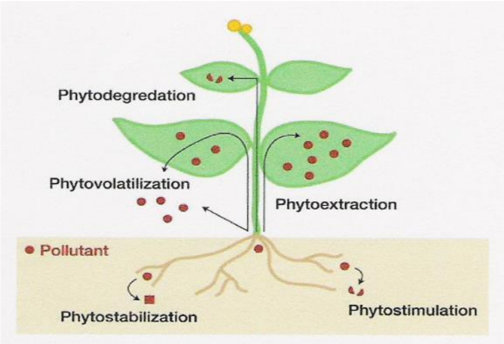
BIOMETRÍA II
CLASE 5
MODELOS CON MÁS DE UNA PREDICTORA

Adriana Pérez
Depto de Ecología, Genética y Evolución
FECN, UBA

Recuperación de suelos empetrolados de Comodoro Rivadavia
Luque, JL - Estación Experimental Chubut del INTA Trelew - 2009

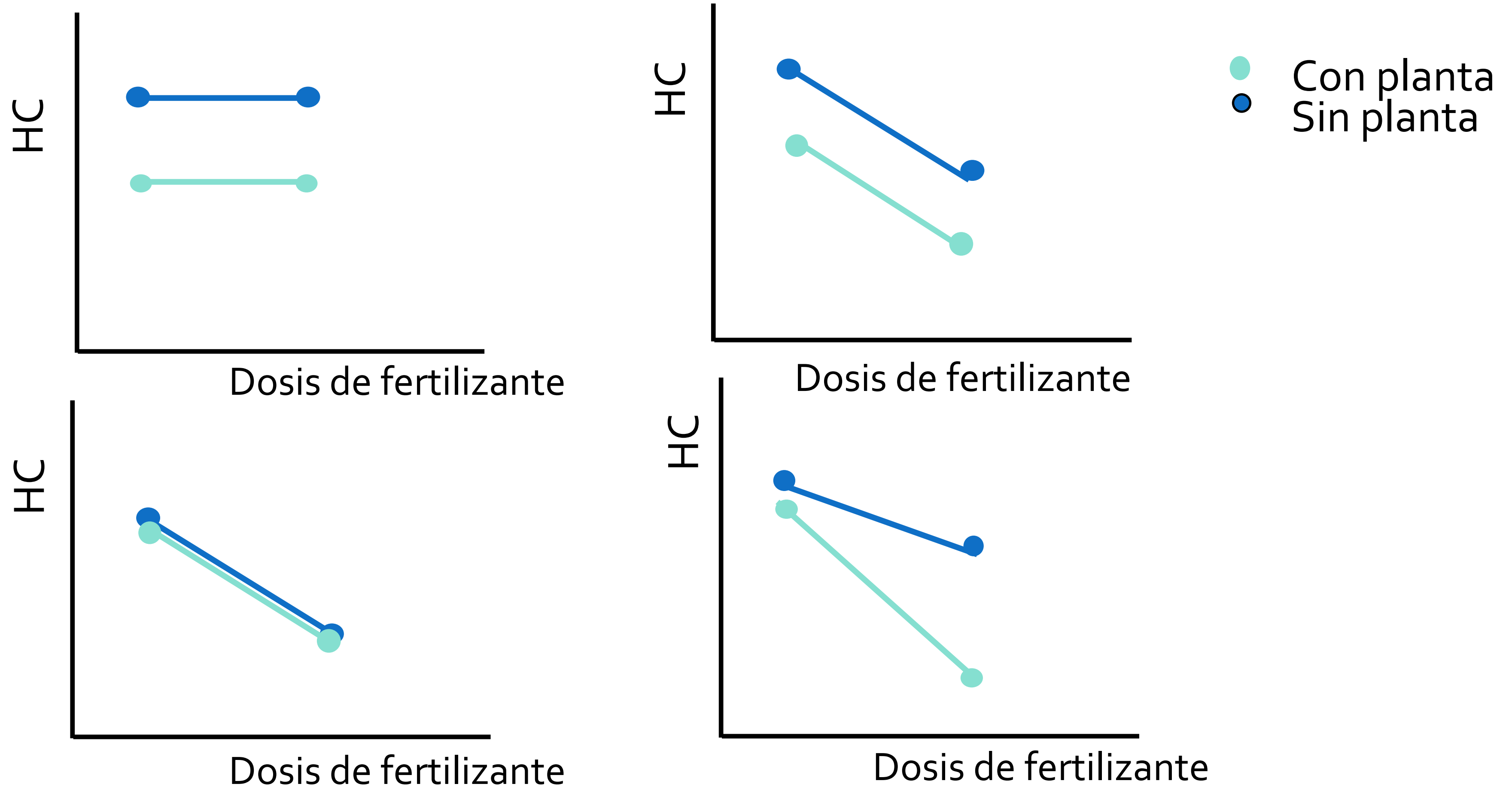
- La fitorremediación es una técnica que emplea vegetales en combinación con microorganismos asociados a la rizósfera para remover, degradar o inmovilizar contaminantes contenidos en suelos, sedimentos y aguas
- Se desea estudiar el desempeño de dos especies vegetales perennes: charcao (*Senecio filaginoides*), nativa, y agropiro alargado (*Thynopiron ponticum*), exótica, para fitorremediar suelos empetrolados de Comodoro Rivadavia.
- La bioestimulación es la adición de nutrientes al suelo para estimular la actividad de microorganismos degradadores del contaminante. Se desea estudiar el efecto de la adición al suelo de un fertilizante (fósforo + nitrógeno).

¿Podemos, en un mismo ensayo, estudiar la efectividad de las plantas y la de los microorganismos en remover los HC del suelo?



Recuperación de suelos empetrolados de C. Rivadavia
Luque, JL - Estación Experimental Chubut del INTA Trelew - 2009

- Se planea un ensayo utilizando macetas conteniendo suelo extraído de la zona petrolera de Cdro Rivadavia conteniendo 4.1% de hidrocarburos (HC) (4,1 g/100 g suelo seco)
- Se desea contestar las siguientes preguntas:
  - ¿Las especies son efectivas en remover los HC? (independientemente del agregado de fertilizante)
  - ¿Cuánto de la remoción de HC es por microorganismos del suelo y no por acción de las plantas?
  - ¿La fertilización es efectiva en promover la remoción de los HC? (independientemente del agregado de plantas)
  - La capacidad de remoción de HC de las especies ¿cambia según el agregado de fertilizante?



