Professor: Héctor Bahamonde, PhD.

e:hibano@utu.fi

w:www.HectorBahamonde.com

Curso: OLS.

TA: Valterri Pulkkinen.

Interaction Terms: An Introduction

Traditional linear models—like the one in Equation 1—shows the effect of a variable x_1 (*schooling*) over y, keeping the control variable x_2 (*man*) constant at its mean.

$$\text{income}_i = \beta_0 + \beta_1 \text{schooling}_i + \beta_2 \text{man}_i + \epsilon_i \tag{1}$$

By the way, Why is it important to even control for gender, like in Equation 1?

An interaction term, however, is used when we want to know the **combined** effect of two (or more) independent variables. The advantage is that interaction terms show the effect on y of the two variables at the same time (x_1 **and** x_2).

For instance, if we wanted to know what's the combined effect of *schooling* (x_1) **and** (that is, in **combination with**) *man* (x_2) over *income* (y_i), we should estimate the following equation:

$$\text{income}_i = \beta_0 + \beta_1 \text{schooling}_i + \beta_2 \text{man}_i + \beta_3 \text{schooling}_i \times \text{man}_i + \epsilon_i \tag{2}$$

Substantive From a substantive point of view, interaction terms are relevant because they shed light on interactive questions. Like in this case, there are very compelling reasons to believe that—unfortunately—if we had two subjects, one male and one female with similarly low levels of schooling, man might be more likely to be better off compared to the female individual. This is particularly in developing contexts. Hence, it makes sense from a substantive point of view to study how our results vary by gender.

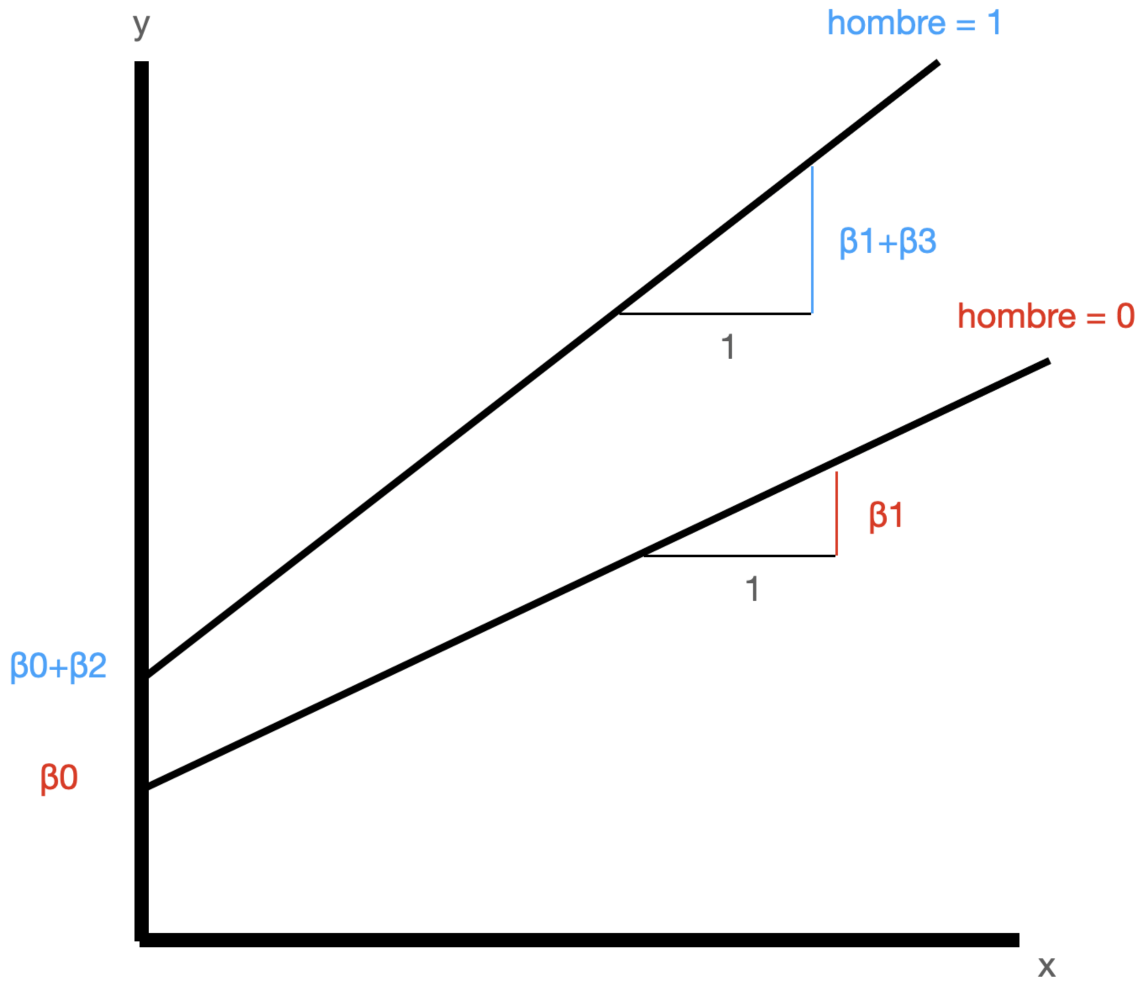
Parametrization The way in which we specify a interactive model is very similar to traditional linear models, but with important differences. Let's see this closely by taking a look at Equation 2 again:

