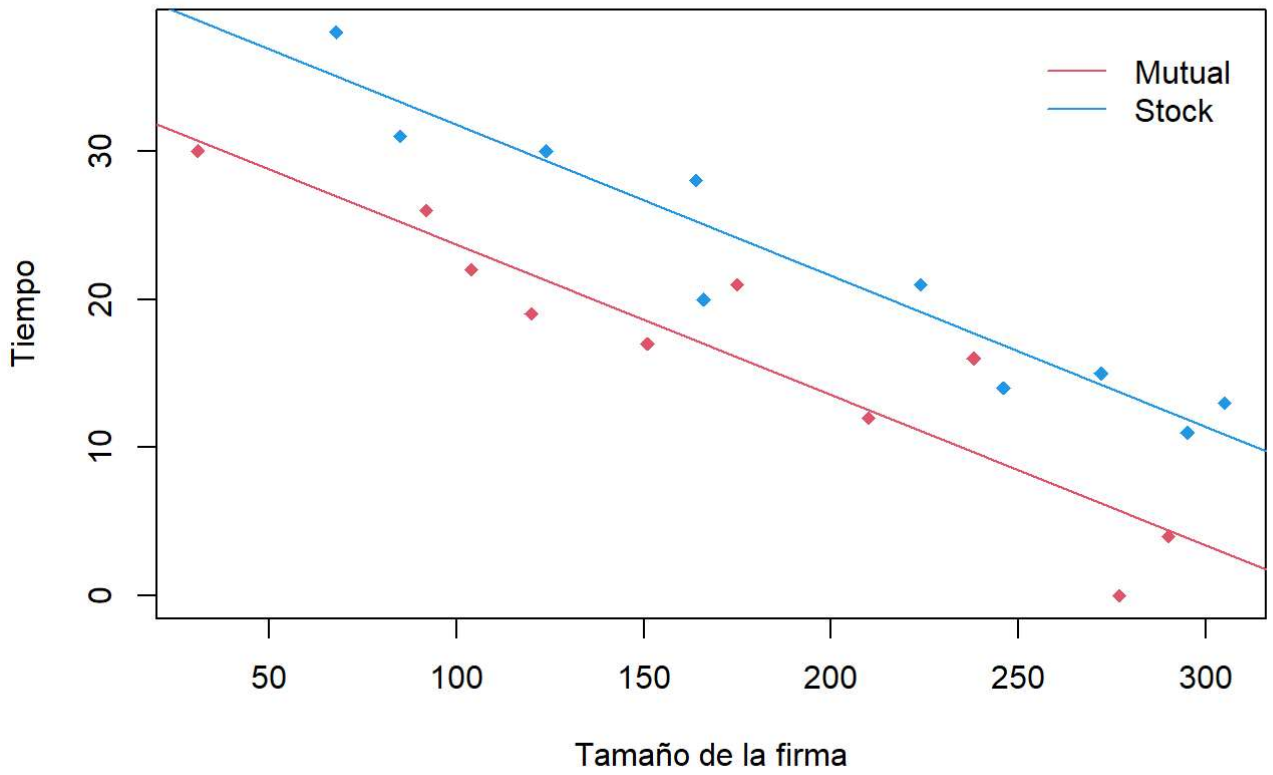
```
#Configuración de la Variable Indicatriz X_2="Tipo.f" (fin)
head(Seguros)
```

```
##   Tiempo Tamano Tipo Tipo.f
## 1     17    151    1 Mutual
## 2     26     92    1 Mutual
## 3     21    175    1 Mutual
## 4     30     31    1 Mutual
## 5     22    104    1 Mutual
## 6      0    277    1 Mutual
```

```
attach(Seguros)
(kol=2*(Tipo.f=="Mutual")+4*(Tipo.f=="Stock"))
```

```
##  [1] 2 2 2 2 2 2 2 2 2 2 4 4 4 4 4 4 4 4 4 4
```

```
plot(Tamano,Tiempo,col=kol,xlab="Tamaño de la firma",ylab="Tiempo",pch=18)
mod1=lm(Tiempo~Tamano,Seguros[Tipo.f=="Mutual",])
mod2=lm(Tiempo~Tamano,Seguros[Tipo.f=="Stock",])
abline(mod1,col=2)
abline(mod2,col=4)
legend(250,max(Tiempo),c("Mutual","Stock"),bty="n",lty=1,col=c(2,4))
```



En la gráfica antes expuesta, se construye lo que presentan (Kutner, Nachtsheim, Neter & Li. Applied Linear Statistical Models, 2005. p. 318). Esto se muestra a continuación.

```
knitr::include_graphics("FOTO2.JPG")
```

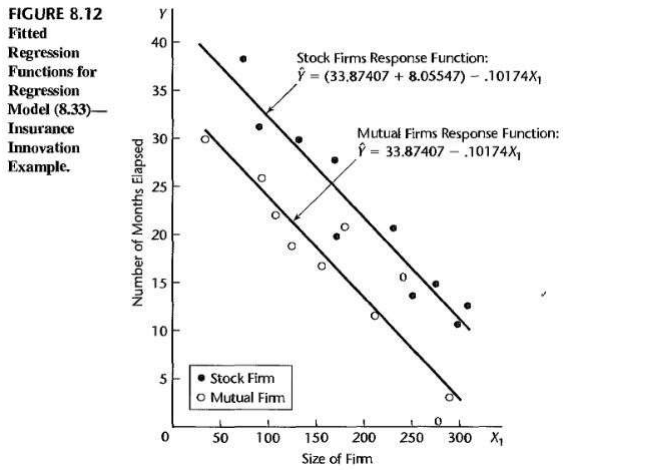


FIGURE 8.12 Fitted Regression Functions for Regression Model (8.33)—Insurance Innovation Example.

```
#Figura 3: Comparación Empírica de Modelos con Predictores Cualitativos
#Fuente:(Kutner, Nachtsheim, Neter & Li. Applied Linear Statistical Models, 2005. p. 318).
```

Análisis Estadístico de los Modelos Explicativos

Introducción

Un investigador podría estar interesado no sólo en comparar descriptivamente dos modelos que tengan en común su intercepto en Y (es decir, que tengan en común el mismo valor para β_0) y el coeficiente β_1 como se hizo antes, sino también de forma inferencial. Para ello, se trabajará con la base de datos original.

La forma general de un caso como el antes descrito se expone a continuación.

```
knitr::include_graphics("FOTO1.JPG")
```