- ✓ Efectos nulos
- ✓ Efectos aditivos
- ✓ Interacción

Interacción entre variables explicatorias

- El efecto de una VE sobre la VR cambia según los valores que tome otra VP
- Es decir que el efecto de una VE depende de / se asocia con el valor que tome otra VE (y viceversa) (modificación de efectos)
- Si hay interacción entre VE, pierde relevancia estimar los efectos de una dada VE independientemente de los valores que tome la otra VE con la que interactúa (principio de marginalidad)
- Las interacciones pueden ser entre cualquier tipo de variables (categóricas con categóricas, cuantitativas con categóricas, cuanti con cuanti…)

Recuperación de suelos empetrolados de C. Rivadavia
Luque, JL - Estación Experimental Chubut del INTA Trelew - 2009

- 30 Macetas conteniendo suelo extraído de la zona petrolera de Cdro Rivadavia conteniendo 4.1% de hidrocarburos (HC) (4,1 g/100 g suelo seco)
- Cada maceta fue asignada al azar a una combinación de Planta (Charcao, Agropiro o Testigo sin vegetación) y Fertilización (con o sin)
- A los 350 días se midió contenido en suelo de HC totales de petróleo (% P/P, g/100 g suelo seco)

- Experimento o estudio observacional?
- UE:
- VR (tipo y potencial distribución de probabilidades):
- VE (tipo):
- Factores y niveles:
- Tratamientos:
- Diseño factorial de 3x2
- Replicación, aleatorización, control del error

Diseños factoriales

Son aquellos que incluyen más de una VE categórica (factor). Modelos de comparación de medias

- Mayor eficiencia en el uso de los recursos, menor error global y mayor potencia que varios unifactoriales
- Permiten evaluar la interacción entre factores

- Los gráficos de perfiles son muy útiles para describir el comportamiento de la VR
- Cuando un experimento tiene dos o más factores, la cantidad de medias que pueden compararse surge de las combinaciones de los niveles de los factores

El modelo es

Y~A*B

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$$

$i=1,2\ldots a$
$j=1,2..b$
$k=1,2\ldots n_{ij}$

- donde Y_{ijk} es la concentración de HC en suelo de cada UE
- μ es la media general o media de la población
- α_i es el efecto del factor i (especie)
- β_j es el efecto del factor j (fertilización)
- $\alpha\beta_{ij}$ es el efecto de la interacción ij
- ε_{ijk} es el error aleatorio de cada UE

$\varepsilon_{ijk} \sim$ NID $(0, \sigma^2)$

tienen que ser cualis

En R: `m1<-lm(HC~veg*fert, bd)`

Hipótesis en Diseño factorial

- Ho: $\gamma_{ij} = 0$

  es decir no existe interacción entre especies y fertilización ⇨ El efecto de la fertilización es independiente de la especie (y viceversa)

> Es la que debe analizarse primero, ya que si existe interacción no hay independencia entre los efectos de A y B y no es correcto analizar los factores por separado

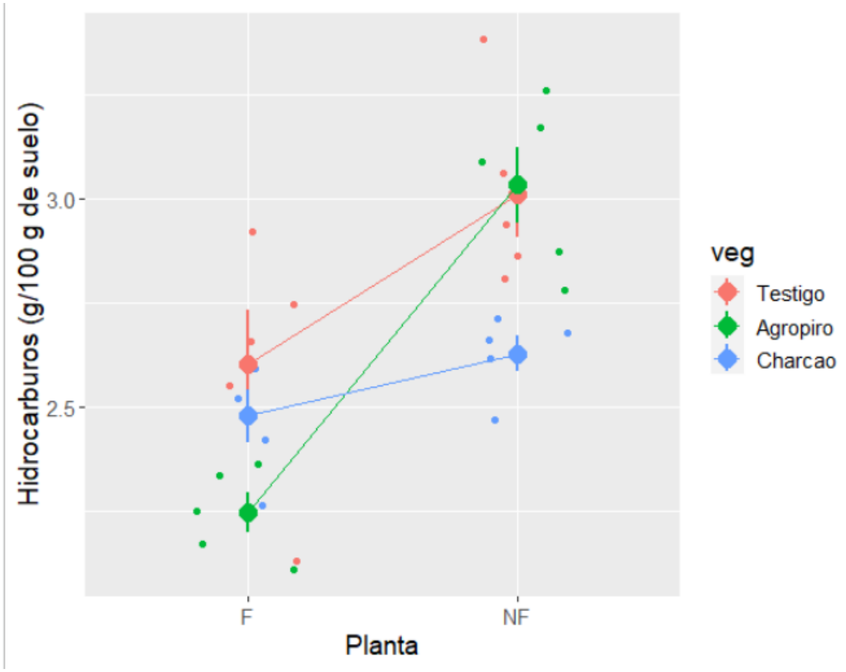
- Ho: $\alpha_i = 0$

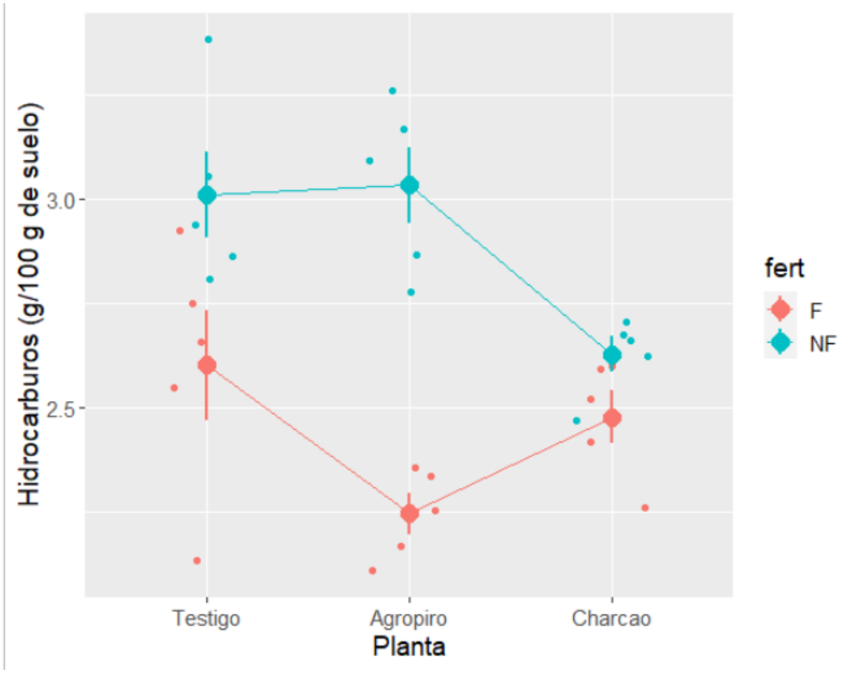
  es decir no existe efecto sobre el contenido de HC del suelo debido a la especie (suponiendo independencia entre A y B)

