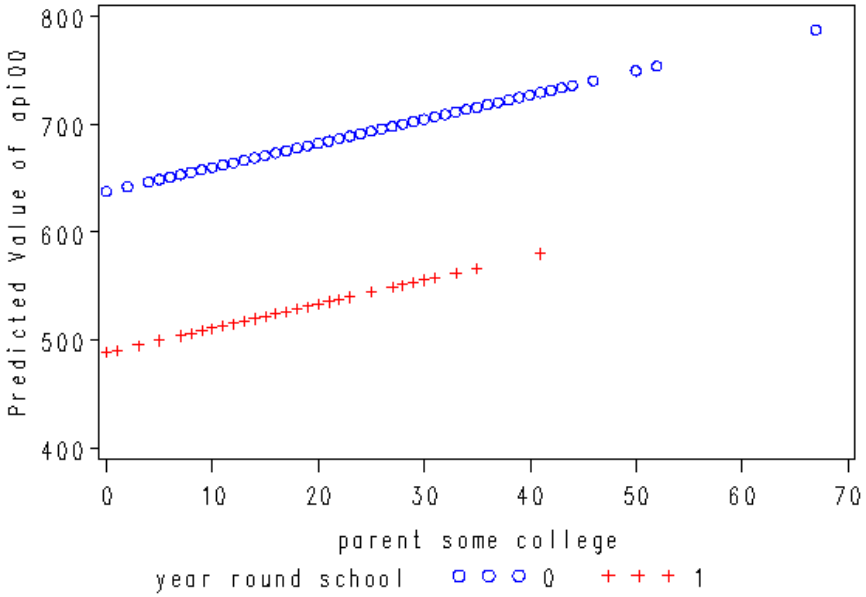
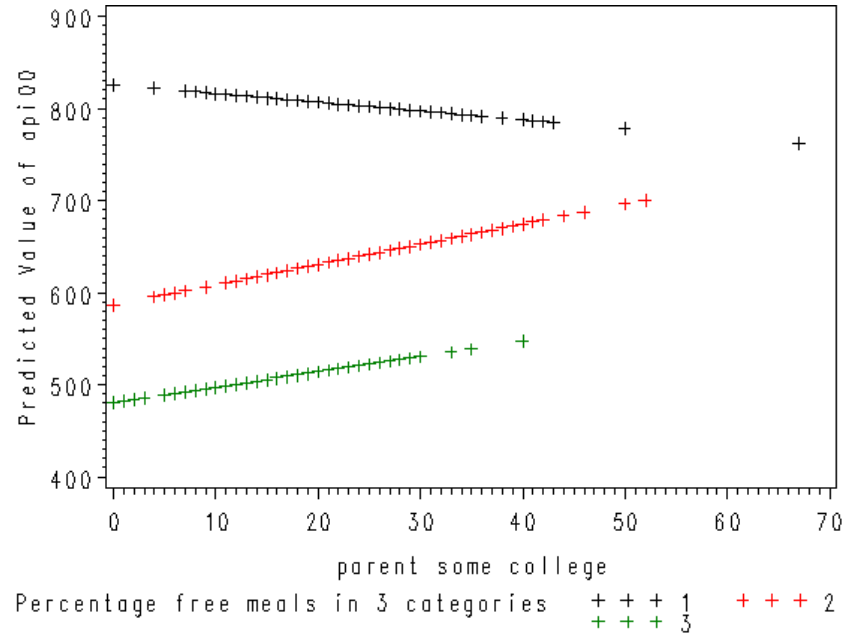




Regresión con variables independientes cualitativas

Óscar Centeno Mora

Preámbulo

- Hasta ahora hemos visto una RLM únicamente con variables cuantitativas, tanto en la variable dependiente, así como en sus predictores.

- ¿Podemos aplicar una RLM con variables cualitativas?

- ¿Variables cualitativas para la variable dependiente o las independientes? Se puede aplicar la técnica de regresión en ambos casos.

- El en presente capítulo nos centraremos en una RLM, en donde esta posee variables cualitativas en sus predictores (variables independientes).

- Más adelante veremos dos métodos de regresión para variables dependientes categóricas: el regresión Logit y la regresión Probit.

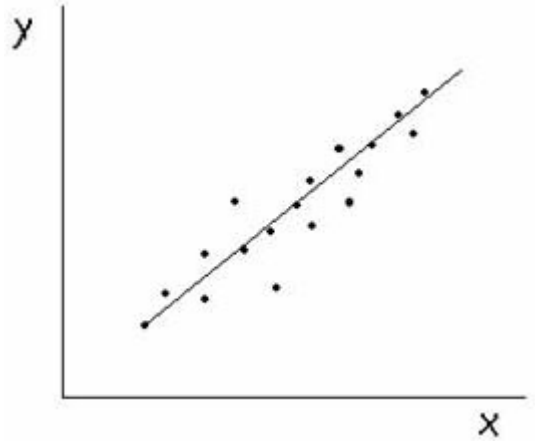
Preámbulo

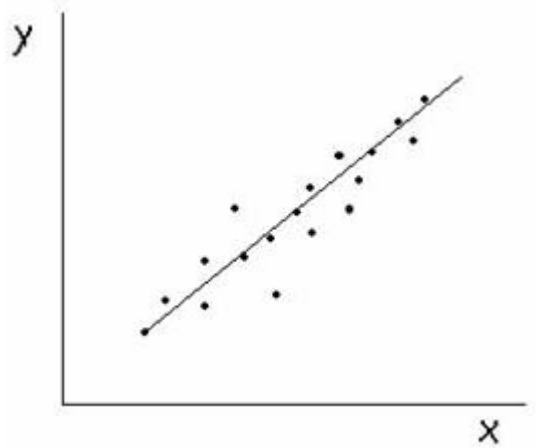
Regresión bivariada

Una variable dependiente (Y)
Una variable independiente (X)

$$Y \sim X$$

$$\begin{bmatrix} Y1 \\ Y2 \\ \cdots \\ Y_n \end{bmatrix} \sim \begin{bmatrix} X1 \\ X2 \\ \cdots \\ X_n \end{bmatrix}$$





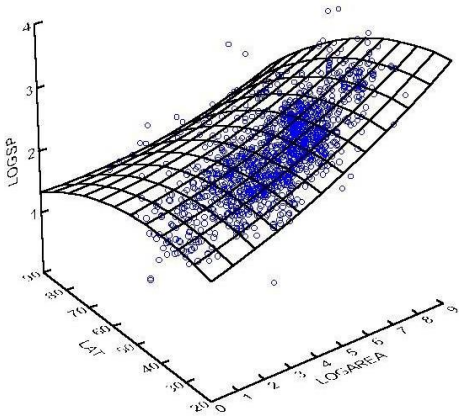
Regresión multivariada

Una variable dependiente
Dos o más variables independientes

$$Y \sim X$$

$$\begin{bmatrix} Y1 \\ Y2 \\ \cdots \\ Y_n \end{bmatrix} \sim \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

¿ cómo podemos analizar una RLM con variables cualitativas en la matriz de X?



Índice

Índice

Introducción

- Al trabajar con variables cualitativas en los predictores, se debería llegar a:

1. Incluir e interpretar variables categóricas en un modelo de regresión lineal mediante variables ficticias.
2. Comprender las implicaciones de usar un modelo con una variable categórica de dos maneras: niveles que sirven como predictores únicos versus niveles que sirven como comparación con una línea de base.
3. Construya e interprete modelos de regresión lineal con términos de interacción.
4. Aplicar la RLM a variables categóricas con 3 o más niveles.

- El uso de las variables cualitativas en la RLM es un caso de aplicar un total de k-1 rectas de RLM, en donde el efecto será analizado a partir de la interacción.

- Empezaremos un el modelo más simple: una RLM con una variable cualitativa de dos categorías.

Índice

Variable categórica de dos niveles

- Usaremos el conjunto de datos integrado ***mtcars*** antes de regresar a nuestro conjunto de datos ***autompg***. El conjunto de datos de mtcars es algo más pequeño, por lo que echaremos un vistazo rápidamente a todo el conjunto de datos.