1. Parameters β_0 and β_1 are the intercept and the slope, respectively of the "reference category." In this case, the reference category is the "base category", or mathematically, when man=0, that is, for "women." This is very intuitive. When the variable man=0, the model is reduced to Equation 3.

2. The intercept is for the other group—man=1, that is, for the man in the dataset. In these case, it should be $\beta_0 + \beta_2$, mientras que la pendiente está dada por $\beta_1 + \beta_3$. De la misma manera, es muy intuitivo. Ve la Equation 4

$$\begin{aligned}
\text{income}_i &= \beta_0 + \beta_1\text{schooling}_i + \beta_2\text{man}_i + \beta_3\text{schooling}_i \times \text{man}_i + \epsilon_i \\
\text{income}_i &= \beta_0 + \beta_1\text{schooling}_i + \beta_2 0 + \beta_3\text{schooling}_i \times 0_i + \epsilon_i \\
\text{income}_i &= \beta_0 + \beta_1\text{schooling}_i + 0 + 0 + \epsilon_i \\
\text{income}_i &= \beta_0 + \beta_1\text{schooling}_i + \epsilon_i
\end{aligned} \tag{3}$$

$$\begin{aligned}
\text{income}_i &= \beta_0 + \beta_1\text{schooling}_i + \beta_2\text{man}_i + \beta_3\text{schooling}_i \times \text{man}_i + \epsilon_i \\
\text{income}_i &= \beta_0 + \beta_1\text{schooling}_i + \beta_2 1 + \beta_3\text{schooling}_i \times 1_i + \epsilon_i \\
\text{income}_i &= \beta_0 + \beta_1\text{schooling}_i + \beta_2 + \beta_3\text{schooling}_i + \epsilon_i \\
\text{income}_i &= (\beta_0 + \beta_2) + \text{schooling}_i \times (\beta_1 + \beta_3) + \epsilon_i
\end{aligned} \tag{4}$$



Buenas Prácticas

Fíjate que cada vez que incluimos un término de interacción ($\text{schooling}_i \times \text{man}_i$), para interpretar su parámetro asociado (β_3), es necesario incluir los sub-términos por separado. Esto es, permitir que la ecuación tenga un parámetro independiente asociado a *schooling* y *man*, esto es, β_1 y β_2 (tal y como aparece en Equation 2). Si estimamos sólo la siguiente ecuación, β_3 estará sesgado. **NO HAGAS LO SIGUIENTE**:

$$\text{income}_i = \beta_0 + \beta_3 \text{man}_i \times \text{schooling}_i + \epsilon_i \tag{5}$$

Estimación en R

De acuerdo a Brambor, Clark, and Golder (2006, p. 73),[1] el efecto marginal en la ecuación Equation 6,

$$\text{income}_i = \beta_0 + \beta_1 \text{schooling}_i + \beta_2 \text{man}_i + \epsilon_i \tag{6}$$

está dado por el siguiente cálculo:

$$\frac{\partial y}{\partial x_1} = \beta_1 \tag{7}$$

Sin embargo el término de interacción en Equation 2 es distinto, y está dado por el siguiente cálculo:

$$\frac{\partial y}{\partial x_1} = \beta_1 + \beta_3 \text{man} \tag{8}$$

En palabras, es *cuánto cambia y cuando cambia x, según niveles de la variable hombre.*

Cambiemos de ejemplo, y estimemos un modelo con tres niveles (no dos, como *man*).

Carguemos los datos:

```
p_load(effects)
data(Duncan)
summary(Duncan)

##    type        income        education         prestige
##  bc  :21   Min.   : 7.00   Min.   :  7.00   Min.   : 3.00
##  prof:18   1st Qu.:21.00   1st Qu.: 26.00   1st Qu.:16.00
##  wc  : 6   Median :42.00   Median : 45.00   Median :41.00
##            Mean   :41.87   Mean   : 52.56   Mean   :47.69
##            3rd Qu.:64.00   3rd Qu.: 84.00   3rd Qu.:81.00
##            Max.   :81.00   Max.   :100.00   Max.   :97.00
```

Estimemos el modelo. Nota que hemos puesto la multiplicación, y `R` "sabe" que debe meter los términos constitutivos.