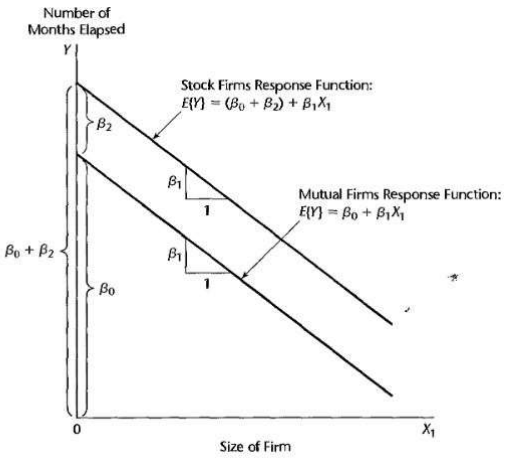


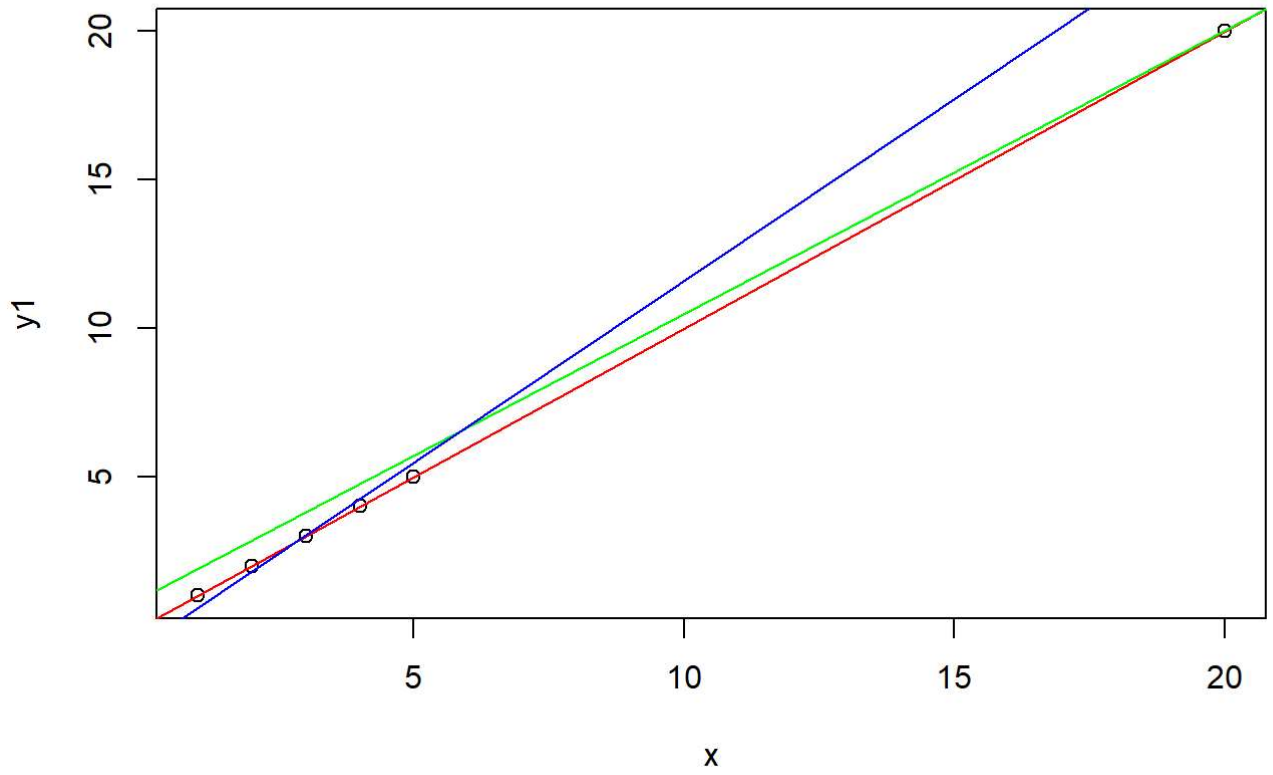
FIGURE 8.11 Illustration of Meaning of Regression Coefficients for Regression Model (8.33) with Indicator Variable X_2—Insurance Innovation Example.

```
#Figura 4: Modelos con predictor cualitativo con igual valor para 6_0 y 6_1
#Fuente: (Kutner, Nachtsheim, Neter & Li. p.316).
```

1. Análisis Descriptivo: Residuos vs Leverage (Influencias)

Como se señala en https://boostedml.com/2019/03/linear-regression-plots-residuals-vs-leverage.html (https://boostedml.com/2019/03/linear-regression-plots-residuals-vs-leverage.html), el significado en términos aplicados de “leverage” en el contexto de regresión es qué tan sensible es un valor condicional de Y (i.e., ajustado por la regresión) a cambios en el valor a priori de Y. Esto ocurre en el contexto de que, como se señala en https://www.wikiwand.com/en/Leverage_(statistics) (https://www.wikiwand.com/en/Leverage_(statistics)), los puntos de alta influencia (“High Leverage”), si los hay, son valores atípicos con respecto a las variables independientes. Esto se comprenderá mejor con el siguiente ejemplo.

```
x=c(1,2,3,4,5,20)
   y1=c(1,2,3,4,5,20)
   y2=c(1,2,7,4,5,20)
   y3=c(1,2,3,4,5,24)
   plot(x,y1)
   abline(lm(y1~x), col='red')
   abline(lm(y2~x),col='green')
   abline(lm(y3~x),col='blue')
```



En la gráfica antes expuesta puede observarse que el valor estimado o condicional de Y cuando $x_i = 20$ está más lejos de otras estimaciones, es decir, posee un alto "levarage". Por lo tanto, cambiar el y_1 asociado (que es a priori) cambiará el ajuste del modelo en mayor grado de lo que lo haría cambiar otros valores y_i (es más sensible esa observación a cambios de modelo -uno a priori y otro posteriori o condicional- que la otra). Para comprobar esto se realizan dos modificaciones sobre la variable de respuesta original y_1, generando con ello y_2 y y_3; la gráfica color rojo corresponde al modelo de regresión construido con y_1. Así, para y_2 se modifica la coordenada, punto de dato u observación en 4 unidades métricas para su coordenada parcial y, lo que equivale a sustituir $x, y = (3, 3)$ por $x, y = (3, 7)$; se representa con una recta color verde. Por su parte, para y_3 se modifica la coordenada, punto de dato u observación en 4 unidades métricas para su coordenada parcial y, que equivale a sustituir $x, y = (20, 20)$ por $x, y = (20, 24)$ y se representa con una recta color azul; no es arbitrario que en ambos casos el incremento sea de 4 unidades métricas, la finalidad de ello es garantizar la estandarización de las condiciones de comparación de las rectas.

Así, como puede observarse, modificar el valor a priori de $y1$ en la coordenada en que x está más lejos de las demás observaciones, tiene un mayor impacto en el ajuste del modelo, reflejado en que el nuevo valor estimado y_3 está más lejano del valor estimado para $y1$ de lo que lo está y_2.

Como se verifica de lo anterior, se está comparando la variación en la respuesta condicional ante cambios en la respuesta a priori en el escenario en el que parece existir un valor atípico (i.e., se sospecha de su existencia) contra el escenario en el que no parece existir, es decir, $x, y = (3, 3)$ contra $x, y = (3, 20)$,

mediante un incremento de 4 unidades métricas en y_i en cada caso. Como se verifica en la última referencia realizada, la distancia de Cook o la D de Cook es una estimación de uso común sobre la influencia de una observación cuando se realiza un análisis de regresión de mínimos cuadrados. En un análisis aplicado de mínimos cuadrados ordinarios, la distancia de Cook se puede utilizar de varias formas: para indicar observaciones influyentes cuya validez merece la pena comprobar; o para indicar regiones del espacio de diseño (construido mediante las covariables) donde sería deseable poder obtener más observaciones. Lleva el nombre del estadístico estadounidense R. Dennis Cook, quien introdujo el concepto en 1977.

Lo anterior puede representarse mediante la sintaxis “plot(mod_i)”, donde “mod_i” es algún modelo construido y almacenado bajo el nombre, por ejemplo, de “mod_1”. Esto se realizará empleando la base de datos original tomada de (Kutner, Nachtsheim, Neter & Li. Applied Linear Statistical Models, 2005. p. 317) y creando un nuevo modelo de regresión lineal múltiple llamado “mod3”.

La comparación entre residuos e influencias tiene como finalidad estudiar cómo cambia la dispersión de los residuos estandarizados a medida que aumenta la influencia o la sensibilidad del modelo ajustado (condicional o a posteriori) a un cambio en y_i. ¿Cuál es la utilidad de esta comparación? En primer lugar, se puede utilizar para detectar la existencia de heterocedasticidad y la no-linealidad. Esto se afirma puesto que la dispersión de los residuos estandarizados no debería cambiar en función de las influencias. aquí parece disminuir, lo que indica heterocedasticidad. En segundo lugar, la eliminación de las observaciones con alta influencia podría cambiar significativamente el modelo. En términos aplicados, la distancia de Cook mide el efecto de eliminar una observación en el vector de parámetros combinados y es representada por las línea rojas punteadas que se verán en la gráfica presentada a continuación, siendo las observaciones fuera de dichas líneas rojas punteadas línea punteada aquellas tienen una gran influencia. En este caso, no hay puntos fuera de tales líneas.