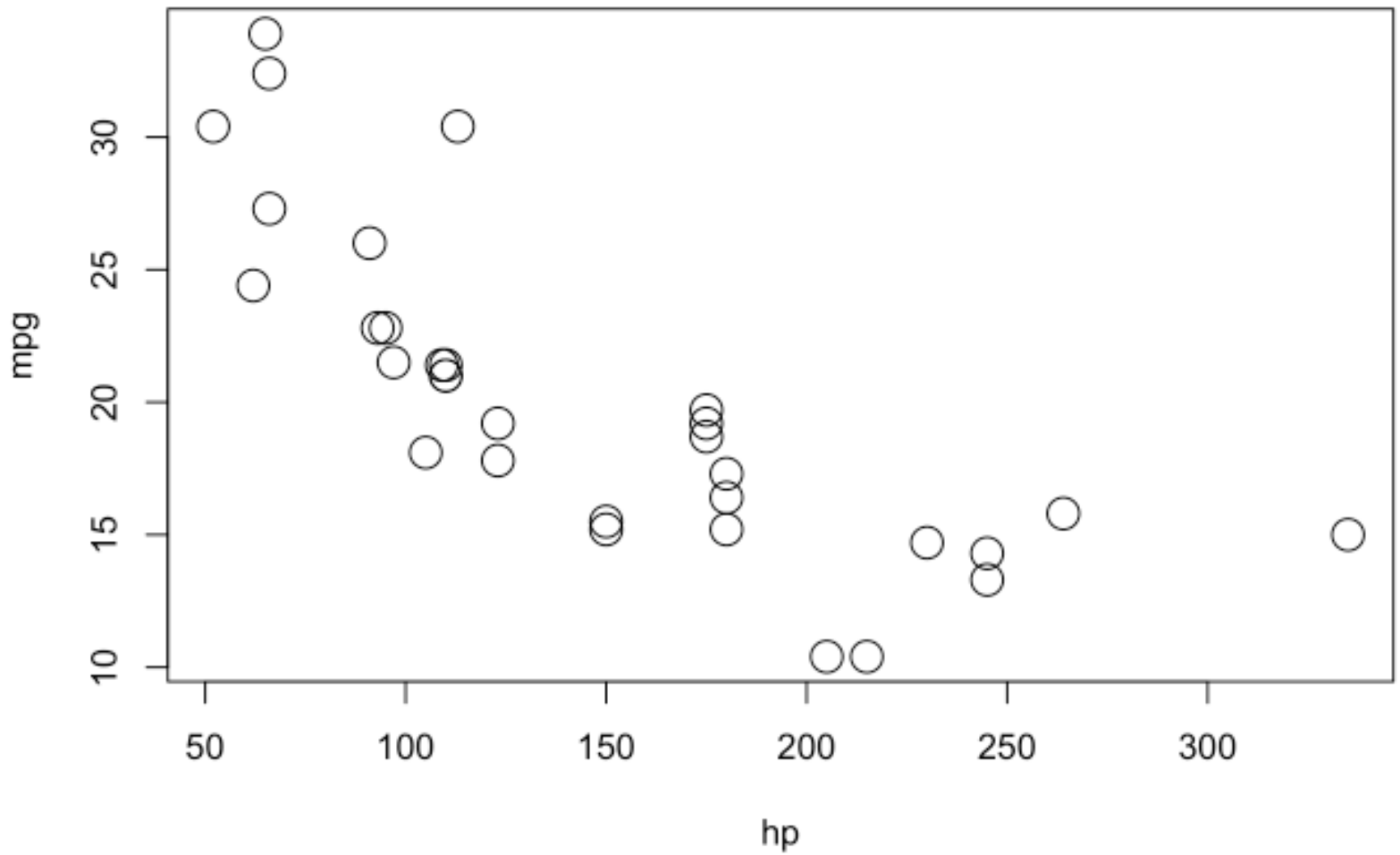
- Nos interesarán tres de las variables:
  - mpg: eficiencia de combustible, en millas por galón.
  - hp: caballos de fuerza, en libras-pie por segundo.
  - am: transmisión. Automático o manual.

- Estamos interesados en mpg como variable dependiente, comenzaremos trazando los datos respuesta y hp como predictor.

- Veamos el gráfico.

```
plot(mpg ~ hp, data = mtcars, cex = 2)
```

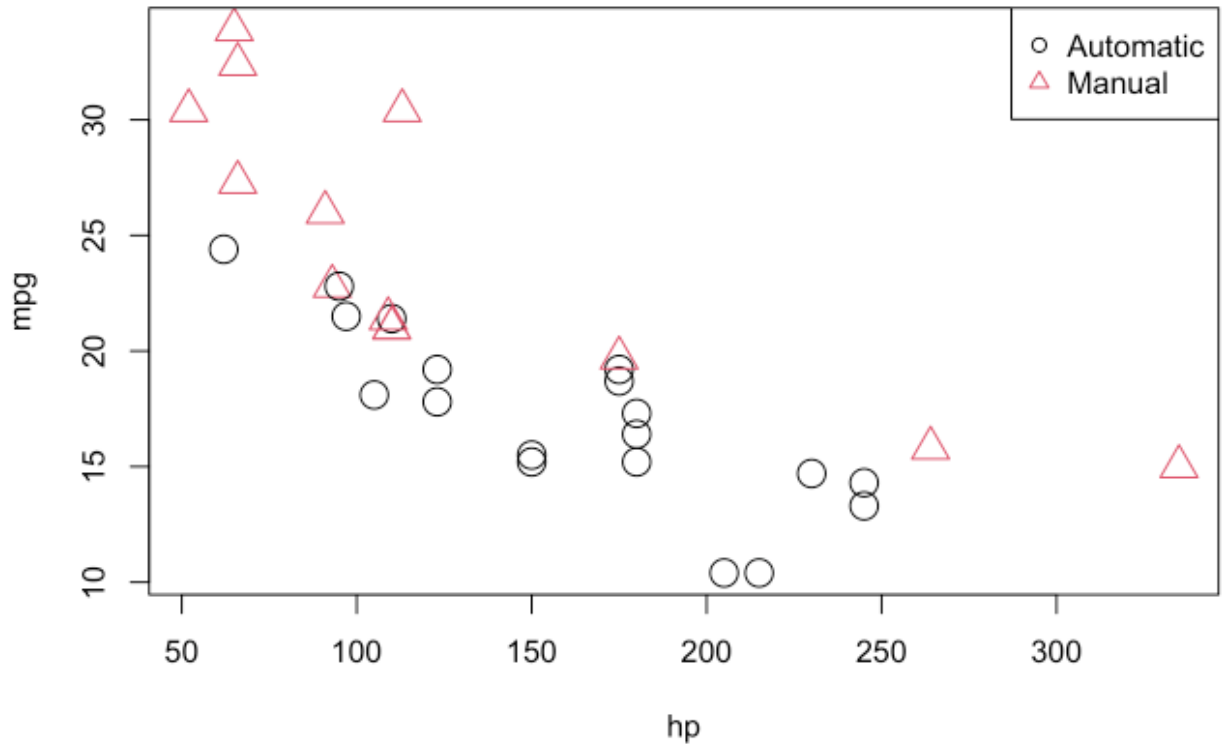
Variable categórica de dos niveles



Variable categórica de dos niveles

- Dado que también estamos interesados en el tipo de transmisión (variable am), también podríamos etiquetar los puntos en consecuencia.

```
plot(mpg ~ hp, data = mtcars, col = am + 1, pch = am + 1, cex = 2)
legend("topright", c("Automatic", "Manual"), col = c(1, 2), pch = c(1, 2))
```



Variable categórica de dos niveles

- Usamos un "truco" común de R al graficar estos datos. La variable am toma dos valores posibles; 0 para transmisión automática y 1 para transmisiones manuales. R puede usar números para representar colores, sin embargo, el color del 0 es el blanco. Así que tomamos el vector am y le sumamos 1. Entonces, las observaciones con transmisiones automáticas ahora están representadas por 1, que es negro en R, y la transmisión manual está representada por 2, que es rojo en R. (Tenga en cuenta, solo estamos agregando 1 dentro de la llamada a plot (), no estamos modificando realmente los valores almacenados en am.)
- Ahora nos adaptamos al modelo de la RLM:

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon$$

En donde Y sería la variable mpg, y la x1 sería el hp.

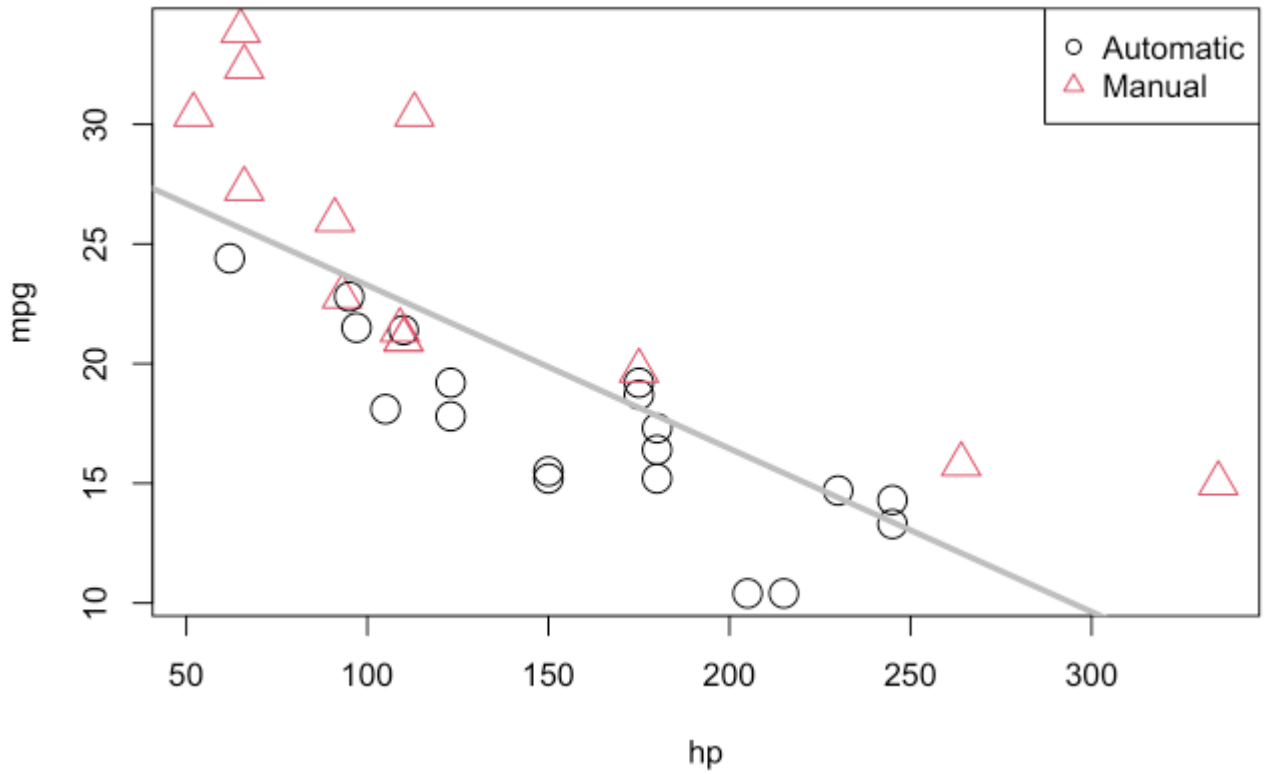
- El modelo de regression se representa en R como:

```
mpg_hp_slr = lm(mpg ~ hp, data = mtcars)
```