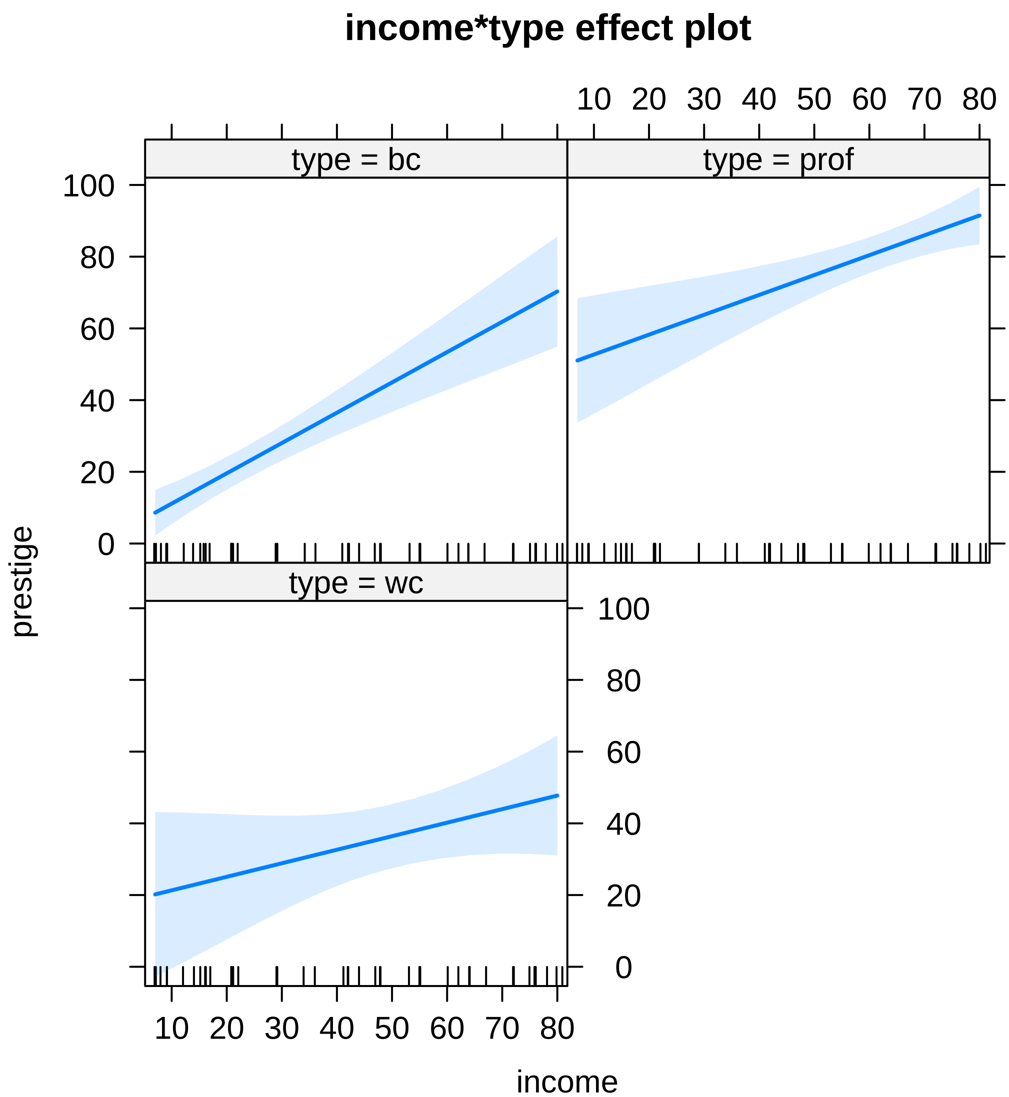
[1] Ecuaciones 11-13.

```r
modelo.1 = lm(prestige ~ income*type, data = Duncan)
summary(modelo.1)
```

```
##
## Call:
## lm(formula = prestige ~ income * type, data = Duncan)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.4405  -6.0480  -0.2787   4.7269  28.1950
##
## Coefficients:
##                 Estimate Std. Error t value     Pr(>|t|)
## (Intercept)       2.6828     3.8182   0.703     0.486450
## income            0.8450     0.1289   6.554 0.0000000882 ***
## typeprof         44.4868    10.3641   4.292     0.000113 ***
## typewc           14.8531    13.4986   1.100     0.277926
## income:typeprof  -0.2909     0.2017  -1.442     0.157148
## income:typewc    -0.4674     0.2736  -1.709     0.095466 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.44 on 39 degrees of freedom
## Multiple R-squared:  0.9026,Adjusted R-squared:  0.8902
## F-statistic: 72.31 on 5 and 39 DF,  p-value: < 0.00000000000000022
```

Como explican Brambor, Clark, and Golder (2006), las tablas de regresión no nos ayudan a interpretar los modelos interactivos. Debemos proceder interpretando como se señala en Equation 8. Afortunadamente existe la librería `effects`.

