BIOMETRÍA II
CLASE 7
REGRESIÓN MÚLTIPLE

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

Valores de referencia para pruebas de función pulmonar

- Continuando con el estudio, en 50 hombres se registró la edad, altura (en cm) y peso (en kg), con el objetivo de estimar la ventilación voluntaria máxima (VVM) (en litros/min)

- VR
- VE
- Modelo

resp.csv

Regresión lineal múltiple

- Una única variable respuesta o dependiente (Y) cuantitativa y más de una variable VE, explicativas o independientes (varias X), que pueden ser cuanti o cualitativas
- Sin interacción (efectos aditivos): el efecto de $X1$ sobre la respuesta media no depende del nivel de $X2$ y viceversa

$$\varepsilon_i \approx NID(0, \sigma^2)$$

$$Y_i = \beta_0 + \beta_1 x_{1_i} + \ldots. + \beta_k X_{k_i} + \varepsilon_i$$

$$E(Y / X_1, \ldots X_k) = \beta_0 + \beta_1 x_{1_i} + \ldots. + \beta_k X_{k_i}$$

- Pueden agregarse al modelo anterior términos cuadráticos, cúbicos, etc:

$$E(Y / X_1, \ldots X_k) = \beta_0 + \beta_1 x_{1_i} + \beta_2 x_{1_i}^{\ 2} + \ldots. + \beta_k X_{k_i}$$

- β_i son los coeficientes de regresión parcial, ya que indican la influencia (parcial) de cada VE sobre Y, cuando se mantiene constante la influencia de las otras VE
- k es la cantidad de VE, por lo tanto la ecuación del modelo posee p parámetros

Múltiples VE

- Lo ideal es que cada una proporcione información "independiente" sobre el comportamiento de Y => modelos sin información redundante, más parsimoniosos
- Si ello ocurre, se dice que las VE son ortogonales

> Dos variables VE son ortogonales (independientes) cuando el conocimiento de una no proporciona información sobre la otra, es decir que no están asociadas

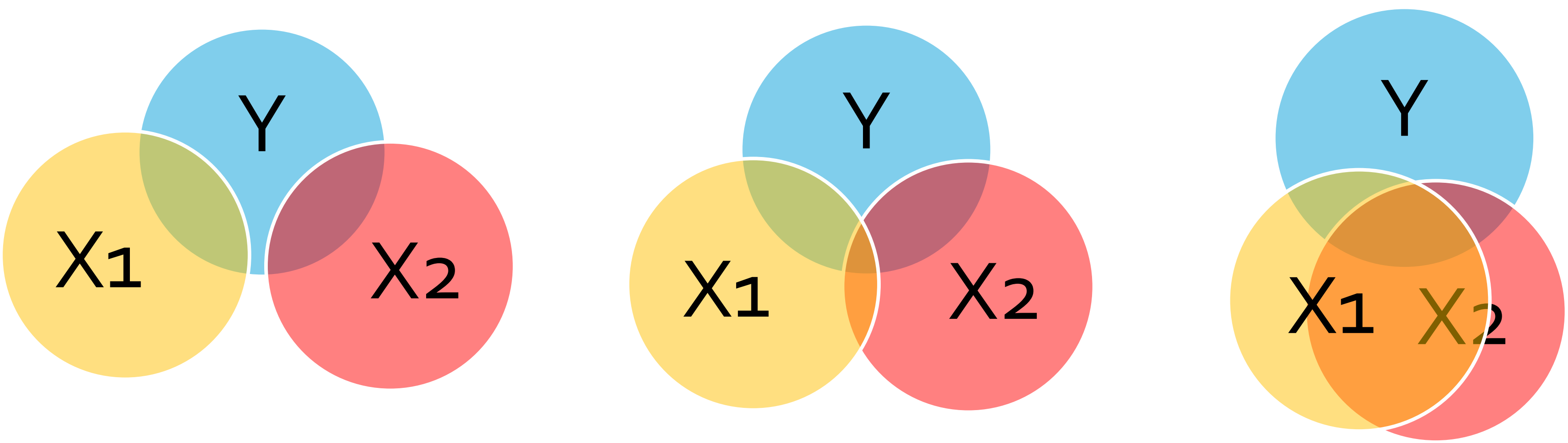
- Es habitual en experimentos diseñados pero casi imposible en estudios observacionales
- Si las VE están asociadas linealmente se habla de colinealidad

> Dos variables VE son colineales cuando están asociadas linealmente. Si una VE es combinación lineal exacta de otra, la colinealidad es perfecta y la estimación por cuadrados mínimos de los coeficientes β no tiene una solución única

La regresión múltiple busca estimar la contribución independiente de X_1 y X_2 a la variación de Y, es decir, estimar β_1 y β_2.