Variable categórica de dos niveles

- Podemos trazar una recta de ajuste, utilizando la regresión estimada:

```r
plot(mpg ~ hp, data = mtcars, col = am + 1, pch = am + 1, cex = 2)
abline(mpg_hp_slr, lwd = 3, col = "grey")
legend("topright", c("Automatic", "Manual"), col = c(1, 2), pch = c(1, 2))
```



Variable categórica de dos niveles

- Deberíamos notar un patrón: las observaciones manuales (rojas) caen en gran medida por encima de la línea, mientras que las observaciones automáticas (negras) están en su mayoría por debajo de la línea. Esto significa que nuestro modelo subestima la eficiencia de combustible de las transmisiones manuales y sobreestima la eficiencia de combustible de las transmisiones automáticas. Para corregir esto, agregaremos un predictor a nuestro modelo, es decir, un x2.

- Nuestro nuevo modelo sería: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$

- En este modelo tenemos que x1 y Y siguen igual, pero temenos para $x_2 = \begin{cases} 1\ transmisión\ manual \\ 0\ transmisión\ auto. \end{cases}$

-

- En este caso, llamamos a x2 una variable dummy. Desafortunadamente, una variable dummy recibe un nombre, ya que de ninguna manera es "tonta". De hecho, es algo (muy!) inteligente. Una variable ficticia es una variable numérica que se utiliza en un análisis de regresión para "codificar" una variable categórica binaria. Veamos cómo funciona esto.

Variable categórica de dos niveles

- Primero, tenga en cuenta que ***am*** ya es una variable dummy (o ficticia), ya que usa los valores 0 y 1 para representar transmisiones automáticas y manuales. A menudo, una variable como am almacenaría los valores de los caracteres auto y man y tendríamos que convertirlos a 0 y 1, o, como veremos más adelante, R se encargará de crear variables ficticias por nosotros.
- Entonces, para ajustar el modelo anterior, lo hacemos como cualquier otro modelo de regresión múltiple que hayamos visto antes.

```
mpg_hp_add = lm(mpg ~ hp + am, data = mtcars)
```

- Veamos la estimación del modelo:

```
##
## Call:
## lm(formula = mpg ~ hp + am, data = mtcars)
##
## Coefficients:
## (Intercept)           hp           am
##    26.58491     -0.05889      5.27709
```

Variable categórica de dos niveles

- Dado que x2 solo puede tomar valores 0 y 1, podemos escribir efectivamente dos modelos diferentes, uno para transmisiones manuales y otro para transmisiones automáticas.

- Para transmisiones automáticas, es decir x2 = 0, tenemos: $Y = \beta_0 + \beta_1 x_1 + \varepsilon$

- Luego, para las transmisiones manuales, es decir x2 = 1, tenemos: $Y = (\beta_0 + \beta_2) + \beta_1 x_1 + \varepsilon$

- Observe que estos modelos comparten la misma pendiente, β_1, pero tienen diferentes intersecciones, que difieren en β_2. Entonces, el cambio en mpg es el mismo para ambos modelos, pero en promedio el mpg difiere en β_2 entre los dos tipos de transmisión. Por una constante β_2 …

Variable categórica de dos niveles

- Ahora calcularemos la pendiente estimada y la intersección de estos dos modelos para poder agregarlos a una gráfica. Tenemos en cuenta que:

  - $\hat{\beta}_0$ = `coef(mpg_hp_add)[1]` = 26.5849137
  - $\hat{\beta}_1$ = `coef(mpg_hp_add)[2]` = -0.0588878
  - $\hat{\beta}_2$ = `coef(mpg_hp_add)[3]` = 5.2770853

- Luego, podemos combinarlos para calcular la pendiente y las intersecciones estimadas.

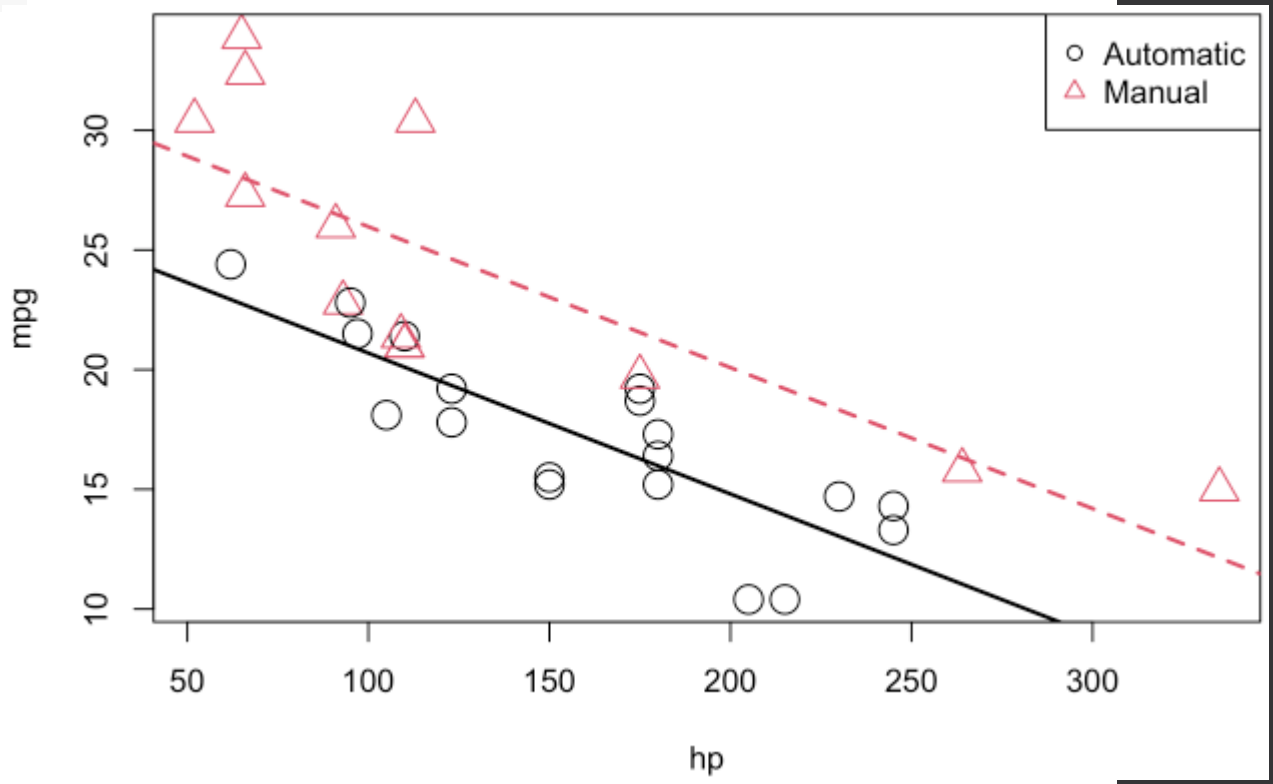
```
int_auto = coef(mpg_hp_add)[1]
int_manu = coef(mpg_hp_add)[1] + coef(mpg_hp_add)[3]


slope_auto = coef(mpg_hp_add)[2]
slope_manu = coef(mpg_hp_add)[2]
```

Variable categórica de dos niveles

- Al volver a graficar los datos, usamos estas pendientes e intersecciones para agregar los "dos" modelos ajustados a la gráfica.

```
plot(mpg ~ hp, data = mtcars, col = am + 1, pch = am + 1, cex = 2)
abline(int_auto, slope_auto, col = 1, lty = 1, lwd = 2) # add line for auto
abline(int_manu, slope_manu, col = 2, lty = 2, lwd = 2) # add line for manual
legend("topright", c("Automatic", "Manual"), col = c(1, 2), pch = c(1, 2))
```



Variable categórica de dos niveles

- Notamos de inmediato que los puntos ya no son sistemáticamente incorrectos. Las observaciones manuales rojas varían alrededor de la línea roja sin un patrón en particular sin subestimar las observaciones como antes. Los puntos negros automáticos varían alrededor de la línea negra, también sin un patrón obvio.