- Ho: $\beta_j = 0$

  es decir no existe efecto sobre el contenido de HC del suelo debido a la fertilización (suponiendo independencia entre A y B)

Gráficos de perfiles





Se pueden reordenar los niveles

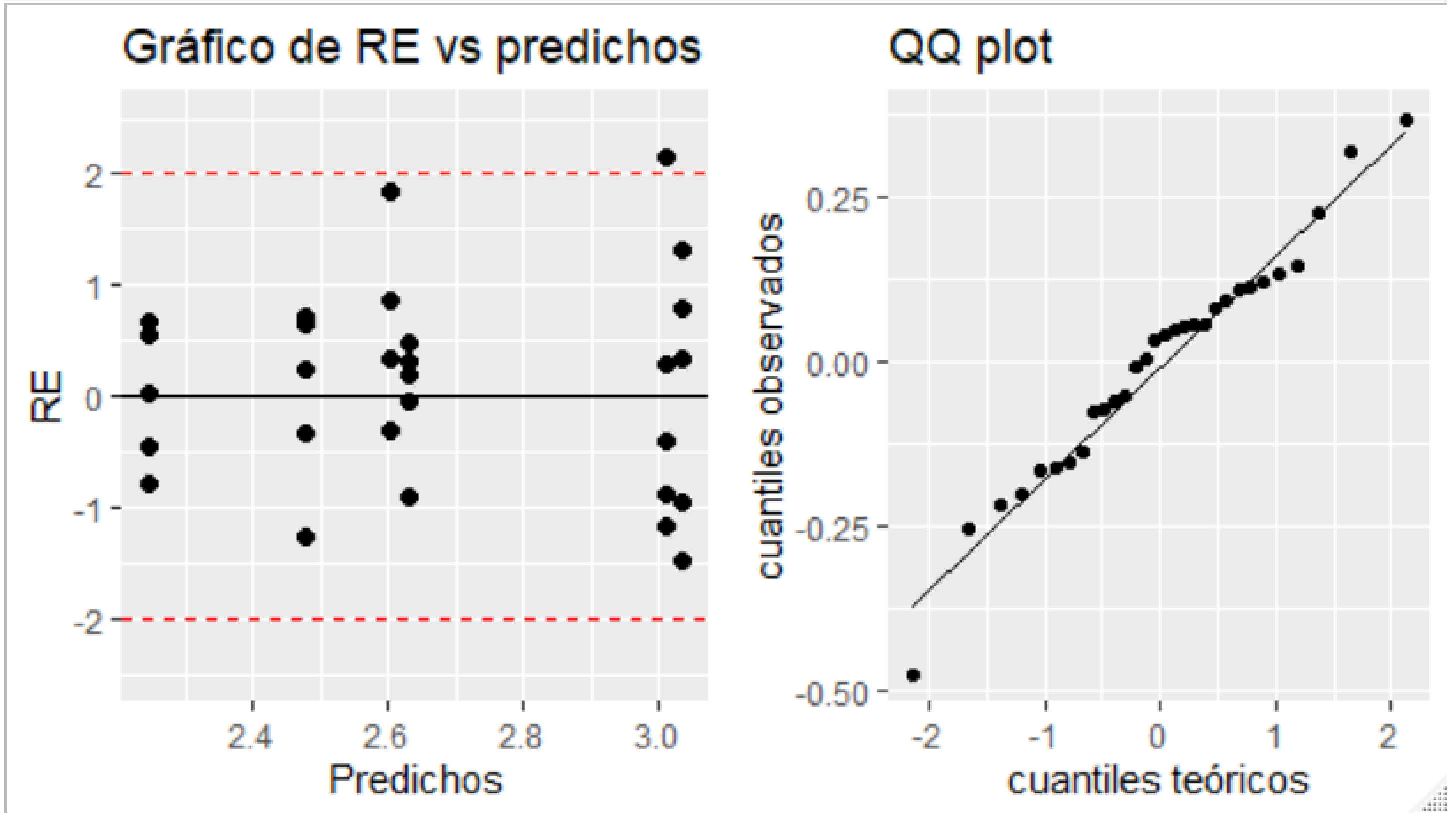
```
factor(bd$veg,levels=c("Testigo","Agro
piro", "Charcao"))
```

Calculando residuos

	Especie	Fertil	HC	Predichos	Residuos	residuos std
1	3	2	3.06	3.010	0.050	0.291
2	3	2	2.86	3.010	-0.150	-0.874
3	3	2	2.81	3.010	-0.200	-1.165
4	3	2	2.94	3.010	-0.070	-0.408
5	3	2	3.38	3.010	0.370	2.155
6	3	1	2.66	2.602	0.058	0.338
7	3	1	2.75	2.602	0.148	0.862
8	3	1	2.55	2.602	-0.052	-0.303
9	3	1	2.92	2.602	0.318	1.852
10	3	1	2.13	2.602	-0.472	-2.749
11	2	2	2.66	2.628	0.032	0.186
12	2	2	2.71	2.628	0.082	0.478
13	2	2	2.62	2.628	-0.008	-0.047
14	2	2	2.68	2.628	0.052	0.303
15	2	2	2.47	2.628	-0.158	-0.920
16	2	1	2.59	2.478	0.112	0.652
17	2	1	2.42	2.478	-0.058	-0.338
18	2	1	2.60	2.478	0.122	0.711
19	2	1	2.52	2.478	0.042	0.245
20	2	1	2.26	2.478	-0.218	-1.270

$$e_{ijk} = y_{ijk} - \bar{y}_{ij}$$

Estudiando los supuestos




```
> leveneTest(HC~veg*fert, FITOR, center=mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  5  1.4057 0.2578
      24
```

```
> shapiro.test(e)

        Shapiro-Wilk normality test

data:  e
W = 0.97461, p-value = 0.6713
```