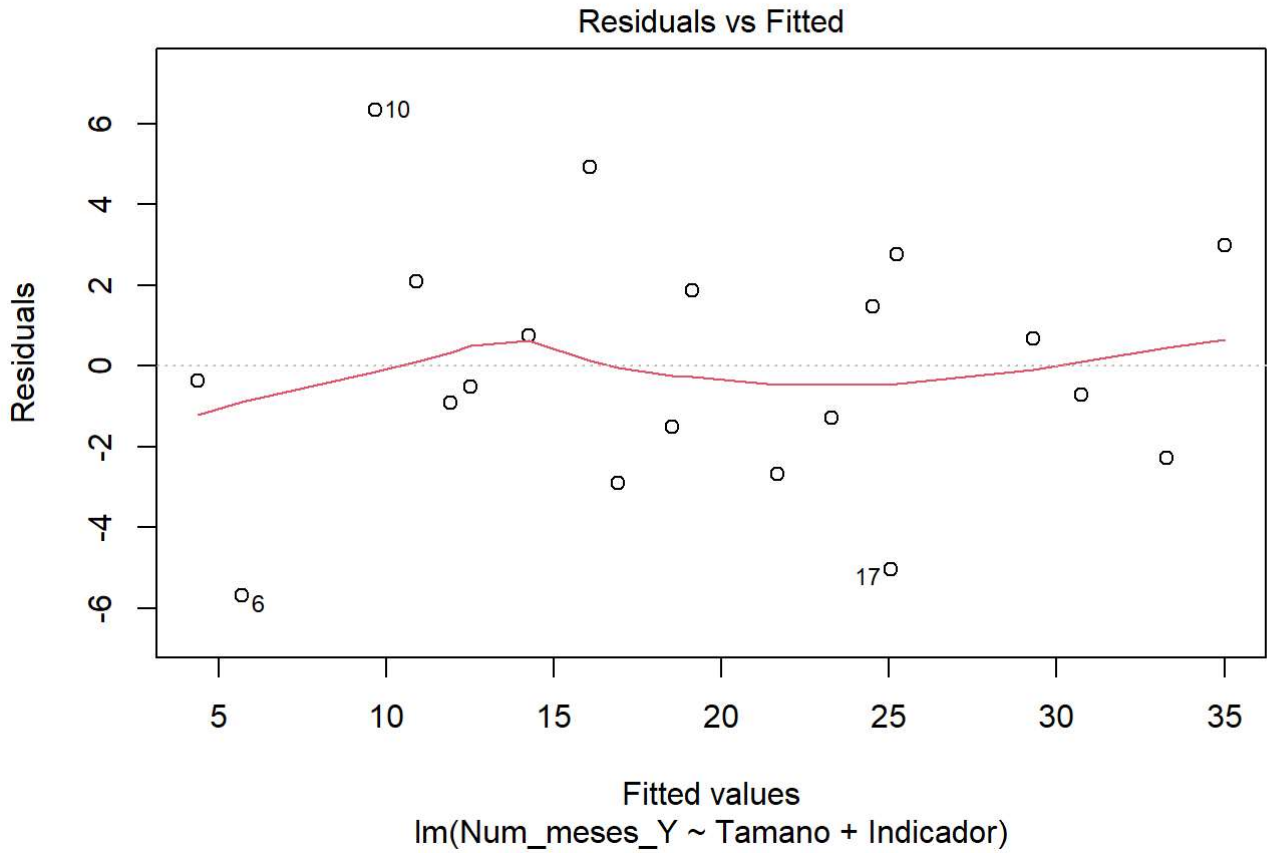
```
library(readxl)
```

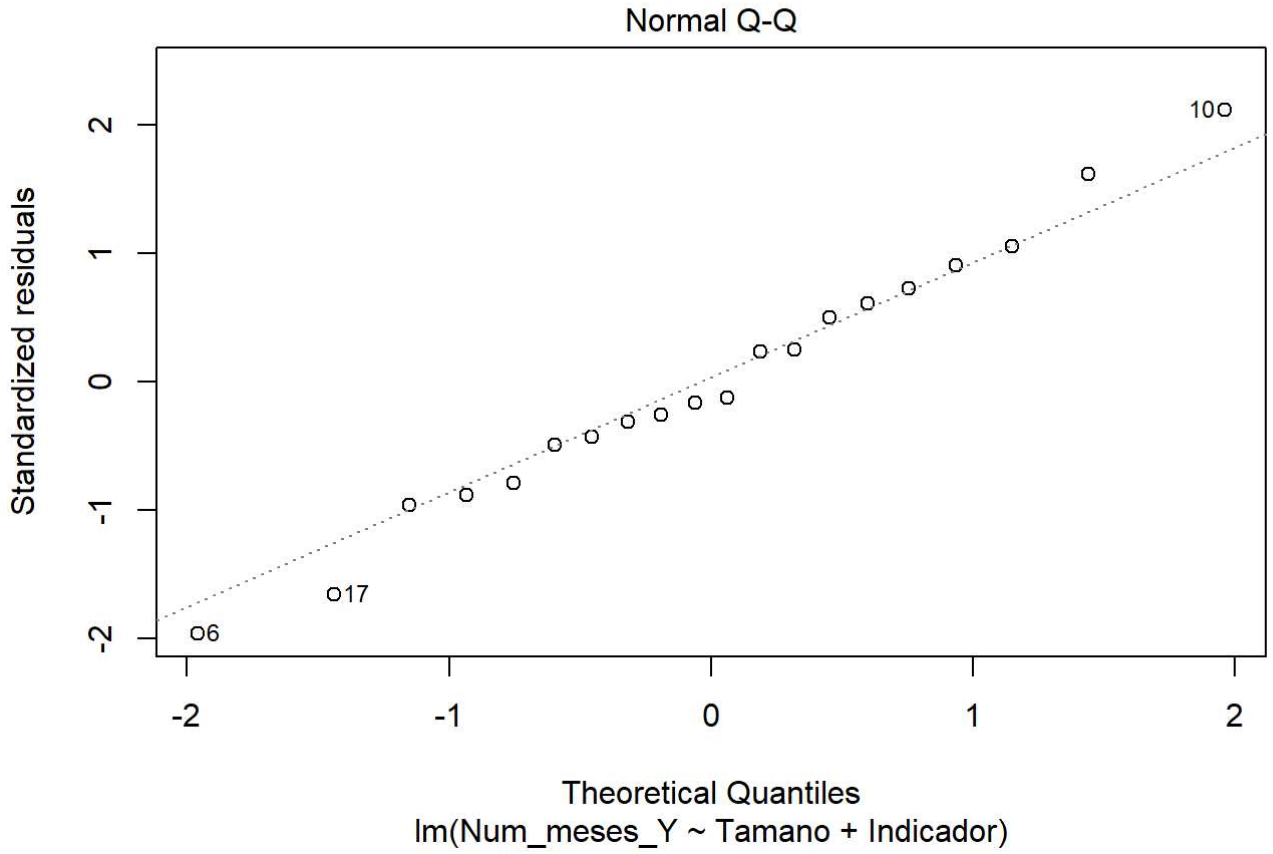
```
## Warning: package 'readxl' was built under R version 4.0.3
```

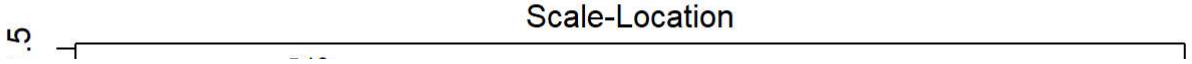
```
Seguros2 <- read_excel("C:/Users/User/Desktop/Carpeta de Estudio/Mis Códigos en R/Seguros2.xl
sx")
attach(Seguros2)
```

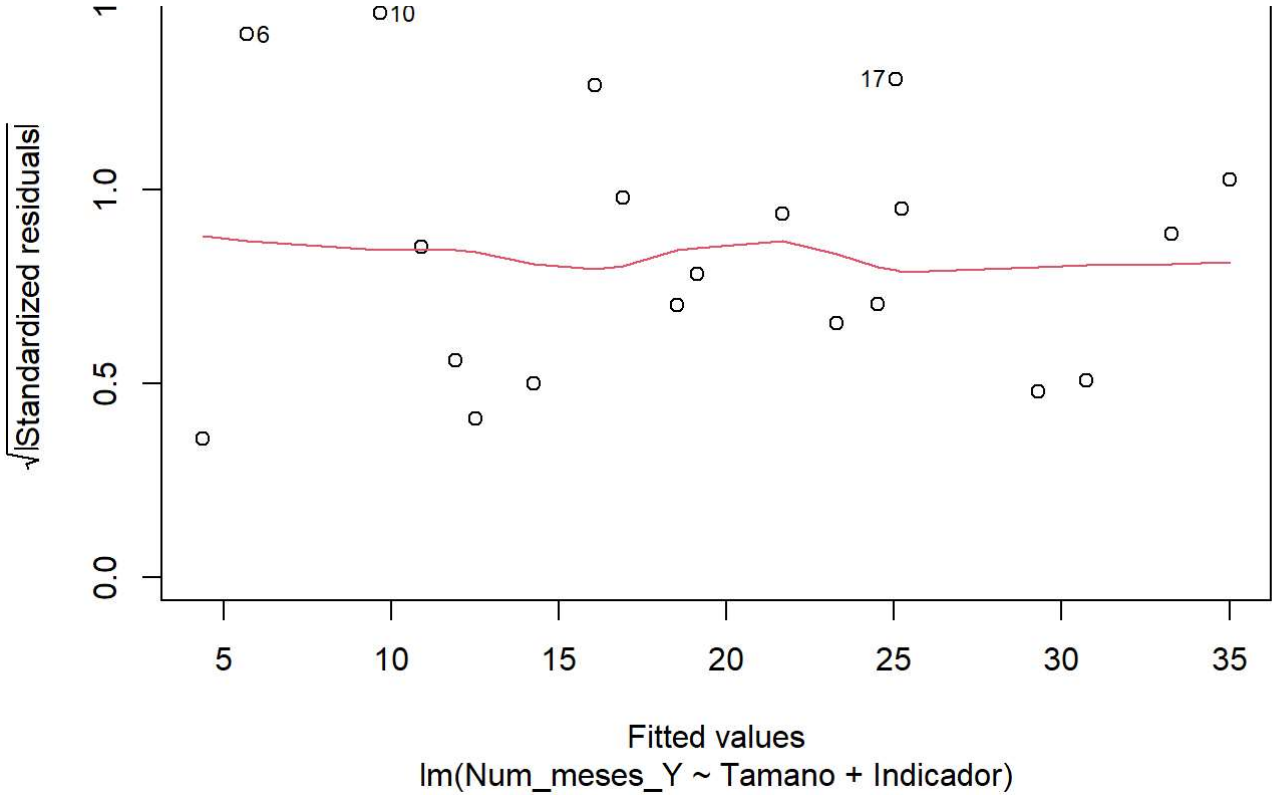
```
## The following objects are masked from Seguros:
##
##     Tamano, Tipo
```