- Una imagen vale más que mil palabras, pero como analistas de información, a veces una imagen vale la pena un análisis completo. La imagen de arriba muestra claramente que β_2 es significativo, pero verifiquémoslo matemáticamente. Básicamente nos gustaría probar:

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

- Esto no es nada nuevo. Nuevamente, las matemáticas son las mismas que las de los análisis de regresión múltiple que hemos visto antes. Podríamos realizar una prueba t o F . La única diferencia es un ligero cambio de interpretación. Podríamos pensar en esto como probar un modelo con una sola línea (h0) contra un modelo que permite dos líneas (h1).

Variable categórica de dos niveles

- Para obtener el estadístico de prueba y el valor p para la prueba t, usaríamos en R lo siguiente:

```
summary(mpg_hp_add)$coefficients["am",]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 5.277085e+00 1.079541e+00 4.888270e+00 3.460318e-05
```

- Para hacer lo mismo con la prueba F, usaríamos

```
anova(mpg_hp_slr, mpg_hp_add)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ hp
## Model 2: mpg ~ hp + am
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 447.67
## 2     29 245.44  1    202.24 23.895 3.46e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tenga en cuenta que estos de hecho están probando lo mismo, ya que los valores p son exactamente iguales. (el estadístico de la prueba F es el estadístico de la prueba t al cuadrado).

Variable categórica de dos niveles

Recapitulando algunas interpretaciones:

- $\beta_0 = 26{,}58$ es el mpg promedio estimado para un automóvil con transmisión automática y 0 hp.
- $\beta_0 + \beta_2 = 31{,}86$ es el mpg promedio estimado para un automóvil con transmisión manual y 0 hp
- $\beta_2 = 5{,}27$, es la diferencia estimada en mpg promedio para autos con transmisión manual en comparación con aquellos con transmisión automática, para cualquier hp
- $\beta_1 = -0{,}05$, es el cambio estimado en mpg promedio para un aumento en un hp, para cualquiera de los tipos de transmisión.

Variable categórica de dos niveles

- Deberíamos prestar especial atención a esos dos últimos betas (β_2 y β_1). En el modelo:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

- vemos que β_1 es el cambio promedio en Y para un aumento en x1, sin importar el valor de x2. Además, β_2 es siempre la diferencia en el promedio de Y para cualquier valor de x1. Estas son dos restricciones que no siempre queremos, por lo que necesitamos una forma de especificar un modelo más flexible.
- Por ahora nos limitamos a un solo predictor numérico x1 y una variable ficticia x2. Sin embargo, el concepto de variable dummy (la cual realmente es inteligente…), se puede utilizar con modelos de regresión múltiple más grandes.
- Aquí solo usamos un único predictor numérico para facilitar la visualización, ya que podemos pensar en la interpretación de "dos líneas". Pero, en general, podemos pensar en una variable dummy como la creación de "dos modelos", uno para cada categoría de una variable categórica binaria.

Índice

La interacción

- Para eliminar la restricción de la "misma pendiente“ (lo cual realmente en pocos caso podría suceder), ahora discutiremos la **interacción**.
- Para ilustrar este concepto, volveremos al conjunto de datos autompg visto en temas anteriores:

```
# read data frame from the web
autompg = read.table(
  "http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data",
  quote = "\"",
  comment.char = "",
  stringsAsFactors = FALSE)
# give the dataframe headers
colnames(autompg) = c("mpg", "cyl", "disp", "hp", "wt", "acc", "year", "origin", "name")
# remove missing data, which is stored as "?"
autompg = subset(autompg, autompg$hp != "?")
# remove the plymouth reliant, as it causes some issues
autompg = subset(autompg, autompg$name != "plymouth reliant")
# give the dataset row names, based on the engine, year and name
rownames(autompg) = paste(autompg$cyl, "cylinder", autompg$year, autompg$name)
# remove the variable for name
autompg = subset(autompg, select = c("mpg", "cyl", "disp", "hp", "wt", "acc", "year", "origin"))
# change horsepower from character to numeric
autompg$hp = as.numeric(autompg$hp)
# create a dummy variable for foreign vs domestic cars. domestic = 1.
autompg$domestic = as.numeric(autompg$origin == 1)
# remove 3 and 5 cylinder cars (which are very rare.)
autompg = autompg[autompg$cyl != 5,]
autompg = autompg[autompg$cyl != 3,]
# the following line would verify the remaining cylinder possibilities are 4, 6, 8
#unique(autompg$cyl)
# change cyl to a factor variable
autompg$cyl = as.factor(autompg$cyl)
```

```
str(autompg)
```

```
## 'data.frame':    383 obs. of  9 variables:
##  $ mpg     : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cyl     : Factor w/ 3 levels "4","6","8": 3 3 3 3 3 3 3 3 3 3 ...
##  $ disp    : num  307 350 318 304 302 429 454 440 455 390 ...
##  $ hp      : num  130 165 150 150 140 198 220 215 225 190 ...
##  $ wt      : num  3504 3693 3436 3433 3449 ...
##  $ acc     : num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year    : int  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ domestic: num  1 1 1 1 1 1 1 1 1 1 ...
```

La interacción

- Eliminamos los automóviles de 3 y 5 cilindros y creamos una nueva variable nacional que indica si un automóvil se fabricó o no en los Estados Unidos. Quitar los cilindros de 3 y 5 es simplemente para facilitar la demostración más adelante en el capítulo y no se haría en la práctica. La nueva variable nacional toma el valor 1 si el automóvil se fabricó en los Estados Unidos y 0 en caso contrario, a lo que nos referiremos como "extranjero". (Estamos usando arbitrariamente a Estados Unidos como punto de referencia aquí). También hemos convertido el ciclo y el origen en variables factoriales, que discutiremos más adelante.
- Ahora nos ocuparemos de tres variables: mpg, disp y doméstico. Usaremos mpg como respuesta. Podemos estimar el siguiente modelo,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

En donde:

- Y: $es\ mpg, la\ eficiencia\ de\ combustible\ en\ millas\ por\ galón$
- x_1: es disp, el desplazamiento en pulgadas cúbicas
- x_2: es doméstico como se describe arriba, que es una variable dummy: $x_2 = \begin{cases} 1\ Domestic \\ 2\ Foreign \end{cases}$

La interacción

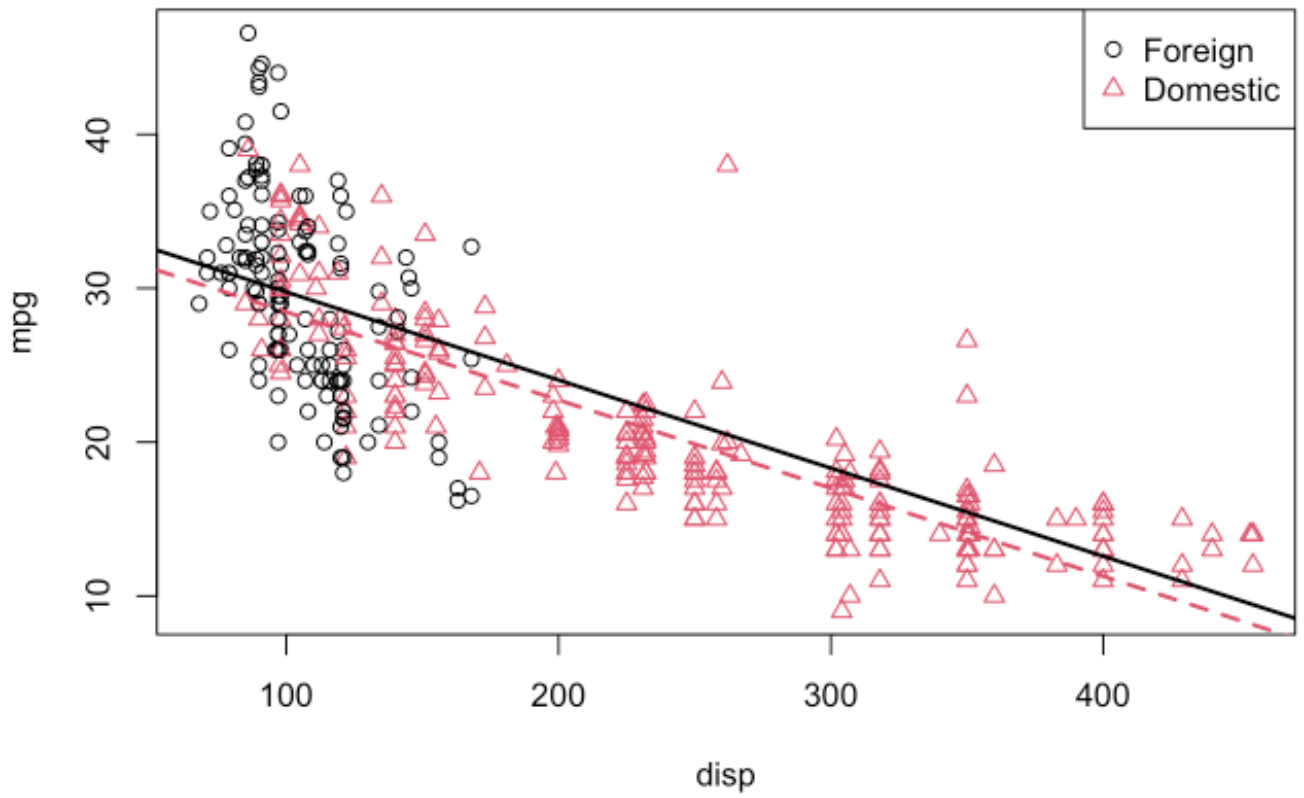
- Ajustaremos este modelo, extraeremos la pendiente y la intersección de las “dos líneas”, trazaremos los datos y sumaremos las líneas.