✔

Tabla de ANOVA

FdV	SC	GL	CM	F
Entre niveles factor A	$\Sigma bn_{ij}(\bar{y}_{i.} - \bar{\bar{y}})^2$	$a\text{-}1$	$\frac{SC_A}{Gl_A}$	$\frac{CM_A}{CM_{error}}$
Entre niveles factor B	$\Sigma an_{ij}(\bar{y}_{.j} - \bar{\bar{y}})^2$	$b\text{-}1$	$\frac{SC_B}{Gl_B}$	$\frac{CM_B}{CM_{error}}$
AxB (interacción)	$\Sigma n_{ij}(\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{\bar{y}})^2$	$(a\text{-}1)(b\text{-}1)$	$\frac{SC_{AB}}{Gl_{AB}}$	$\frac{CM_{AB}}{CM_{error}}$
Error o dentro	$\Sigma(y_{ijk} - \bar{y}_{ij})^2$	$n\text{-}ab$	$\frac{SC_{error}}{GL_{error}}$	
Total	$\Sigma(y_{ijk} - \bar{\bar{y}})^2$	$n\text{-}1$		

Anova

```
m1<-lm(HC~veg*fert, bd)
anova(m1)
```

```
Analysis of Variance Table

Response: HC
          Df  Sum Sq Mean Sq F value    Pr(>F)    
veg        2 0.33045 0.16522  4.4833  0.022164 *  
fert       1 1.50976 1.50976 40.9668 1.283e-06 ***
veg:fert   2 0.51501 0.25750  6.9872  0.004061 ** 
Residuals 24 0.88448 0.03685                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

No es correcto analizar efectos principales

Pruebas globales

Efectos principales y simples

Efectos principales o **marginales** de un factor son las comparaciones entre los niveles de un factor promediados para todos los niveles del otro factor. Es decir, *independientemente* del otro factor.

Efectos simples o comparaciones de interacción (de celdas) son comparaciones entre distintos niveles de un factor fijando los niveles del otro factor.

Medias HC	Especie			$\bar{y}_{.j}$
Fertilización	Testigo	Charcao	Agropiro	
no	3,01	2,63	3,03	**2,890**
sí	2,60	2,48	2,25	**2,443**
$\bar{y}_{i.}$	**2,805**	**2,555**	**2,640**	**2,667**

Efectos simples (celdas; dentro de la tabla)

Efectos principales (marginales)

Principio de Marginalidad

- No deben interpretarse los efectos principales de las VE que interactúan
- No deben plantearse modelos con términos de interacción sin incluir los efectos principales

Comparaciones múltiples

- Los métodos disponibles y el procedimiento es el mismo que para anova de un factor
- Si la interacción fue no significativa se comparan los efectos principales y se efectúan comparaciones entre niveles de cada factor
- Si la interacción fue significativa no se comparan efectos principales, sino comparaciones entre celdas

Comparaciones de interacción

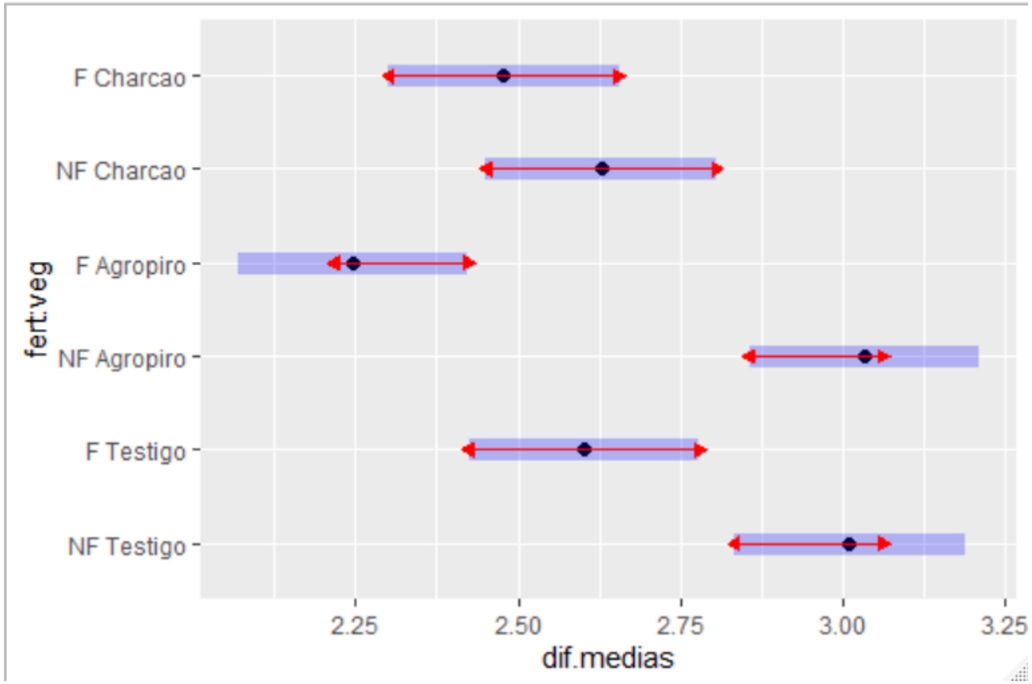
(entre todas las combinaciones)

IC para la diferencia de medias:
¿Cero pertenece al IC?
¿Cuál es la magnitud del efecto?

```
library(eemeans)
emmeans(m1, pairwise ~ fert*veg)
```

```
Confidence level used: 0.95

$contrasts
 contrast                   estimate    SE df t.ratio p.value  lower.CL upper.CL
 NF,Testigo - F,Testigo        0.408 0.121 24   3.360  0.0277    0.0326   0.7834
 NF,Testigo - NF,Agropiro     -0.024 0.121 24  -0.198  1.0000   -0.3994   0.3514
 NF,Testigo - F,Agropiro       0.764 0.121 24   6.293  <.0001    0.3886   1.1394
 NF,Testigo - NF,Charcao       0.382 0.121 24   3.146  0.0445    0.0066   0.7574
 NF,Testigo - F,Charcao        0.532 0.121 24   4.382  0.0024    0.1566   0.9074
 F,Testigo - NF,Agropiro      -0.432 0.121 24  -3.558  0.0176   -0.8074  -0.0566
 F,Testigo - F,Agropiro        0.356 0.121 24   2.932  0.0701   -0.0194   0.7314
 F,Testigo - NF,Charcao       -0.026 0.121 24  -0.214  0.9999   -0.4014   0.3494
 F,Testigo - F,Charcao         0.124 0.121 24   1.021  0.9062   -0.2514   0.4994
 NF,Agropiro - F,Agropiro      0.788 0.121 24   6.490  <.0001    0.4126   1.1634
 NF,Agropiro - NF,Charcao      0.406 0.121 24   3.344  0.0287    0.0306   0.7814
 NF,Agropiro - F,Charcao       0.556 0.121 24   4.579  0.0015    0.1806   0.9314
 F,Agropiro - NF,Charcao      -0.382 0.121 24  -3.146  0.0445   -0.7574  -0.0066
 F,Agropiro - F,Charcao       -0.232 0.121 24  -1.911  0.4200   -0.6074   0.1434
 NF,Charcao - F,Charcao        0.150 0.121 24   1.235  0.8153   -0.2254   0.5254