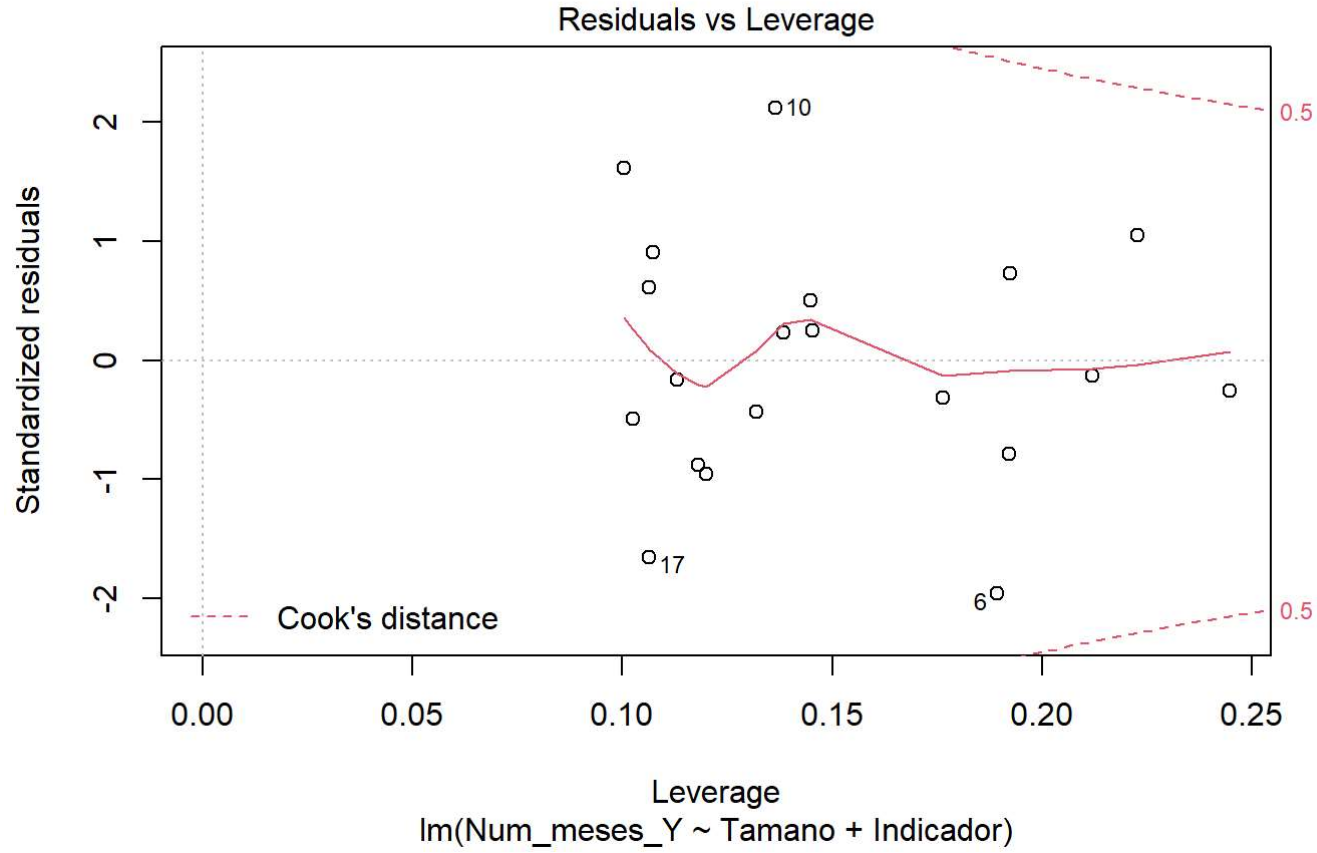
```
mod3=lm(Num_meses_Y ~ Tamano+Indicador)
plot(mod3)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Num_meses_Y ~ Tamano + Indicador)
10
6
17


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Num_meses_Y ~ Tamano + Indicador)
10
17
6


Scale-Location

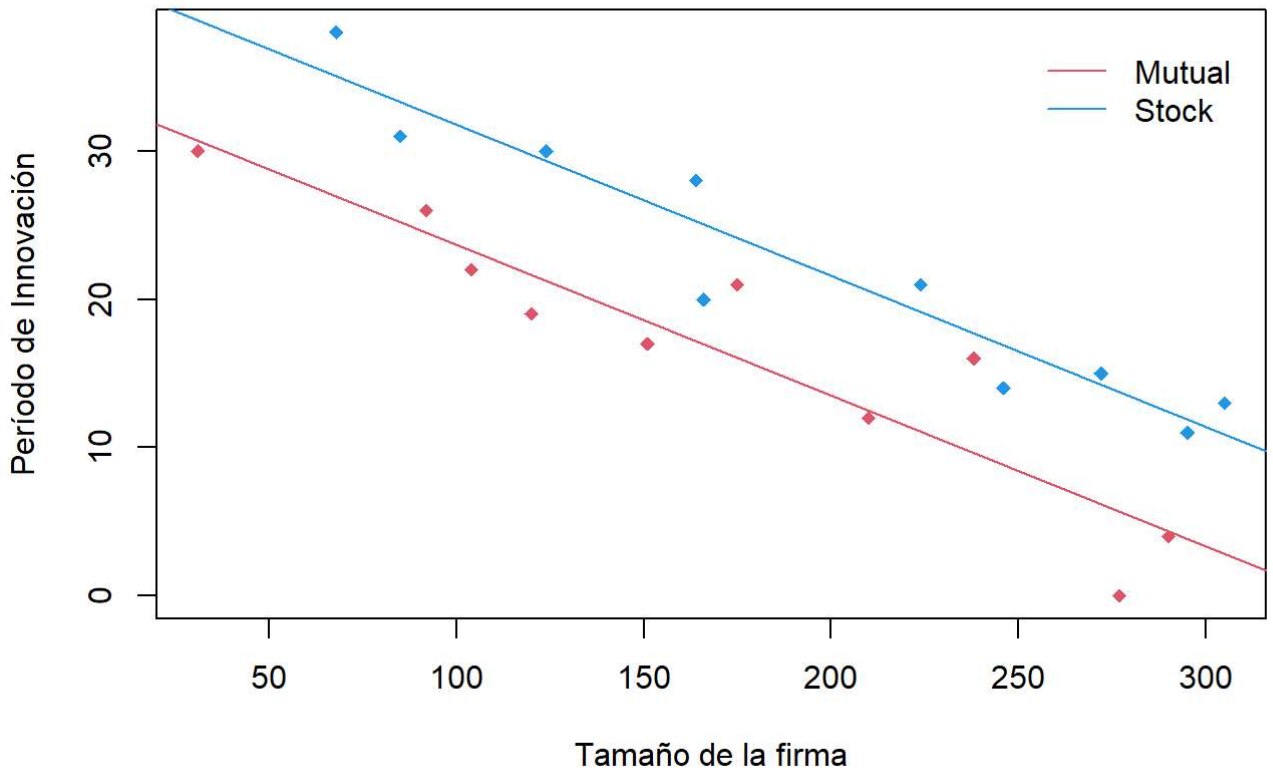





```
plot(Tamano,Num_meses_Y,col=kol,xlab="Tamaño de la firma",ylab="Período de Innovación",pch=18
)
(beta=mod3$coef)
```

```
## (Intercept)      Tamano   Indicador
##  33.8740690  -0.1017421   8.0554692
```

```
abline(beta[1],beta[2],col=2)
abline(beta[1]+beta[3],beta[2],col=4)
legend(250,max(Num_meses_Y),c("Mutual","Stock"),bty="n",lty=1,col=c(2,4))
```



2. Análisis Inferencial de los Modelos Explicativos ###### 2.1. Aspectos Generales Nótese que mod3 permite expresar de forma general la ecuación de regresión. A continuación, se presentarán por separado las ecuaciones de regresión $E[Y] = \beta_0 + \beta_1 X_1$ (aquí $X_2 = 0$) y $E[Y] = (\beta_0 + \beta_2) + \beta_1 X_1$ (aquí $X_2 = 1$) mediante las sintaxis mod4=lm(Num_meses_Y~Tamano,Seguros[Tipo==“Mutual”,]) y mod5=lm(Num_meses_Y~Tamano,Seguros[Tipo==“Stock”,]), respectivamente; nótese que estas sintaxis son modificaciones de las sintaxis antes utilizadas para realizar el análisis gráfico antes expuesto.