```r
mpg_disp_add = lm(mpg ~ disp + domestic, data = autompg)

int_for = coef(mpg_disp_add)[1]
int_dom = coef(mpg_disp_add)[1] + coef(mpg_disp_add)[3]

slope_for = coef(mpg_disp_add)[2]
slope_dom = coef(mpg_disp_add)[2]

plot(mpg ~ disp, data = autompg, col = domestic + 1, pch = domestic + 1)
abline(int_for, slope_for, col = 1, lty = 1, lwd = 2) # add line for foreign cars
abline(int_dom, slope_dom, col = 2, lty = 2, lwd = 2) # add line for domestic cars
legend("topright", c("Foreign", "Domestic"), pch = c(1, 2), col = c(1, 2))
```



La interacción

- Este es un modelo que permite dos líneas paralelas, lo que significa que el mpg puede ser diferente en promedio entre automóviles extranjeros y nacionales del mismo desplazamiento de motor, pero el cambio en el mpg promedio para un aumento en el desplazamiento es el mismo para ambos. Podemos ver que este modelo no está funcionando muy bien aquí. La línea roja se ajusta bastante bien a los puntos rojos, pero a la línea negra no le va muy bien con los puntos negros, claramente debería tener una pendiente más negativa. Básicamente, nos gustaría un modelo que permita dos pendientes diferentes.
- Considere el siguiente modelo,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

donde $x1$, $x2$ e Y son los mismos que antes, pero hemos agregado un nuevo término de interacción $x1x2$ que multiplica $x1$ y $x2$, por lo que también tenemos un parámetro β adicional β_3.

- Este modelo esencialmente crea dos pendientes y dos intersecciones, siendo β_2 la diferencia en las intersecciones y β_3 la diferencia en las pendientes. Para ver esto, desglosaremos el modelo en dos "submodelos" para automóviles nacionales y extranjeros.

La interacción

- Para automóviles extranjeros, es decir $x_2 = 0$, tenemos: $Y = \beta_0 + \beta_1 x_1 + \epsilon$.
- Para automóviles domésticos, es decir $x_2 = 1$, tenemos: $Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \epsilon$.

Estos dos modelos tienen pendientes e intersecciones diferentes.

- β_0 es el mpg promedio para un automóvil extranjero con 0 disp
- β_1 es el cambio en mpg promedio para un aumento de un disp, para automóviles extranjeros.
- $\beta_0 + \beta_2$ es el mpg promedio para un automóvil nacional con 0 disp
- $\beta_1 + \beta_3$ es el cambio en mpg promedio para un aumento de un disp, para automóviles nacionales.

- ¿Cómo estimamos este modelo en R? Hay diferentes maneras.

La interacción

- Un método sería simplemente crear una nueva variable y luego ajustar un modelo como cualquier otro.

```
autompg$x3 = autompg$disp * autompg$domestic # THIS CODE NOT RUN!
do_not_do_this = lm(mpg ~ disp + domestic + x3, data = autompg) # THIS CODE NOT RUN!
```

- Se prefiere no tener que modificar nuestros datos simplemente para ajustarlos a un modelo... Podemos decirle a R que nos gustaría usar los datos existentes con un término de interacción, que creará automáticamente cuando usemos el operador “**:**” .

```
mpg_disp_int = lm(mpg ~ disp + domestic + disp:domestic, data = autompg)
```

- Pero hay uno que lo hace todo ! Un método alternativo que se ajusta exactamente al mismo modelo anterior, sería utilizar el operador “*****”. Este método crea automáticamente el término de interacción, así como cualquier "término de orden inferior", que en este caso son los términos de primer orden para disp y domestic.

```
mpg_disp_int2 = lm(mpg ~ disp * domestic, data = autompg)
```

La interacción

- Verifiquemos los coeficientes de ambos métodos y el resumen del modelo estimado:

Método con "**:**"

```
coef(mpg_disp_int)
```

```
##   (Intercept)          disp      domestic disp:domestic
##    46.0548423    -0.1569239   -12.5754714     0.1025184
```

Método con "*****"

```
coef(mpg_disp_int2)
```

```
##   (Intercept)          disp      domestic disp:domestic
##    46.0548423    -0.1569239   -12.5754714     0.1025184
```

```
## Call:
## lm(formula = mpg ~ disp + domestic + disp:domestic, data = autompg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8332  -2.8956  -0.8332   2.2828  18.7749
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    46.05484    1.80582  25.504  < 2e-16 ***
## disp           -0.15692    0.01668  -9.407  < 2e-16 ***
## domestic      -12.57547    1.95644  -6.428 3.90e-10 ***
## disp:domestic   0.10252    0.01692   6.060 3.29e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.308 on 379 degrees of freedom
## Multiple R-squared:  0.7011, Adjusted R-squared:  0.6987
## F-statistic: 296.3 on 3 and 379 DF,  p-value: < 2.2e-16
```

→ ¿Qué podemos decir?

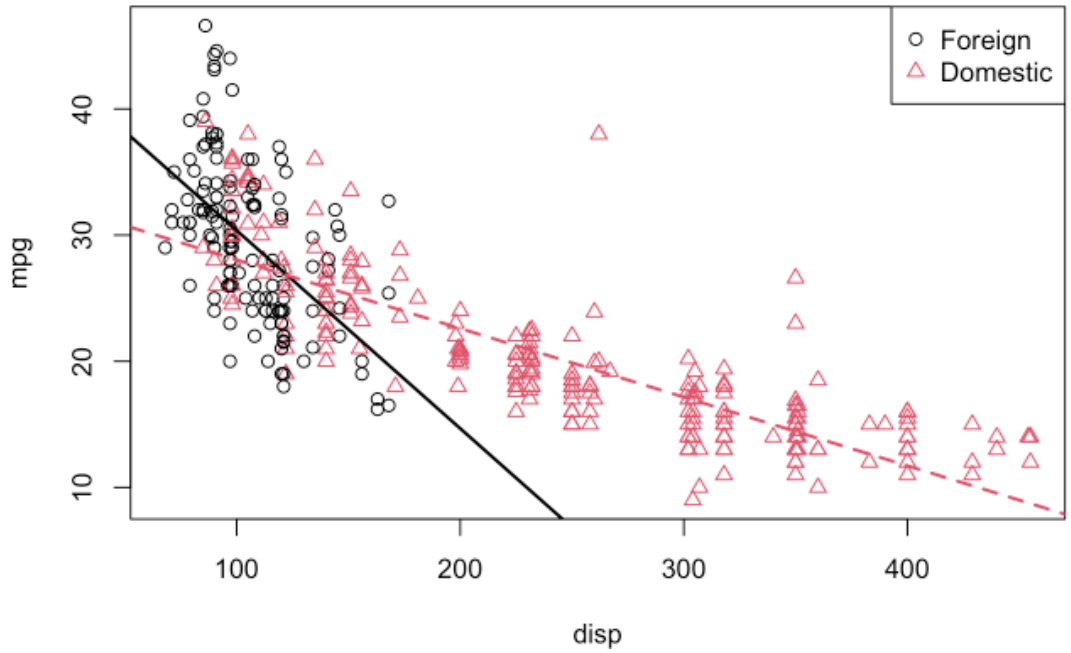
La interacción

- Veamos el efecto de introducir la interacción en la nueva regresión. Vemos que hay una diferencia entre los autos domésticos y los extranjeros. Sin la interacción, se suponía "paralelidad". Y esto no era correcto.

```
int_for = coef(mpg_disp_int)[1]
int_dom = coef(mpg_disp_int)[1] + coef(mpg_disp_int)[3]

slope_for = coef(mpg_disp_int)[2]
slope_dom = coef(mpg_disp_int)[2] + coef(mpg_disp_int)[4]
```

```
plot(mpg ~ disp, data = autompg, col = domestic + 1, pch = domestic + 1)
abline(int_for, slope_for, col = 1, lty = 1, lwd = 2) # line for foreign cars
abline(int_dom, slope_dom, col = 2, lty = 2, lwd = 2) # line for domestic cars
legend("topright", c("Foreign", "Domestic"), pch = c(1, 2), col = c(1, 2))
```



Vemos que estas líneas se ajustan mucho mejor a los datos, lo que coincide con el resultado de nuestras pruebas.

La interacción

- ¿Qué sucede con variables de tipo nominales (H Hombre y M mujer) tipo factores o variables ordinales (M Mucho A algo P Poco)?

- ¿Podemos realizar interacciones entre variables cuantitativas?

Índice

Tres o más categorías en la variable cuali.

- Consideremos ahora una variable factorial con más de dos niveles. En este conjunto de datos, cyl es un ejemplo.

- Aquí la variable cyl tiene tres niveles posibles: 4, 6 y 8. Quizás se pregunte, ¿por qué no usar simplemente cyl como variable numérica? Ciertamente se podría… pero es una variable ordinal…

- Sin embargo, eso obligaría a que la diferencia en mpg promedio entre 4 y 6 cilindros sea la misma que la diferencia en mpg promedio entre 6 y 8 cilindros. Eso suele tener sentido para una variable continua, pero no para una variable discreta con tan pocos valores posibles. En el caso de esta variable, no existe un motor de 7 cilindros o un motor de 6.23 cilindros en los vehículos personales. Por estas razones, simplemente consideraremos que cyl es categórico. Esta es una decisión que normalmente deberá tomarse con variables ordinales. A menudo, con un gran número de categorías, la decisión de tratarlas como variables numéricas es apropiada porque, de lo contrario, se necesita una gran cantidad de variables dummy para representar estas variables.

Tres o más categorías en la variable cuali.

- Definamos tres variables dummy relacionadas con la variable del factor cyl.

$$v_1 = \begin{cases} 1 & \text{4 cylinder} \\ 0 & \text{not 4 cylinder} \end{cases}$$

$$v_2 = \begin{cases} 1 & \text{6 cylinder} \\ 0 & \text{not 6 cylinder} \end{cases}$$

$$v_3 = \begin{cases} 1 & \text{8 cylinder} \\ 0 & \text{not 8 cylinder} \end{cases}$$

- Ahora, ajustemos un modelo aditivo en R, usando mpg como respuesta y disp y cyl como predictores. Este debe ser un modelo que use "tres líneas de regresión" para modelar mpg, una para cada uno de los posibles niveles de cilindros. Todos tendrán la misma pendiente (ya que es un modelo aditivo), pero cada uno tendrá su propia intersección.