P value adjustment: tukey method for comparing a family of 6 esti
mates
```

```
emmeans(m1, pairwise ~ fert*veg)
```



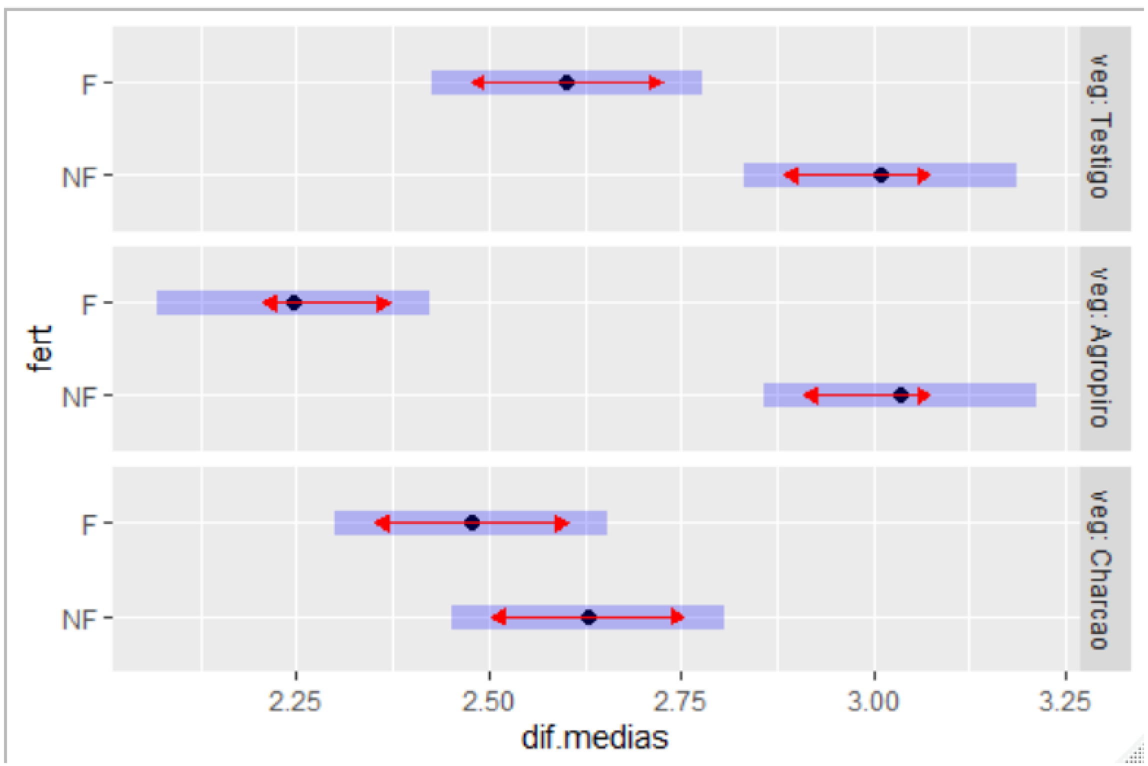
- Las barras grises son los IC para las medias
- las flechas rojas son para las comparaciones entre grupos. Si una flecha de un grupo se superpone a una flecha de otro grupo, la diferencia no es significativa
- Nota: Nunca usar IC para una media para realizar comparaciones, pueden ser muy engañosos

```
> CLD(comp1)
 fert veg      emmean         SE df lower.CL upper.CL .group
 F    Agropiro  2.246 0.08585259 24 2.068809 2.423191  1
 F    Charcao   2.478 0.08585259 24 2.300809 2.655191  12
 F    Testigo   2.602 0.08585259 24 2.424809 2.779191  12
 NF   Charcao   2.628 0.08585259 24 2.450809 2.805191   2
 NF   Testigo   3.010 0.08585259 24 2.832809 3.187191    3
 NF   Agropiro  3.034 0.08585259 24 2.856809 3.211191    3

Confidence level used: 0.95
P value adjustment: tukey method for comparing a family of 6 estimates
significance level used: alpha = 0.05
```

2- Otra posibilidad: efectos simples

```
emmeans(m1,  ~ fert | veg)
```

```
$contrasts
veg = Agropiro:
 contrast estimate        SE df t.ratio p.value
 F - NF     -0.788 0.1214139 24  -6.490  <.0001

veg = Charcao:
 contrast estimate        SE df t.ratio p.value
 F - NF     -0.150 0.1214139 24  -1.235  0.2286