```
mod4=lm(Num_meses_Y~Tamano,Seguros2[Tipo=="Mutual",])
summary(mod4)
```

```
##
## Call:
## lm(formula = Num_meses_Y ~ Tamano, data = Seguros2[Tipo == "Mutual",
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7144 -1.4502 -0.6039  1.0282  6.3259
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.83837    2.73918  12.353 1.72e-06 ***
## Tamano      -0.10153    0.01465  -6.931 0.000121 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.726 on 8 degrees of freedom
## Multiple R-squared:  0.8572, Adjusted R-squared:  0.8394
## F-statistic: 48.04 on 1 and 8 DF,  p-value: 0.0001207
```

```
mod5=lm(Num_meses_Y~Tamano,Seguros2[Tipo=="Stock",])
summary(mod5)
```

```
##
## Call:
## lm(formula = Num_meses_Y ~ Tamano, data = Seguros2[Tipo == "Stock",
##     ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.046 -1.952  0.716  2.060  2.963
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41.96962    2.33998  17.936 9.57e-08 ***
## Tamano      -0.10195    0.01108  -9.205 1.57e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.857 on 8 degrees of freedom
## Multiple R-squared:  0.9137, Adjusted R-squared:  0.9029
## F-statistic: 84.73 on 1 and 8 DF,  p-value: 1.57e-05
```

Como señalan (Kutner, Nachtsheim, Neter & Li. p.318), “El lector puede preguntarse por qué no ajustamos simplemente regresiones separadas para las sociedades anónimas y las mutuales en nuestro ejemplo, y en su lugar adoptamos el enfoque de ajustar una regresión con una variable indicatriz. Hay dos razones para esto. Dado que el modelo asume pendientes iguales y la misma varianza del término de error constante para cada tipo de empresa, la pendiente común β_1 se puede estimar mejor agrupando los dos tipos de empresas. Además, otras inferencias, tal como las hechas para β_0 y β_2, se pueden hacer de manera más precisa trabajando con un modelo de regresión que contenga una variable indicatriz, ya que entonces se asociarán más grados de libertad con MSE.” Lo anterior lo afirman los autores por cuanto la MSE es el resultado de dividir la suma de cuadrados totales residuales (SCTR) entre el número de grados de libertad, obtenidos a través de la diferencia entre n-k, donde n es el tamaño de muestra y k el número de parámetros a estimar. De

ello se desprende que una variable indicatriz consumirá más grados de libertad por cuanto, como se señala en https://stats.stackexchange.com/questions/187667/degrees-of-freedom-binary-vector (https://stats.stackexchange.com/questions/187667/degrees-of-freedom-binary-vector), “Un vector binario de longitud N tiene 2^N estados posibles, por lo que una distribución de probabilidad sobre todos los estados posibles requiere la especificación de las respectivas 2^N probabilidades. Sin embargo, dado que las probabilidades deben sumar uno, la especificación de cualquier probabilidad $2^N - 1$ resume completamente la distribución.” Así, los grados de libertad entre los cuales se divide la SCTR es mayor y, por consiguiente, la varianza disminuye considerablemente, lo que implica incrementos considerables en la precisión de las estimaciones; una mayor cantidad de grados de libertad se explica por el hecho de tener una muestra global más grande. Por supuesto, lo anterior es válido en el escenario de que un modelo de regresión lineal cumpla con determinados supuestos (véase https://marxianstatistics.com/2021/09/24/supuestos-del-modelo-clasico-de-regresion-lineal-y-de-los-modelos-lineales-generalizados/ (https://marxianstatistics.com/2021/09/24/supuestos-del-modelo-clasico-de-regresion-lineal-y-de-los-modelos-lineales-generalizados/)); en caso contrario, como se verá en la sección 2.4., los resultados podrían ser ambigüos porque no la estructura de datos estudiada no cumple con la suficiente precisión con la estructura relacional especificada a través de la ecuación de regresión.

2.2. Cambios en el nivel de referencia

Conceptos Preliminares

La sintaxis “contr.treatment” contrasta cada nivel del factor con el nivel base o de referencia (especificado con “base=” dentro de la sintaxis “contr.treatment()”), pero se omite el nivel base del factor. Debe tomarse en cuenta que esto no produce “contrastes” como se define en la teoría estándar para modelos lineales, ya que no son ortogonales a la intersección.

Como se señala en http://faculty.nps.edu/sebuttre/home/r/contrasts.html (http://faculty.nps.edu/sebuttre/home/r/contrasts.html), los contrastes establecidos por defecto en R determinan cómo se manejan las variables categóricas en los modelos en que se utilizan. El esquema más común en el contexto del análisis de regresión se llama “contrastes de tratamiento”: con los contrastes de tratamiento, al primer nivel de la variable categórica se le asigna el valor 0, y luego otros niveles miden el cambio desde el primer nivel. Esto es necesario porque al disponerse de k categorías, únicamente se necesitan $k - 1$ piezas de información para representar cualquiera de los valores correspondientes a las k categorías. Por ejemplo, si los valores son “Rojo”, “Blanco” y “Azul”, se podría tener una columna denominada “Rojo”, que contiene 1 para los elementos rojos y 0 para los elementos que no son rojos, y una columna denominada “Blanco,”que contiene 1 para los artículos blancos y 0 para los artículos que no son blancos. Eso es todo lo que se requiere: si se observa un elemento con ceros tanto para rojo como para blanco, debe ser azul. Entonces, se necesita una restricción en los contrastes a realizar o, dicho de otra manera, se requieren k-1 columnas para representar una variable categórica con k niveles. ####### Ejemplo Puede desearse realizar, por diversos motivos, cambios en el nivel de referencia o cambios en las categorías de referencia (“categoría” en el sentido bioestadístico, no de los tipos de estructura de datos en R) de un factor. Existen ocasiones en que un usuario de R ingresa variables de tipo factor con las categorías del factor definidas cualitativamente (i.e., como cadenas de texto) con la finalidad de, por ejemplo, hacer una regresión. En el ejemplo que aquí se ha trabajado, para el factor “Tipo.f” se tiene “Mutual” y “Stock”, pero ¿qué codificaciones le asigna R para que sea posible realizar una regresión (puesto que las regresiones son cuantitativas, necesita entonces hacer un tipo de sustitución -no confundir con cohersión- que le permita realizarla)? Esto puede determinarse utilizando la sintaxis “contrasts(Tipo.f)”. Tras ello, puede asignarse la columna “Tipo.f” del dataframe “Seguros” a un vector numérico mediante la sintaxis “Tipo.f2 = Tipo.f”, con la finalidad de facilitar los cambios en los niveles de referencia. Se puede observar que al ejecutar Posteriormente, contrasts(Tipo.f) se obtiene que el programa R asume el valor “0” para “Mutual” y “1” para “Stock”. Sin embargo, esto puede invertirse mediante el uso de la sintaxis “Tipo.f2 = Tipo.f” (que indica que “Tipo.f” se sustituirá por “Tipo.f2”) y luego redefinir la categoría base (i.e., la que será 0) con la sintaxis “(contrasts(Tipo.f2)=contr.treatment(levels(Tipo.f),base=2))” para finalmente comprobar que los cambios en las codificaciones fueron hechos efectivos puede utilizarse la sintaxis “contrasts(Tipo.f2)”.