```
(mpg_disp_add_cyl = lm(mpg ~ disp + cyl, data = autompg))
```

→

```
## Call:
## lm(formula = mpg ~ disp + cyl, data = autompg)
##
## Coefficients:
## (Intercept)         disp         cyl6         cyl8
##    34.99929     -0.05217     -3.63325     -2.03603
```

Tres o más categorías en la variable cuali.

- La pregunta es, ¿cuál es el modelo que R ha estimado aquí? Veamos el caso de un modelo estima sin interacción:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 v_2 + \beta_3 v_3 + \epsilon.$$

tenemos que

- Y es mpg, la eficiencia de combustible en millas por galón,
- X es disp, el desplazamiento en pulgadas cúbicas,
- v_2 y v_3 son las variables dummy definidas arriba

Por qué R no utiliza un v1? Básicamente porque no es necesario... Para crear tres líneas, solo necesita dos variables ficticias ya que está usando un nivel de referencia, que en este caso es un automóvil de 4 cilindros. Los tres "submodelos" son entonces:

- 4 Cylinder: $Y = \beta_0 + \beta_1 x_1 + \epsilon$
- 6 Cylinder: $Y = (\beta_0 + \beta_2) + \beta_1 x_1 + \epsilon$
- 8 Cylinder: $Y = (\beta_0 + \beta_3) + + \beta_1 x_1 + \epsilon$

Tres o más categorías en la variable cuali.

Observe que todas ecuaciones anteriores tienen la misma pendiente. Sin embargo, usando las dos variables dummy, logramos las tres intersecciones necesarias.

- β_0 es el mpg promedio para un automóvil de 4 cilindros con 0 disp.
- $\beta_0 + \beta_2$ is the average mpg for a 6 cylinder car with 0 disp.
- $\beta_0 + \beta_3$ is the average mpg for a 8 cylinder car with 0 disp.

- Entonces, debido a que 4 cilindros es el nivel de referencia, β_0 es específico de 4 cilindros, pero β_2 y β_3 se utilizan para representar cantidades relativas a 4 cilindros.

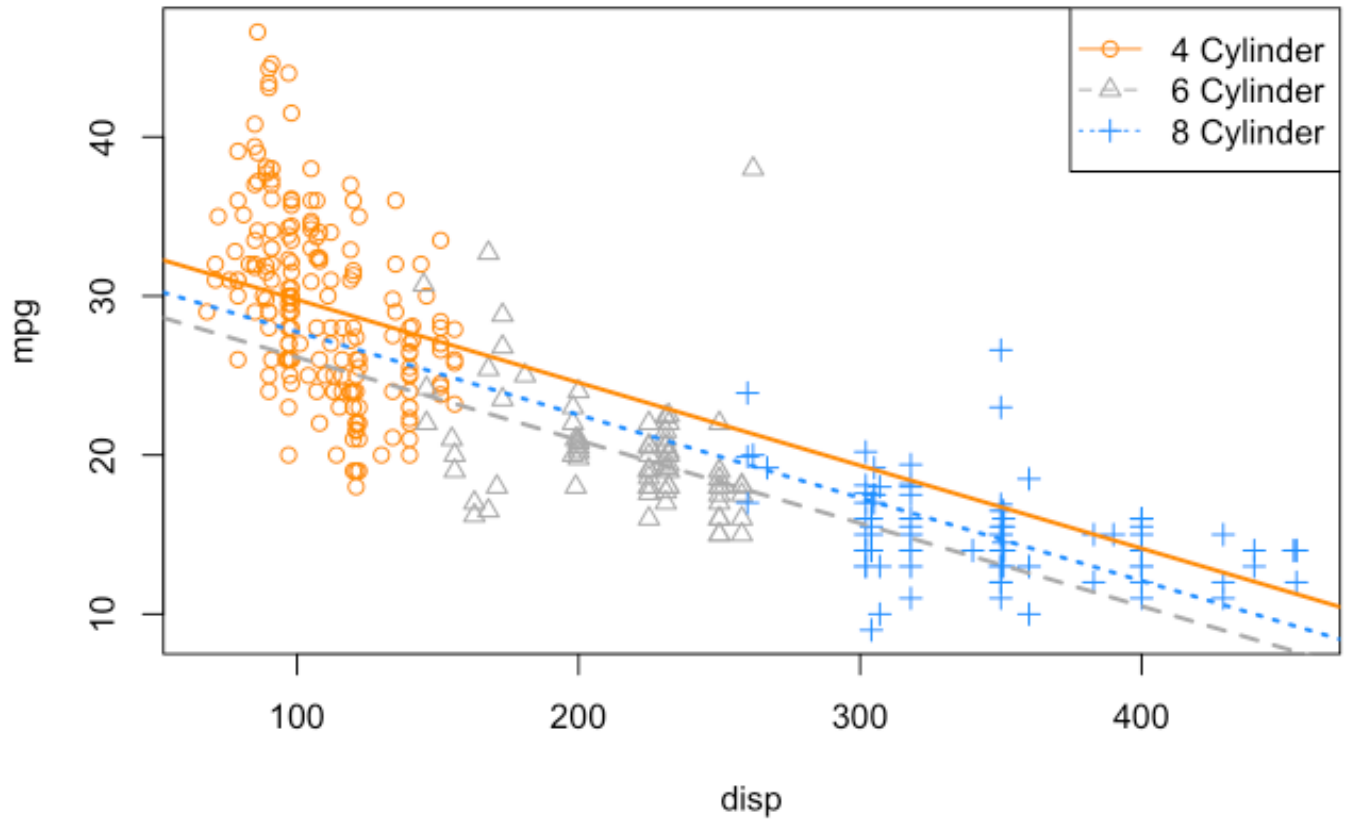
- Como hemos hecho antes, podemos extraer estas intersecciones y pendientes para las tres líneas y trazarlas en consecuencia (siguiente diapositiva).

Tres o más categorías en la variable cuali.

```
int_4cyl = coef(mpg_disp_add_cyl)[1]
int_6cyl = coef(mpg_disp_add_cyl)[1] + coef(mpg_disp_add_cyl)[3]
int_8cyl = coef(mpg_disp_add_cyl)[1] + coef(mpg_disp_add_cyl)[4]

slope_all_cyl = coef(mpg_disp_add_cyl)[2]

plot_colors = c("Darkorange", "Darkgrey", "Dodgerblue")
plot(mpg ~ disp, data = autompg, col = plot_colors[cyl], pch = as.numeric(cyl))
abline(int_4cyl, slope_all_cyl, col = plot_colors[1], lty = 1, lwd = 2)
abline(int_6cyl, slope_all_cyl, col = plot_colors[2], lty = 2, lwd = 2)
abline(int_8cyl, slope_all_cyl, col = plot_colors[3], lty = 3, lwd = 2)
legend("topright", c("4 Cylinder", "6 Cylinder", "8 Cylinder"),
       col = plot_colors, lty = c(1, 2, 3), pch = c(1, 2, 3))
```



Tres o más categorías en la variable cuali.

- ¡El resultado extraño aquí es que estamos estimando que los autos de 8 cilindros tienen una mejor eficiencia de combustible que los autos de 6 cilindros en cualquier desplazamiento! La línea azul punteada siempre está por encima de la línea gris punteada. Eso no parece correcto. Tal vez para motores de cilindrada muy grande eso podría ser cierto, pero parece incorrecto para motores de cilindrada media a baja… por supuesto, tenemos que tener en cuenta la interacción…
- Para intentar arreglar esto, intentaremos usar un modelo de interacción, es decir, en lugar de simplemente tres intersecciones y una pendiente, permitiremos tres pendientes. Veamos el modelo estimado:

```
(mpg_disp_int_cyl = lm(mpg ~ disp * cyl, data = autompg))
```

```
## Call:
## lm(formula = mpg ~ disp * cyl, data = autompg)
##
## Coefficients:
## (Intercept)         disp         cyl6         cyl8    disp:cyl6    disp:cyl8
##    43.59052     -0.13069    -13.20026    -20.85706      0.08299      0.10817
```

Tres o más categorías en la variable cuali.

- Se ha optado por utilizar 4 cilindros como nivel de referencia, pero esto también tiene ahora un efecto sobre los términos de interacción. Se ajustó el modelo.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 v_2 + \beta_3 v_3 + \gamma_2 x v_2 + \gamma_3 x v_3 + \epsilon.$$

- Estamos usando γ como un parámetro ß para simplificar, de modo que, por ejemplo, ß2 y γ2 están asociados con v2

Ahora lo que tenemos es son “3 sub modelos”:

- 4 Cylinder: $Y = \beta_0 + \beta_1 x + \epsilon.$
- 6 Cylinder: $Y = (\beta_0 + \beta_2) + (\beta_1 + \gamma_2) x + \epsilon.$
- 8 Cylinder: $Y = (\beta_0 + \beta_3) + (\beta_1 + \gamma_3) x + \epsilon.$

- ¿Cómo probamos si debemos aplicar o no interacciones?

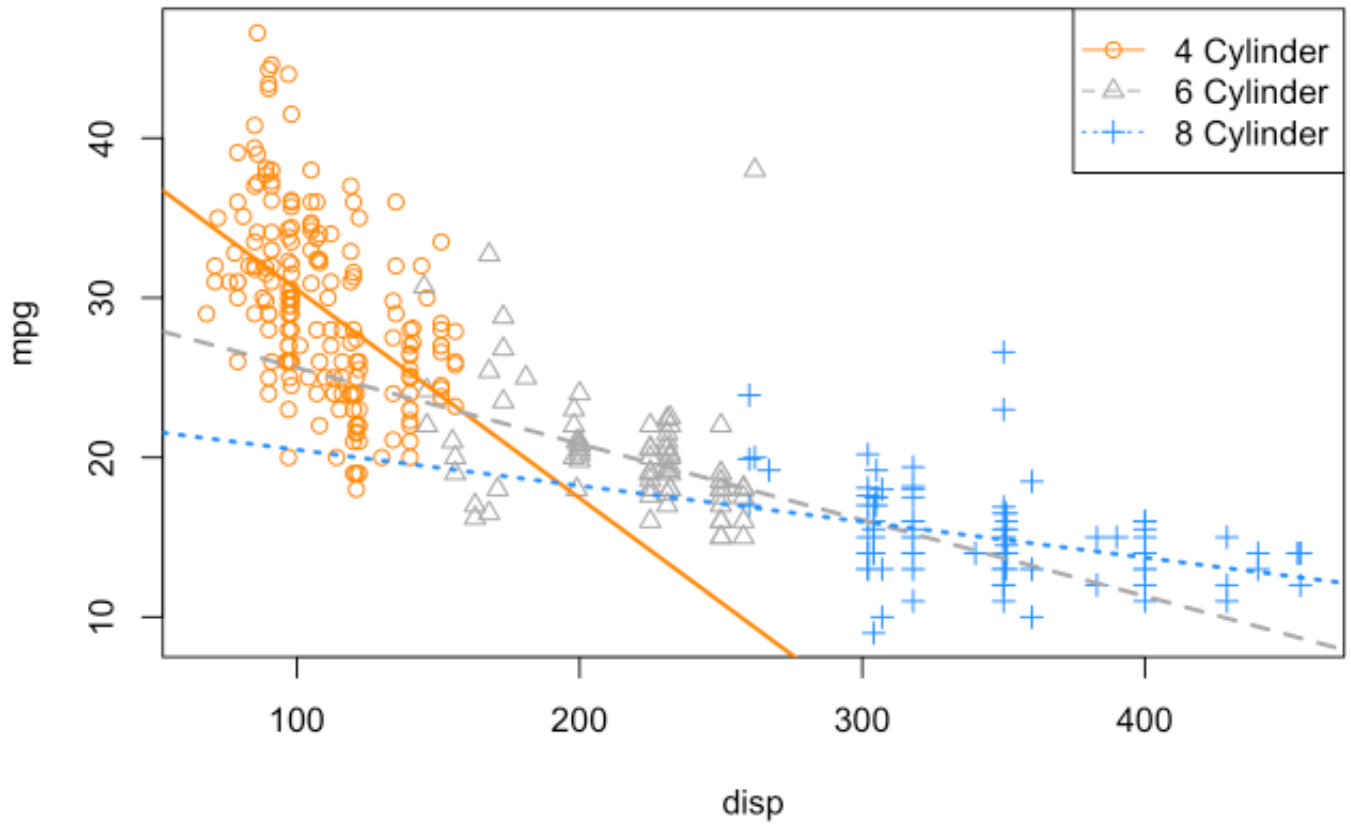
- Veamos los modelos con la interacción de estos tres sub modelos.

Tres o más categorías en la variable cuali.

```r
int_4cyl = coef(mpg_disp_int_cyl)[1]
int_6cyl = coef(mpg_disp_int_cyl)[1] + coef(mpg_disp_int_cyl)[3]
int_8cyl = coef(mpg_disp_int_cyl)[1] + coef(mpg_disp_int_cyl)[4]

slope_4cyl = coef(mpg_disp_int_cyl)[2]
slope_6cyl = coef(mpg_disp_int_cyl)[2] + coef(mpg_disp_int_cyl)[5]
slope_8cyl = coef(mpg_disp_int_cyl)[2] + coef(mpg_disp_int_cyl)[6]

plot_colors = c("Darkorange", "Darkgrey", "Dodgerblue")
plot(mpg ~ disp, data = autompg, col = plot_colors[cyl], pch = as.numeric(cyl))
abline(int_4cyl, slope_4cyl, col = plot_colors[1], lty = 1, lwd = 2)
abline(int_6cyl, slope_6cyl, col = plot_colors[2], lty = 2, lwd = 2)
abline(int_8cyl, slope_8cyl, col = plot_colors[3], lty = 3, lwd = 2)
legend("topright", c("4 Cylinder", "6 Cylinder", "8 Cylinder"),
       col = plot_colors, lty = c(1, 2, 3), pch = c(1, 2, 3))
```



Tres o más categorías en la variable cuali.

- ¡Esto se ve mucho mejor!

- Podemos ver que para los coches de cilindrada media, los coches de 6 cilindros ahora funcionan mejor que los de 8 cilindros, lo que parece mucho más razonable que antes.

- Para justificar completamente el modelo de interacción (es decir, una pendiente única para cada nivel de cilindro) en comparación con el modelo aditivo (pendiente única), podemos realizar una prueba F. En primer lugar, observe que no existe una prueba t que pueda hacer esto, ya que la diferencia entre los dos modelos no es un solo parámetro… (teoría estadística …).

- Tenemos que probar lo siguiente: $H_0 : \gamma_2 = \gamma_3 = 0$

que representa las líneas de regresión paralelas que vimos antes,

$$Y = \beta_0 + \beta_1 x + \beta_2 v_2 + \beta_3 v_3 + \epsilon.$$

Tres o más categorías en la variable cuali.

- Veamos la prueba en R:

```r
anova(mpg_disp_add_cyl, mpg_disp_int_cyl)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ disp + cyl
## Model 2: mpg ~ disp * cyl
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    379 7299.5
## 2    377 6551.7  2    747.79 21.515 1.419e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Y vemos que efectivamente, la interacción si tiene una razón de ser

Índice

Otras extensiones

- Ahora que hemos visto cómo incorporar predictores categóricos, así como términos de interacción, podemos comenzar a construir modelos mucho más grandes y flexibles que potencialmente pueden ajustarse mejor a los datos.