veg = Testigo:
 contrast estimate        SE df t.ratio p.value
 F - NF     -0.408 0.1214139 24  -3.360  0.0026
```



- Mayor potencia, ya que son menos comparaciones
- La elección depende de los objetivos del ensayo

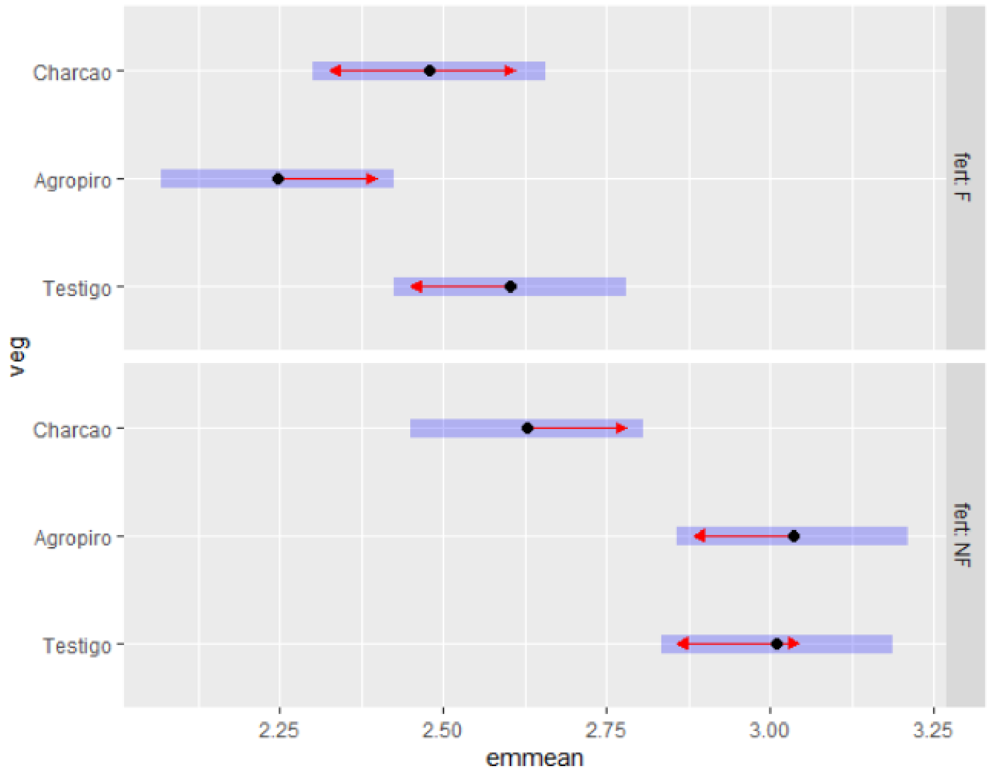
✓ Para cada tipo de planta ¿es efectiva la fertilización?

✓ O podría interesar: Para cada nivel de fertilización ¿cuál es la especie más efectiva?

2- efectos simples en el otro sentido

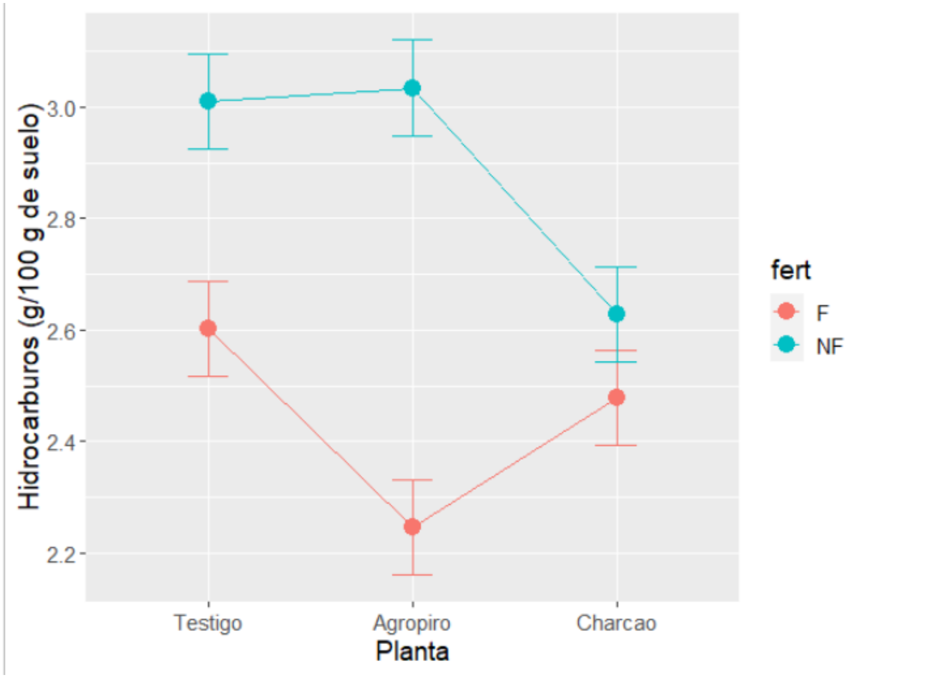
`emmeans(m1,  ~ veg | fert)`

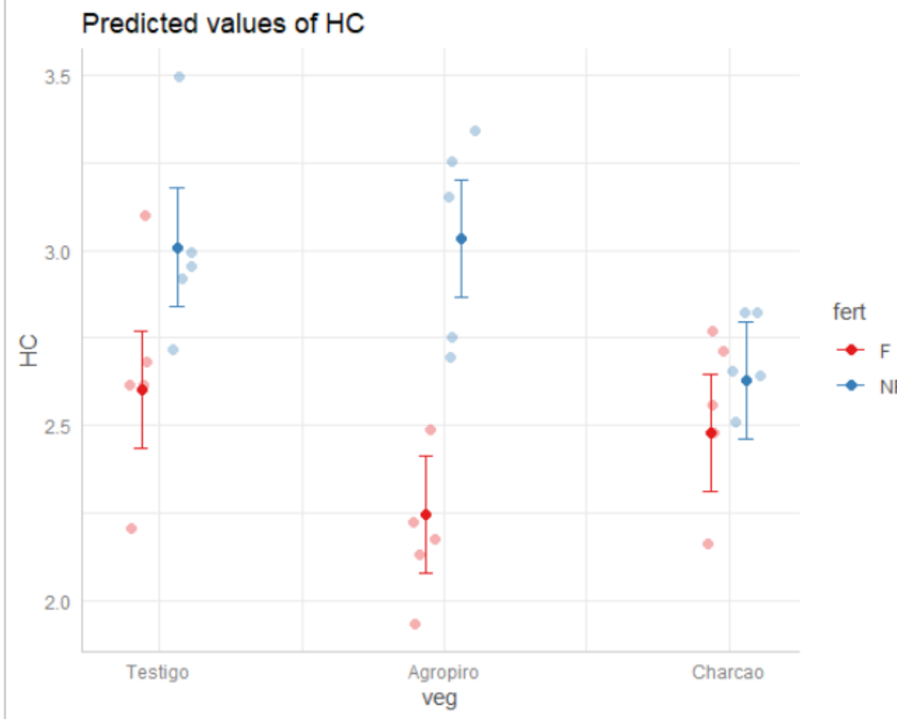
```
$contrasts
fert = F:
 contrast            estimate    SE df t.ratio p.value
 Testigo - Agropiro     0.356 0.121 24   2.932  0.0192
 Testigo - Charcao      0.124 0.121 24   1.021  0.5709
 Agropiro - Charcao    -0.232 0.121 24  -1.911  0.1574