```
attach(Seguros)
```

```
## The following objects are masked from Seguros2:
##
##     Tamano, Tipo
```

```
## The following objects are masked from Seguros (pos = 5):
##
##     Tamano, Tiempo, Tipo, Tipo.f
```

```
contrasts(Tipo.f)
```

```
##        Stock
## Mutual     0
## Stock      1
```

```
Tipo.f2 = Tipo.f
(contrasts(Tipo.f2)=contr.treatment(levels(Tipo.f),base=2))
```

```
##        Mutual
## Mutual      1
## Stock       0
```

```
contrasts(Tipo.f2)
```

```
##        Mutual
## Mutual      1
## Stock       0
```

Lo anteriormente realizado puede replicarse para el caso de un análisis de regresión lineal múltiple, por ejemplo, para estudiar la relación lineal entre el tiempo requerido por parte de una compañía de seguros para adoptar una innovación financiera. Para ello, puede construirse un modelo 6 de la forma “mod6 = lm(Tiempo ~ Tamano+Tipo.f2)”, que representa el escenario “Tipo.f2” en el que “Mutual=0” y “Stock=1”, mientras que un modelo 6.1 de la forma mod6.1 = lm(Tiempo ~ Tamano+relevel(Tipo.f, ref = “Stock”)), en donde se indica mediante la sintaxis “relevel(Tipo.f, ref =”Stock“)” que la variable cero será “Stock” en lugar de “Mutual”.

```
mod6 = lm(Tiempo ~ Tamano+Tipo.f2)
summary(mod6)
```

```
##
## Call:
## lm(formula = Tiempo ~ Tamano + Tipo.f2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6915 -1.7036 -0.4385  1.9210  6.3406
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   41.929538   2.010101  20.859 1.50e-13 ***
## Tamano        -0.101742   0.008891 -11.443 2.07e-09 ***
## Tipo.f2Mutual -8.055469   1.459106  -5.521 3.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.221 on 17 degrees of freedom
## Multiple R-squared:  0.8951, Adjusted R-squared:  0.8827
## F-statistic:  72.5 on 2 and 17 DF,  p-value: 4.765e-09
```

```
mod6.1 = lm(Tiempo ~ Tamano+relevel(Tipo.f, ref = "Stock"))
summary(mod6.1)
```

```
##
## Call:
## lm(formula = Tiempo ~ Tamano + relevel(Tipo.f, ref = "Stock"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6915 -1.7036 -0.4385  1.9210  6.3406
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        41.929538   2.010101  20.859 1.50e-13 ***
## Tamano                             -0.101742   0.008891 -11.443 2.07e-09 ***
## relevel(Tipo.f, ref = "Stock")Mutual -8.055469   1.459106  -5.521 3.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.221 on 17 degrees of freedom
## Multiple R-squared:  0.8951, Adjusted R-squared:  0.8827
## F-statistic:  72.5 on 2 and 17 DF,  p-value: 4.765e-09
```

2.3. Modelo con efectos de interacciones

Como se señala en https://statisticsbyjim.com/regression/interaction-effects/ (https://statisticsbyjim.com/regression/interaction-effects/), los efectos de interacción ocurren cuando el efecto de una variable explicativa sobre la respuesta depende del valor de otra variable explicativa. Los efectos de interacción son comunes en el análisis de regresión, ANOVA y diseño de experimentos. En cualquier estudio, ya sea una prueba de sabor o un proceso de fabricación, muchas variables pueden afectar el resultado. Cambiar estas variables puede afectar el resultado directamente. Por ejemplo, cambiar el condimento de la comida en una prueba de sabor puede afectar el disfrute general. De esta manera, los analistas utilizan modelos para evaluar la relación entre cada variable independiente y la variable dependiente. Este tipo de efecto se llama **efecto principal**. Sin embargo, puede ser un error evaluar solo los efectos principales de un

modelo. En áreas de estudio más complejas, las variables independientes pueden interactuar entre sí. Los efectos de interacción indican que una tercera variable influye en la relación entre una variable independiente y una dependiente. Este tipo de efecto, conocido como **efecto de interacción**, hace que el modelo sea más complejo, pero si el mundo real (lo que subyace a la realidad, que es lo real observado) se comporta de esta manera, es fundamental incorporarlo en el modelo. Por ejemplo, la relación entre condimentos y placer probablemente dependa del tipo de comida, por lo que podría adoptar la forma "Satisfacción = Tipo de Comida + Condimentación + Tipo de Comida*Condimentación".

Por ejemplo, utilizando la base de datos inicial “Seguros.Rdata”, puede plantearse una interacción de la forma $Y_2 = Y + X_2 * X_1$, mediante la sintaxis “Tiempo2=Tiempo+(Tipo.f==”Mutual“)*Tamano*0.05” para el caso de empresas mutuales y estudiar esta relación con interacciones mediante un análisis de regresión lineal múltiple usando la sintaxis “mod7 = lm(Tiempo2 ~ Tamano+Tipo.f+Tamano:Tipo.f)”.

```
attach(Seguros)
```

```
## The following objects are masked from Seguros (pos = 3):
##
##     Tamano, Tiempo, Tipo, Tipo.f
```

```
## The following objects are masked from Seguros2:
##
##     Tamano, Tipo
```

```
## The following objects are masked from Seguros (pos = 6):
##
##     Tamano, Tiempo, Tipo, Tipo.f
```

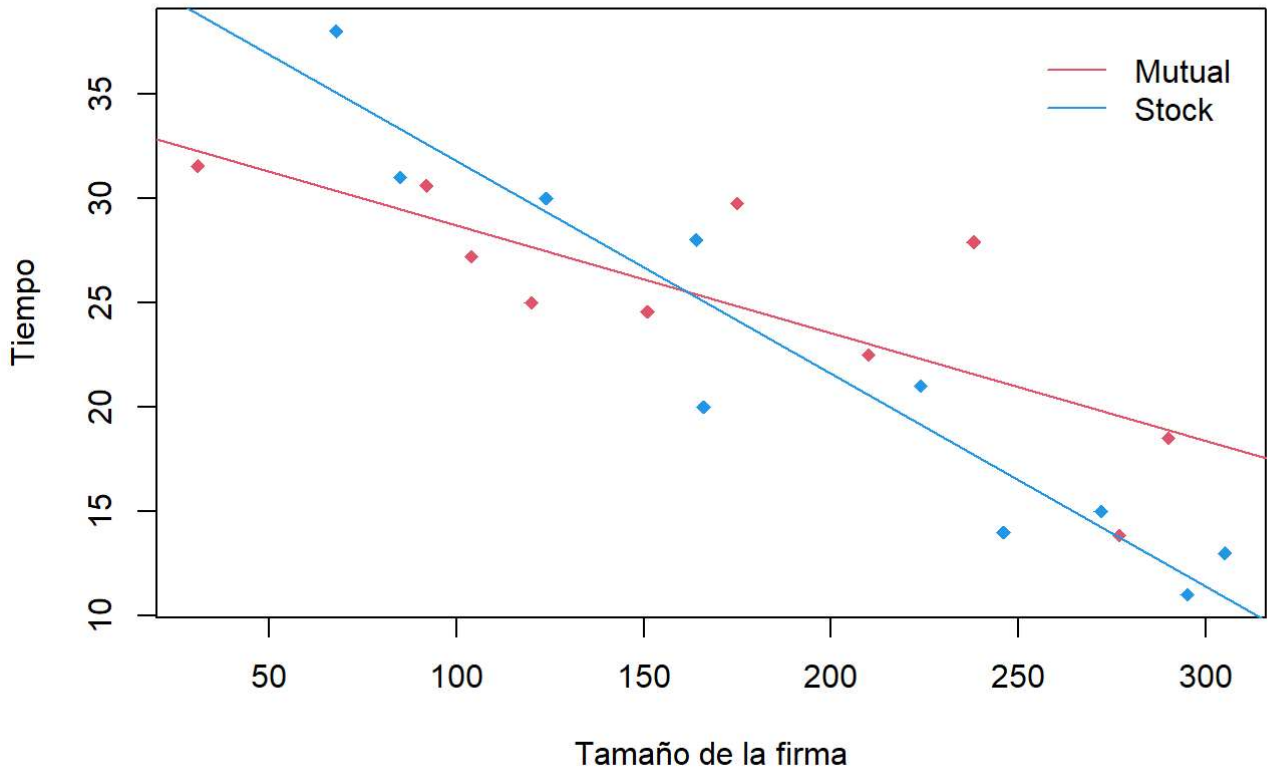
```
contrasts(Tipo.f) = contr.treatment(levels(Tipo.f),base=1)
str(Seguros)
```

```
## 'data.frame':    20 obs. of  4 variables:
##  $ Tiempo: num  17 26 21 30 22 0 12 19 4 16 ...
##  $ Tamano: num  151 92 175 31 104 277 210 120 290 238 ...
##  $ Tipo  : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Tipo.f: Factor w/ 2 levels "Mutual","Stock": 1 1 1 1 1 1 1 1 1 1 ...
```