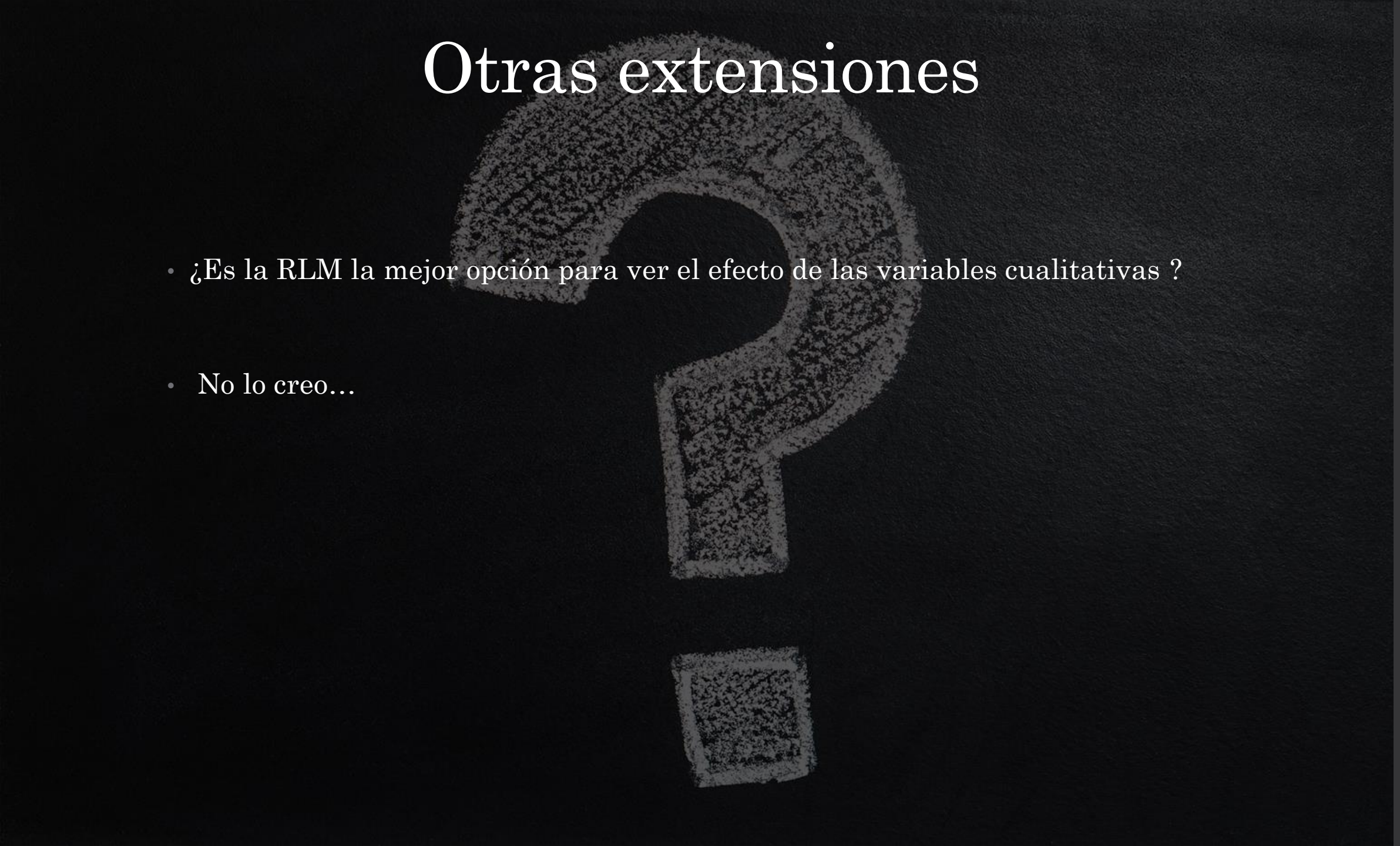
- Definamos un modelo "grande“,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \beta_7 x_1 x_2 x_3 + \epsilon.$$

- El procedimiento es similar…

- Podemos probar de igual forma la interacción para 2 o más variables cuantitativas a la vez…

Otras extensiones

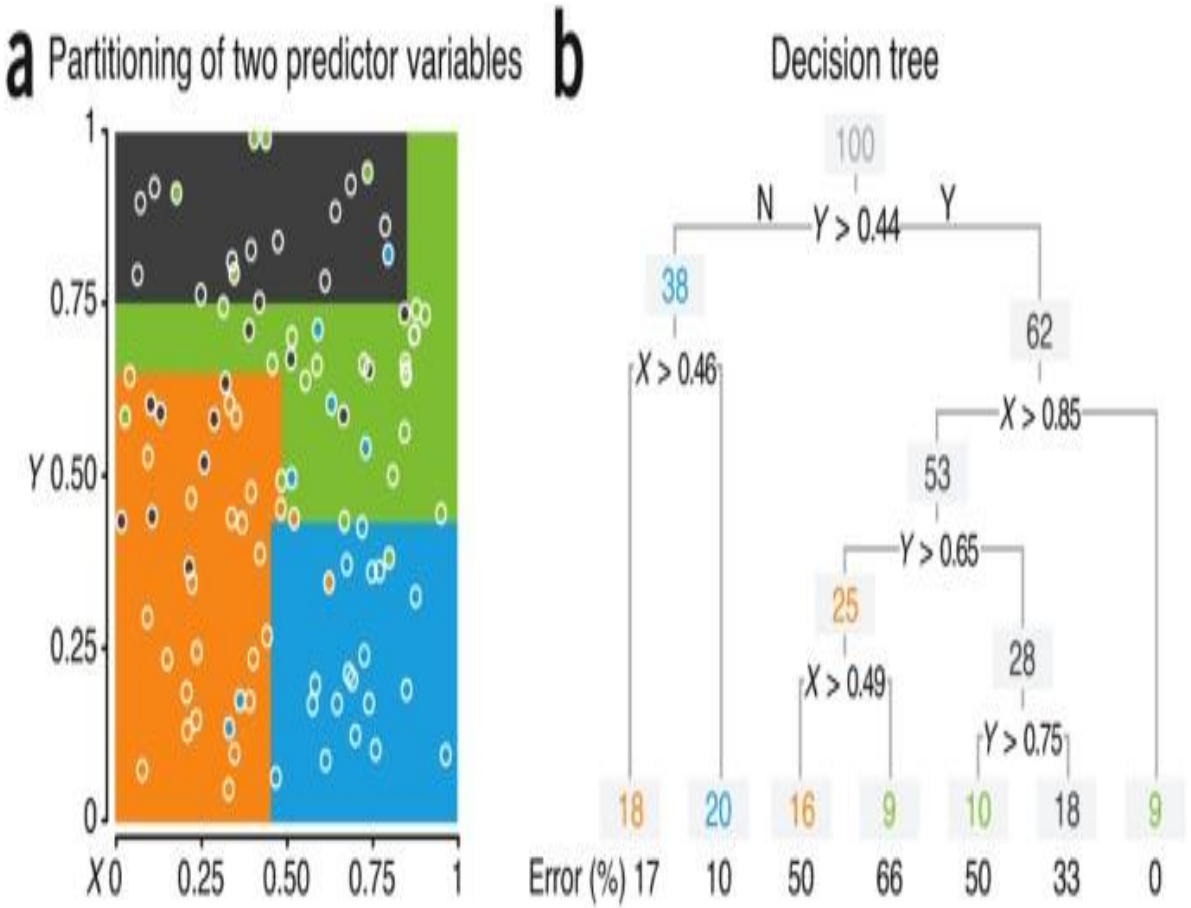


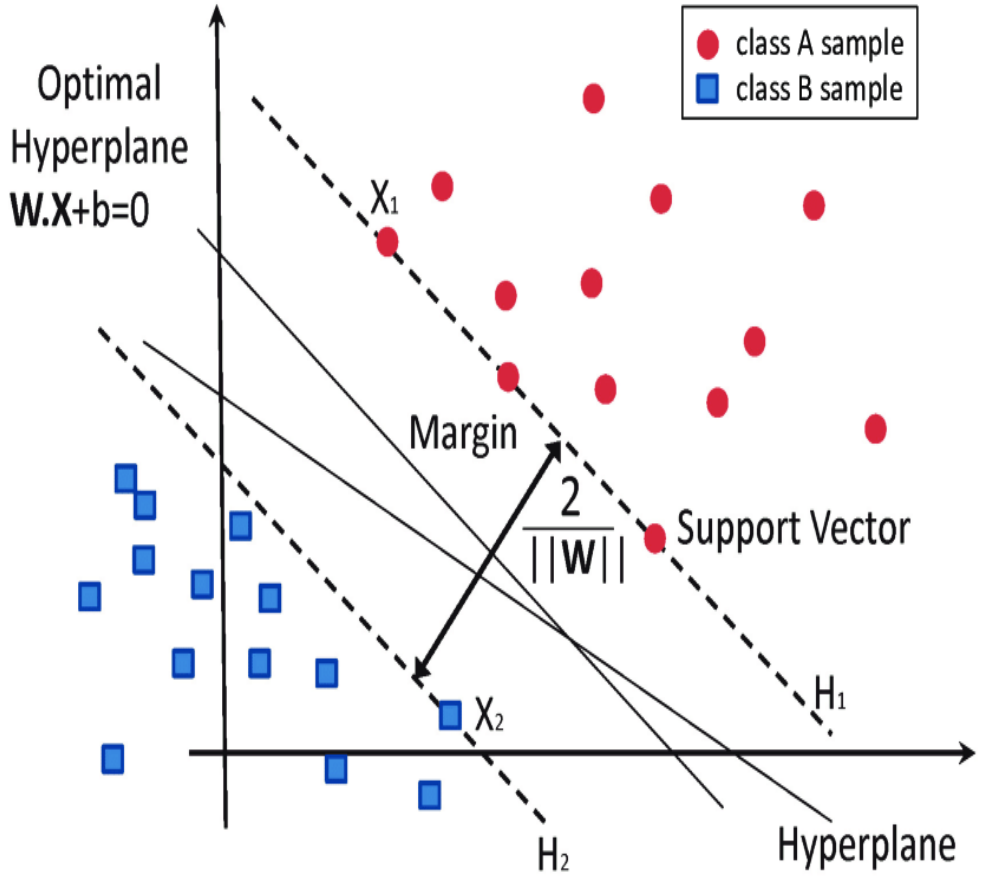
- ¿Es la RLM la mejor opción para ver el efecto de las variables cualitativas ?

- No lo creo…

Otras extensiones

- Al trabajar con variables cualitativas, prefiero otras opciones como los árboles de regresión o los SVM… me parece que extraen y brindan mejores resultados.





Conclusión

El presente capítulo estudió el incorporar variables cualitativas a la matriz X en la regresión.

Se estudio el caso de una variable dummy, luego el caso de 2 o más variables, y el uso de las interacciones.

¿Podemos aplicar modelos de regresión pero para variables dependientes cualitativas?

Claro que si, más adelante veremos los modelos dicotómicos de Logit y Probit.


The End