fert = NF:
 contrast            estimate    SE df t.ratio p.value
 Testigo - Agropiro    -0.024 0.121 24  -0.198  0.9787
 Testigo - Charcao      0.382 0.121 24   3.146  0.0117
 Agropiro - Charcao     0.406 0.121 24   3.344  0.0073
```

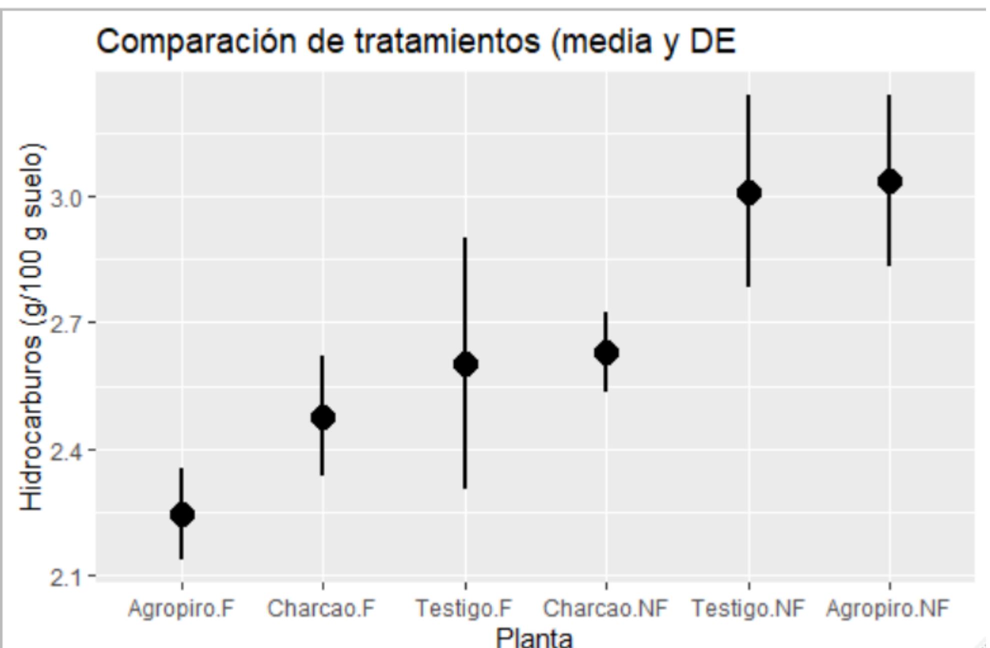


- Mayor potencia, ya que son menos comparaciones
- La elección depende de los objetivos del ensayo

- ✓ Para cada nivel de fertilización ¿cuál es la especie más efectiva?
- ✓ O uno u otro, pero no ambos







¿Cuál es la pregunta a responder?

- ✓ Para cada nivel de vegetación ¿es efectiva la fertilización?
- ✓ Para cada nivel de fertilización ¿cuál es la especie más efectiva?
- ✓ ¿Cuál es el mejor de estos tratamientos para restaurar el suelo?
- ✓ ¿Cuál es la magnitud de los efectos?

```
> CLD(comp1)
 fert veg       .group
 F    Agropiro  1
 F    Charcao   12
 F    Testigo   12
 NF   Charcao    2
 NF   Testigo     3
 NF   Agropiro    3
```

Test de Tukey
si la interacción no es significativa

Efectos principales:

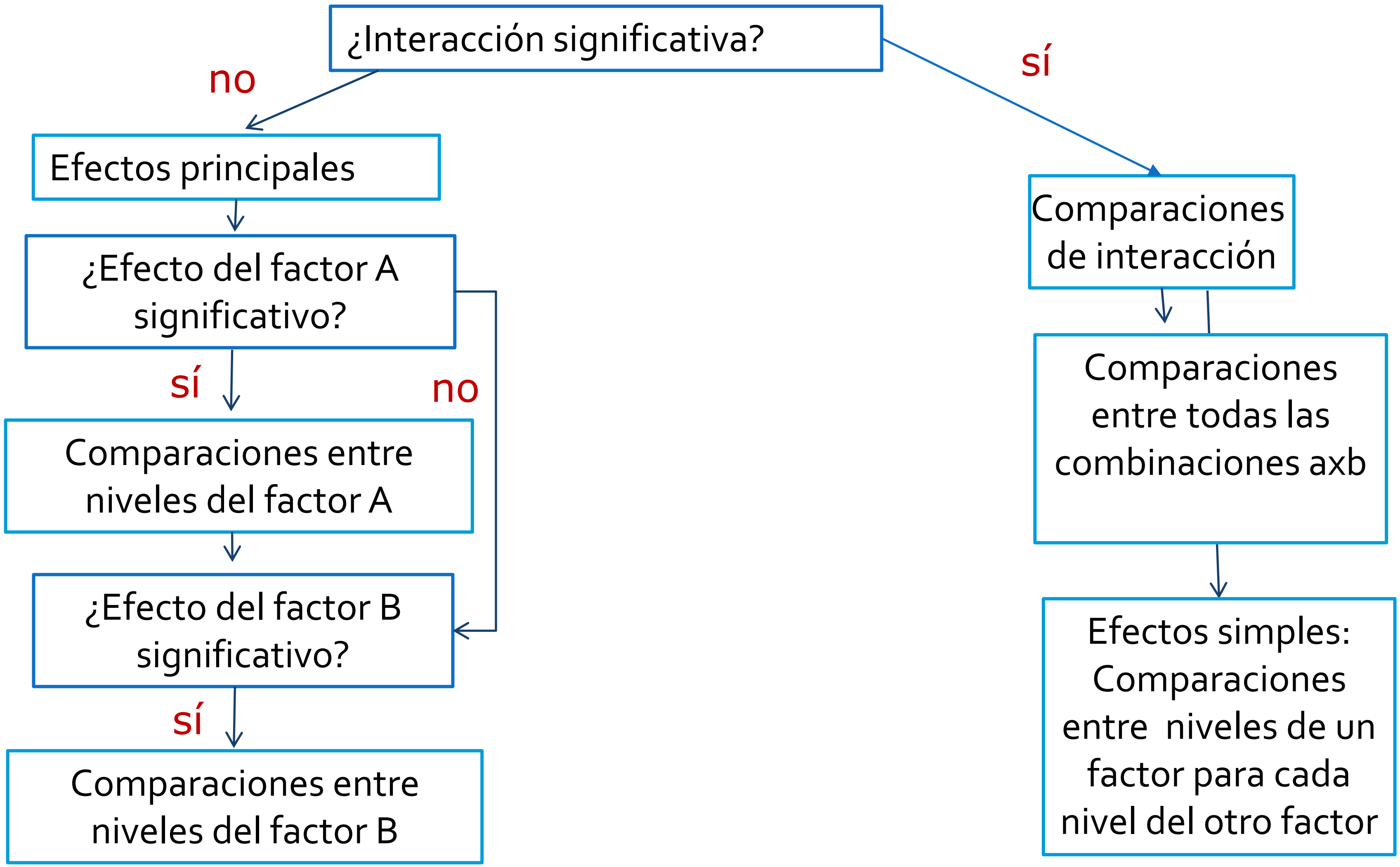
Comparaciones entre niveles del factor fert