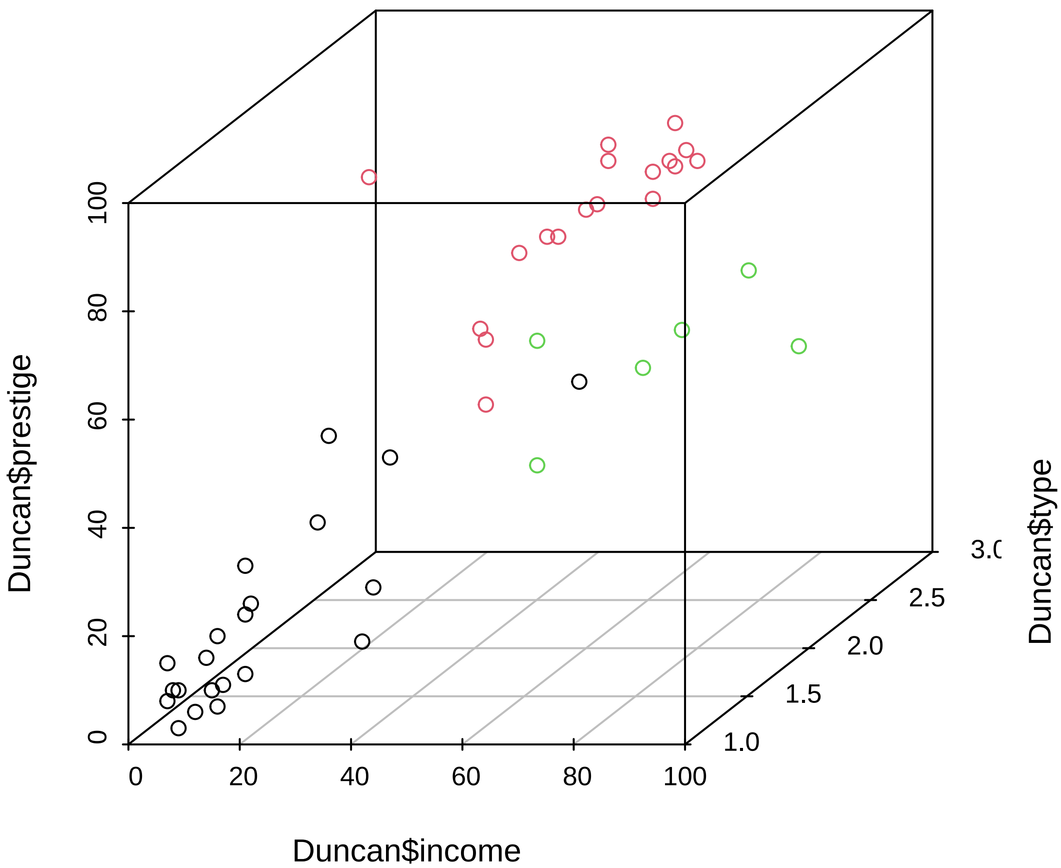
```
term.int <- effect("income*type", modelo.1)
plot(term.int, as.table=T)
```



Si te das cuenta, los efectos no son los mismos. Las derivadas (en Equation 8) no tienen por qué dar lo mismo. Es por esto que no debemos mirar la tabla de regresión. En un sentido espacial, un término de interacción es el análisis de tres planos. En la Equation 2: y, x_1, x_2.

Veamos de qué se trata:

```
p_load(scatterplot3d)
scatterplot3d(Duncan$income, Duncan$type, Duncan$prestige, color = as.numeric(Duncan$type))
```



Usemos una base de datos donde todas las variables son continuas (no como en el ejemplo donde *man* es dicotómica):

```r
p_load(car,rgl)
data(iris)
sep.l <- iris$Sepal.Length
sep.w <- iris$Sepal.Width
pet.l <- iris$Petal.Length
scatter3d(x = sep.l, y = pet.l, z = sep.w, groups = iris$Species)
```

Correr esto último en R.

```
## Error in parse_block(g[-1], g[1], params.src, markdown_mode):  Duplicate chunk label
'setup', which has been used for the chunk:
## if (!require("pacman")) install.packages("pacman"); library(pacman)
## p_load(knitr)
## set.seed(2020)
## options(scipen=9999999)
## if (!require("pacman")) install.packages("pacman"); library(pacman)

## Writing to file Clase_14.R
```