```
Tiempo2=Tiempo+(Tipo.f=="Mutual")*Tamano*0.05
mod7=lm(Tiempo2 ~ Tamano+Tipo.f+Tamano:Tipo.f)
summary(mod7)
```

```
##
## Call:
## lm(formula = Tiempo2 ~ Tamano + Tipo.f + Tamano:Tipo.f)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7144 -1.7064 -0.4557  1.9311  6.3259
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        33.83837    2.44065  13.864 2.47e-10 ***
## Tamano             -0.05153    0.01305  -3.948  0.00115 **
## Tipo.fStock         8.13125    3.65405   2.225  0.04079 *
## Tamano:Tipo.fStock -0.05042    0.01833  -2.750  0.01422 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.32 on 16 degrees of freedom
## Multiple R-squared:  0.8376, Adjusted R-squared:  0.8071
## F-statistic: 27.51 on 3 and 16 DF,  p-value: 1.495e-06
```

```
plot(Tamano,Tiempo2,col=kol,xlab="Tamaño de la firma",ylab="Tiempo",pch=18)
beta7=mod7$coef
abline(beta7[1],beta7[2],col=2)
abline(beta7[1]+beta7[3],beta7[2]+beta7[4],col=4)
legend(250,max(Tiempo),c("Mutual","Stock"),bty="n",lty=1,col=c(2,4))
```



En la función “Tiempo2=Tiempo+(Tipo.f==”Mutual“)*Tamano*0.05” el elemento “(Tipo.f==”Mutual“)” lleva el símbolo “==” porque, como se señala en https://towardsdatascience.com/the-ultimate-guide-to-relational-operators-in-r-6d8489d9d947 (https://towardsdatascience.com/the-ultimate-guide-to-relational-operators-in-r-6d8489d9d947), “Los operadores relacionales, o comparadores, son operadores que nos ayudan a ver cómo un objeto R se relaciona con otro. Por ejemplo, puede verificarse si dos objetos son iguales (igualdad) usando un signo igual doble ==. El resultado de la consulta de igualdad es un valor lógico (TRUE o FALSE)”. Así, en caso la categoría en la que se ubique alguna observación particular de la variable “Tipo.f” sea la de “Mutual”, la variable “Tamano” será multiplicada por 0.5 (que es el peso de la categoría “Mutual” de las dos categorías que puede adoptar la variable dicotómica).

Los resultados anteriores muestran que para diferentes tamaños de la firma, el tiempo que tarda un tipo de firma en absorber una innovación de la forma antes definida (con efectos de interacción) puede ser mayor que el que dura otra firma en adoptar una innovación.

A pesar de ser un ejercicio pedagógico debe decirse (en función de evitar equívocos) que el 0.5 que multiplica a “Tamano” no representa la probabilidad de que la empresa seleccionada sea “Mutual” o “Stock” (lo que podría derivarse de una interpretación rigurosamente frecuentista), sino simplemente su peso cuantitativo en la muestra tomada de la variable “Tipo”.

2.4. Modelos de efectos separados

La variable antes definida como “Tiempo2=Tiempo+(Tipo.f==”Mutual“)*Tamano*0.05” puede representarse mediante modelos separados. Para llevar a cabo esto, se construirán los modelos mod8a y mod8b, en donde mod8a representa el escenario en el que la observación de “Tiempo2” corresponde a una compañía de tipo “Mutual” y mod8b representa el escenario en el que la observación “Tiempo2” corresponde a una compañía de tipo “Stock” o de acciones.

```
Seguros3=data.frame(Seguros,Tiempo2)
attach(Seguros3)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     Tiempo2, Tipo.f
```

```
## The following objects are masked from Seguros (pos = 3):
##
##     Tamano, Tiempo, Tipo, Tipo.f
```

```
## The following objects are masked from Seguros (pos = 4):
##
##     Tamano, Tiempo, Tipo, Tipo.f
```

```
## The following objects are masked from Seguros2:
##
##     Tamano, Tipo
```

```
## The following objects are masked from Seguros (pos = 7):
##
##     Tamano, Tiempo, Tipo, Tipo.f
```

```
mod8a=lm(Tiempo2~Tamano,Seguros3[Tipo.f=="Mutual",])
mod8b=lm(Tiempo2~Tamano,Seguros3[Tipo.f=="Stock",])
summary(mod8a)
```

```
##
## Call:
## lm(formula = Tiempo2 ~ Tamano, data = Seguros3[Tipo.f == "Mutual",
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7144 -1.4502 -0.6039  1.0282  6.3259
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.83837    2.73918  12.353 1.72e-06 ***
## Tamano      -0.05153    0.01465  -3.518  0.00787 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.726 on 8 degrees of freedom
## Multiple R-squared:  0.6073, Adjusted R-squared:  0.5583
## F-statistic: 12.37 on 1 and 8 DF,  p-value: 0.007873
```

```
summary(mod8b)
```

```
##
## Call:
## lm(formula = Tiempo2 ~ Tamano, data = Seguros3[Tipo.f == "Stock",
##     ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.046 -1.952  0.716  2.060  2.963
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41.96962    2.33998  17.936 9.57e-08 ***
## Tamano      -0.10195    0.01108  -9.205 1.57e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.857 on 8 degrees of freedom
## Multiple R-squared:  0.9137, Adjusted R-squared:  0.9029
## F-statistic: 84.73 on 1 and 8 DF,  p-value: 1.57e-05
```

Nótese que al comparar las ecuaciones de regresión de los modelos separados para “Tiempo2” con la ecuación de regresión que considera efectos de interacción los resultados son equivalentes en términos de sus coeficientes de regresión. Sin embargo, nótese también que los errores estándar son diferentes, así como también los coeficientes de determinación de Pearson R^2. Así, como se adelantó, el modelo que emplea más variables no tiene una menor varianza del error (MSE, que es la desviación del error) que ambos modelos que emplean menos variables, puesto que mod7 únicamente tiene menor MSE que el modelo 8a, mientras que su MSE es mayor que el del modelo 8b.

2.4. Modelos de Múltiples Categorías

Introducción

Como se señala en https://bookdown.org/jboscomendoza/r-principiantes4/coercion.html (https://bookdown.org/jboscomendoza/r-principiantes4/coercion.html), en R los datos pueden ser coercionados, es decir, forzados, para transformarlos de un tipo a otro. La coerción es muy importante. Cuando se solicita a R ejecutar una operación, intentará coercionar de manera implícita, sin avisar a al usuario, los datos de su tipo original al tipo correcto que permita realizarla. Habrá ocasiones en las que R tenga éxito y la operación ocurra sin problemas, y otras en las que falle y se obtenga un error. Lo anterior ocurre porque no todos los tipos de datos pueden ser transformados a los demás, para ello se sigue una regla general: $lógico- > entero- > numérico- > texto$. La coerción de tipos se realiza de los tipos de datos más restrictivos a los más flexibles. Las coerciones no pueden ocurrir en orden inverso, es decir, es posible coercionar un dato de tipo entero a uno numérico, pero no uno de cadena de texto a numérico. Como los datos de tipo lógico sólo admiten dos valores (TRUE y FALSE), estos son los más restrictivos; mientras que los datos de cadena de texto, al admitir cualquier cantidad y combinación de caracteres, son los más flexibles. Los factores son un caso particular para la coerción. Dado que son valores numéricos con etiquetas, pueden ser coercionados a tipo numérico y cadena de texto; y los datos numéricos y cadena de texto pueden ser coercionados a factor. Sin embargo, al coercionar un factor tipo numérico, se pierden sus niveles. A continuación, se presentan los tipos de coerción explícita que es posible realizar con la familia de sintaxis conocida como “as()”.