```
emmeans(modelo2, pairwise ~ fert)
```

Comparaciones entre niveles del factor veg

```
emmeans(modelo2, pairwise ~ veg)
```

Análisis en diseños factoriales



Diseños más complejos

- Si se tienen 3 VE (A, B, C) hay una interacción triple y 3 dobles, además de los efectos principales (A, B, C, A*B, A*C, B*C, A*B*C)
- El modelo debe respetar el principio de marginalidad: si se incluye una interacción, se deben incluir los efectos principales de las VE que la constituyen

Fuente de variación	Escenario 1	Escenario 2	Escenario 3	Escenario 4
A				
B				
C				
A*B		S	S	NS
A*C		S	NS	NS
B*C		S	NS	NS
A*B*C	S	NS	NS	NS

Ej de efectos simples en escenario 1 y 2:

```
emmeans(m1, pairwise~ A*B|C, adjust= "tukey")
```

Selección de modelos con más de una v.explicatoria

- Experimentos diseñados:
  - Las VE son en general cualitativas (factores)
  - el modelo viene dado por el diseño experimental, no se lo debería simplificar
  - VE generalmente ortogonales

- Estudios observacionales:
  - VE cuanti y/o cualitativas (regresión múltiple)
  - es necesario simplificar el modelo => métodos de selección de modelos
  - VE casi nunca ortogonales

> VE ortogonales: La variabilidad explicada por un factor es la misma, independientemente de si el otro factor es tenido en cuenta o no. Verdadera partición de la variabilidad

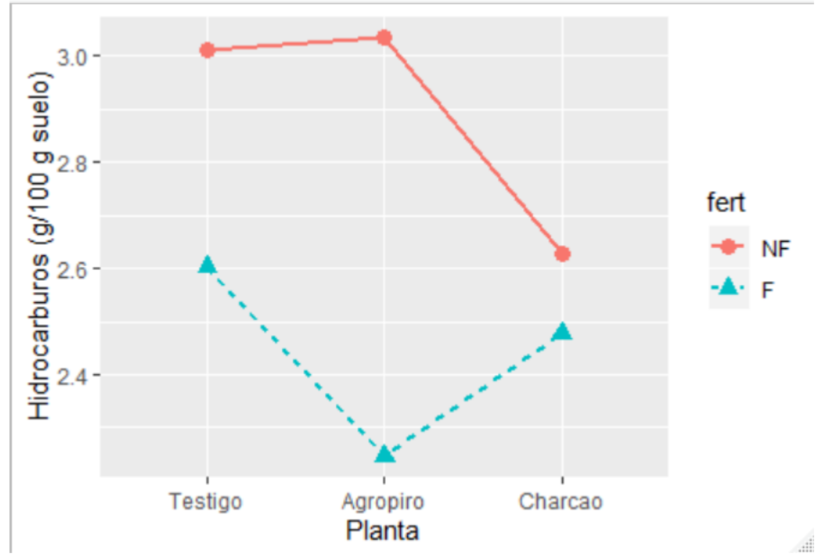
Diseños desbalanceados

- Cuando los diseños factoriales están desbalanceados (distinta cantidad de réplicas en las combinaciones) las SC dejan de ser ortogonales y los resultados pueden diferir según cómo se calculen
- La pérdida de ortogonalidad puede darse también por asociación entre las VE (infrecuente en experimentos pero muy frecuente en estudios observacionales)
- Existen distintos métodos para calcularlas (tipo I, tipo III), que difieren en cómo se calculan las medias marginales

- Simulo desbalanceo en el diseño

```
FITOR[sample(1:nrow(FITOR), 20, replace=FALSE),]
```

- Calculo SC tipo I y tipo III




```
> anova(modelo3)  #solo tipo I
Analysis of Variance Table

Response: HC
          Df  Sum Sq Mean Sq F value    Pr(>F)
veg        2  1603.5   801.8  1.8792    0.1893
fert       1 13041.8 13041.8 30.5677 7.431e-05 ***
veg:fert   2  3297.3  1648.7  3.8642    0.0461 *
Residuals 14  5973.1   426.7
```

```
fert      veg HC
   F Agropiro  3
  NF Agropiro  4
   F  Charcao  2
  NF  Charcao  5
   F  Testigo  3
  NF  Testigo  3
```

```
> Anova(modelo3, type="III") #paquete car
Anova Table (Type III tests)

Response: HC
            Sum Sq Df  F value    Pr(>F)
(Intercept) 154587  1 362.3254   2.1e-11 ***
veg            980  2   1.1487 0.3451918
fert         11109  1  26.0376 0.0001609 ***
veg:fert      3297  2   3.8642 0.0461000 *
Residuals     5973 14
```

Sumas de cuadrados

- **SC secuenciales o Tipo I**
  - Particionan la SC del modelo según la secuencia de incorporación de términos => el orden importa
  - Miden la contribución de una VE siendo que las VE que la preceden en el modelo ya están incluidas en el mismo
  - Como pesa las medias marginales por la cantidad de observaciones, el tamaño de las celdas importa
  - Es una verdadera partición de la SC total

$$SC_{X_1}$$

$$SC_{X_2/X_1}$$

$$SC_{X_3/X_1,X_2}$$

- **SC parciales, ajustadas o Tipo III**
  - Miden la contribución de una VE siendo que todas las otras VE ya están incluidas en el modelo => el orden no importa
  - se basa en medias marginales sin ponderar (les da el mismo peso independientemente de la cantidad de observaciones)
  - No es una verdadera partición de la SC total

$$SC_{X_1/X_2,X_3}$$

$$SC_{X_2/X_1,X_3}$$

$$SC_{X_3/X_1,X_2}$$

Resumiendo:

Cuando hay ortogonalidad entre las VE:

- La variabilidad explicada por un factor es la misma, independientemente si la otra VE es tenida en cuenta o no
- SC total puede descomponerse en fuentes de variación independientes y aditivas
- SC secuencial = SC parcial o ajustada
- En experimentos diseñados y balanceados, las SC son ortogonales. En regresión múltiple sólo se daría si las VE fueran estrictamente independientes entre sí (correlación nula). Imposible en estudios observacionales…

Cuando no hay ortogonalidad entre las VE:

- La significación de la interacción es la misma en ambos métodos
- La SC tipo III respeta el principio de marginalidad, más recomendable
- Alternativamente los datos pueden analizarse utilizando un diseño de celdas (comparando a x b tratamientos)