```
knitr::include_graphics("CUADRO1.JPG")
```

Función	Tipo al que hace coerción
as.integer()	Entero
as.numeric()	Numerico
as.character()	Cadena de texto
as.factor()	Factor
as.logical()	Lógico
as.null()	NULL

```
#Figura 5: Coerción explícita con la familia "as()"
#Fuente: https://bookdown.org/jboscomendoza/r-principiantes4/coercion.html
```

De lo anterior se desprende que los factores en R son estructuras de datos utilizadas para manipular dos o más variables categóricas, aquellos factores generados mediante la coerción “as.factor()” no son la excepción. Esto significa que para coercionar una variable a factor hace falta involucrar a otra variable, lo que debe reflejarse en la construcción específica de la sintaxis “as.factor()”. Por ejemplo, se podría tener interés en vincular un determinado tipo de prueba de resistencia (“trat”) realizada a distintos tipos de tela de algodón de algún peso (“peso”) con la finalidad de determinar su resistencia “resist”. Con este fin puede invocarse al vector numérico columna “resist” de forma directa (puesto que está almacenado en el mismo directorio especificado al inicio del documento, al igual que las imágenes insertadas) mediante la sintaxis “base=read.table(”algodon.txt“,col.names=”resist“)”. Posteriormente, puede crearse una variable con (por ejemplo) cinco números cualesquiera que se repitan cinco veces cada uno mediante la sintaxis “base=read.table(”algodon.txt“,col.names=”resist“)”; esto sentará las bases para construir un predictor de tipo factor en el siguiente paso. Acto seguido, es posible crear una variable “trat” vinculada explícitamente con el peso de la tela de algodón sometida a prueba usando la sintaxis “as.factor(base$peso)”. Al usar esta sintaxis, se está diciendo que la variable “trat” es una variable de tipo factor en función de “peso”, lo que implica (puesto que “peso” se construyó de tal forma que se repitiesen los cinco números seleccionados de cinco en cinco -con el fin que coincidiera con el tamaño de la base de datos invocada-) que será una variable de tipo

factor con cinco niveles, lo cual se puede verificar mediante la sintaxis “str(base)”. Debido a que se trabajará con una nueva base de datos y antes se habían almacenado varias y computado distintos modelos, podría ser recomendable reiniciar el entorno global mediante la sintaxis “rm(list=ls(all=T))”.

Tras realizar lo anterior, puede construirse un modelo de regresión lineal simple que explique la resistencia “resist” en función del tratamiento aplicado “trat” mediante la sintaxis “mod9 = lm(resist~trat)” y posteriormente utilizar la sintaxis “tapply(resist,trat,mean)”, esto último con el fin de crear un resumen de las variables involucradas en el modelo basado en los niveles del factor “trat” que se creó; en este caso, el resumen consiste en que obtenga la media “mean” de cada uno de los niveles del factor creado.

```
rm(list=ls(all=T))
base=read.table("algodon.txt",col.names="resist")
base$peso=rep(c(15,20,25,30,35),each=5)
base$trat=as.factor(base$peso)
str(base)
```

```
## 'data.frame':    25 obs. of  3 variables:
##  $ resist: int  7 7 15 11 9 12 17 12 18 18 ...
##  $ peso  : num  15 15 15 15 15 20 20 20 20 20 ...
##  $ trat  : Factor w/ 5 levels "15","20","25",..: 1 1 1 1 1 2 2 2 2 2 ...
```

```
head(base)
```

```
##   resist peso trat
## 1      7   15   15
## 2      7   15   15
## 3     15   15   15
## 4     11   15   15
## 5      9   15   15
## 6     12   20   20
```

```
attach(base)
mod9 = lm(resist~trat)
anova(mod9)
```

```
## Analysis of Variance Table
##
## Response: resist
##           Df Sum Sq Mean Sq F value    Pr(>F)
## trat       4 475.76  118.94  14.757 9.128e-06 ***
## Residuals 20 161.20    8.06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(mod9)
```

```
##
## Call:
## lm(formula = resist ~ trat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -3.8   -2.6    0.4    1.4    5.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.800      1.270   7.719 2.02e-07 ***
## trat20         5.600      1.796   3.119 0.005409 **
## trat25         7.800      1.796   4.344 0.000315 ***
## trat30        11.800      1.796   6.572 2.11e-06 ***
## trat35         1.000      1.796   0.557 0.583753
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.839 on 20 degrees of freedom
## Multiple R-squared:  0.7469, Adjusted R-squared:  0.6963
## F-statistic: 14.76 on 4 and 20 DF,  p-value: 9.128e-06
```

```
tapply(resist,trat,mean)
```

```
##   15   20   25   30   35
##  9.8 15.4 17.6 21.6 10.8
```

Finalmente, como se señala en https://wikidiff.com/category/class (https://wikidiff.com/category/class), la diferencia entre categoría y clase es que la categoría es un grupo, a menudo nombrado o numerado, al que se asignan elementos (en este caso observaciones numéricas) en función de la similitud o criterios definidos (no deben ser grupos cuyos elementos sean numerables) , mientras que la clase es un grupo, colección, categoría o conjunto rigurosamente numerable que comparte características o atributos. En las sintaxis antes expuestas el criterio definido para los grupos fue que se repitiesen en orden y cinco veces cada uno de los números que se seleccionó a discreción. Estas categorías pueden expresarse como clases, es decir, como agrupaciones de observaciones que comparten la característica de ser iguales a determinado valor y que posean alguna jerarquía definida, la cual para este caso viene dada por el orden de los números naturales. Así, tomando como base la categoría de los elementos iguales a 15, la primera clase será la conformada por los valores iguales a 20, la segunda por los iguales a 25, la tercera por los iguales a 30 y la cuarta por los iguales a 35.

Esto tiene como finalidad dos cosas. La primera es introducir a los conceptos de “categoría” en la Bioestadística y al concepto de “clase”, que es un concepto que históricamente nace en la filosofía, llega a las ciencias formales a través de las matemáticas y en la actualidad es utilizado en diversas ramas de las ciencias, sea por su vinculación con las metodologías cuantitativas o heredadas directamente de alguna rama filosófica. La segunda es utilizar tales clases para construir un modelo de regresión lineal múltiple que explique la respuesta “resist” en función de las mismas y comparar su ajuste con el ajuste del modelo que utiliza directamente al factor (i.e., con mod9) compuesto por 5 categorías.

Para construir las clases se requiere, además de las categorías ya establecidas, de la función indicatriz y del símbolo “==”. La función indicatriz 1_A, construida mediante la sintaxis "1*()“, asigna el valor de 0 a un elemento que no pertenece a determinada agrupación de elementos y el valor de 1 cuando sí pertenece a dicha agrupación (en lenguaje computacional, (la sintaxis anterior se usa para codificar las observaciones en términos de 1 y 0), lo cual tiene como finalidad construir vectores de datos que, dado su papel como variables

explicativas en el modelo de regresión, aporten valores nulos en las casillas correspondientes a valores que no pertenecen a la clase i-ésima (donde i=1,2,3,4); esto equivale a que cada $clase$ (que son vectores numéricos), aunque cuente dentro de sí con un conjunto de 25 obseraciones, no aporta valor explicativo en aquellas casillas de su respectivo vector numérico que provienen de elementos que no pertenecen a la clase (y por consiguiente, tampoco a la categoría con base en la cual se construyó). Esto es posible debido a que un valor nulo es diferente de vacío o un N/A, por lo que sí cuenta como observación y ese es el sentido de la manipulación antes hecha. Finalmente, mediante el uso del signo doble igual"==" se condiciona la pertenencia a la clase con base en el criterio antes definido.

```
trat
```

```
##  [1] 15 15 15 15 15 20 20 20 20 20 25 25 25 25 25 30 30 30 30 30 35 35 35 35 35
## Levels: 15 20 25 30 35
```

```
CLASS1 = 1*(trat==20)
CLASS2 = 1*(trat==25)
CLASS3 = 1*(trat==30)
CLASS4 = 1*(trat==35)
head(cbind(resist,CLASS1,CLASS2,CLASS3,CLASS4))
```

```
##      resist CLASS1 CLASS2 CLASS3 CLASS4
## [1,]      7      0      0      0      0
## [2,]      7      0      0      0      0
## [3,]     15      0      0      0      0
## [4,]     11      0      0      0      0
## [5,]      9      0      0      0      0
## [6,]     12      1      0      0      0
```

```
mod10 = lm(resist~CLASS1+CLASS2+CLASS3+CLASS4)
summary(mod10)
```

```
##
## Call:
## lm(formula = resist ~ CLASS1 + CLASS2 + CLASS3 + CLASS4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -3.8   -2.6    0.4    1.4    5.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.800      1.270   7.719 2.02e-07 ***
## CLASS1         5.600      1.796   3.119 0.005409 **
## CLASS2         7.800      1.796   4.344 0.000315 ***
## CLASS3        11.800      1.796   6.572 2.11e-06 ***
## CLASS4         1.000      1.796   0.557 0.583753
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.839 on 20 degrees of freedom
## Multiple R-squared:  0.7469, Adjusted R-squared:  0.6963
## F-statistic: 14.76 on 4 and 20 DF,  p-value: 9.128e-06
```

Como se verifica, los modelos son rigurosamente idénticos o, lo que es lo mismo, perfectamente equivalentes. Así, se dejan sentadas las condiciones para estudiar la temática de las medias marginales, es decir, los promedios de la variable dependiente según diferentes niveles de una o más variables predictoras categóricas.