(1) X_1 y X_2 son independientes, ortogonales. La asociación entre ellas es nula
(2) X_1 y X_2 están asociadas débilmente
(3) X_1 y X_2 están asociadas fuertemente

(1) (2) (3)



variabilidad conjunta

Aunque la variación de Y explicada por X_1 y X_2 es similar a (1), cuanto mayor es la asociación entre X_1 y X_2 menor es la contribución independiente de cada variable y eventualmente se convierte no significativa

¿Cómo estudiamos asociación entre variables cuantitativas?

- Gráficamente: matrices de diagramas de dispersión
- Analíticamente: Coeficiente de correlación lineal de Pearson ρ
- Mide el grado de asociación lineal entre dos variables aleatorias
- No depende de las unidades de medida de las variables originales
- Toma valores entre [-1,1]. Cuanto más cerca esté de +1 o -1 más fuerte será el grado de relación lineal (siempre que no existan datos anómalos)
- Su **signo** nos indica si la posible relación es directa o inversa
- Su estimador muestral es:

$$r = \frac{S_{Y_1Y_2}}{S_{Y_1}S_{Y_2}} = \frac{\sum_{i=1}^{n}(y_{i1}-\bar{y}_1)(y_{i2}-\bar{y}_2)}{\sqrt{\sum_{i=1}^{n}(y_{i1}-\bar{y}_1)^2\sum_{i=1}^{n}(y_{i2}-\bar{y}_2)^2}}$$

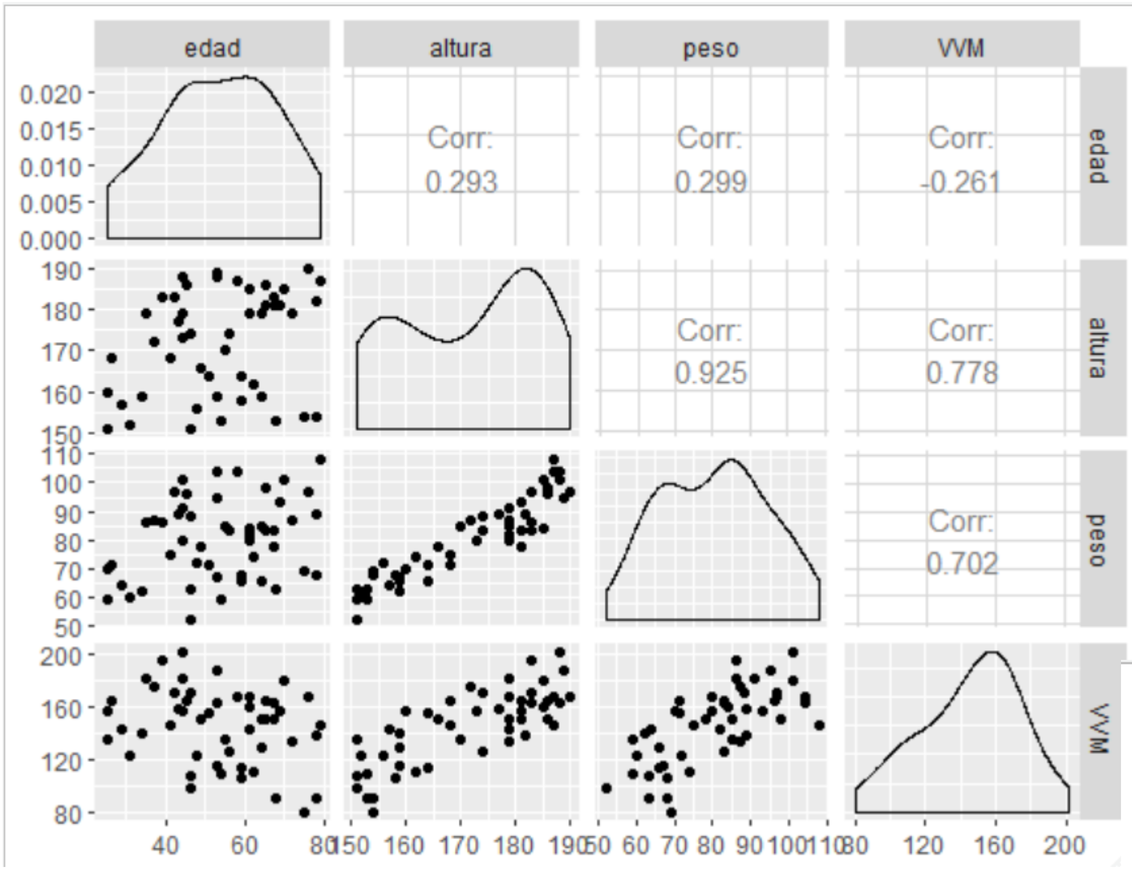
En un experimento diseñado

- Se prueban 4 dosis de un nitrógeno (0, 10, 20 y 30 mg) y 2 temperaturas (20 y 30 C) en plántulas, y se mide la longitud de la plántula (en cm) al mes de iniciado el tratamiento

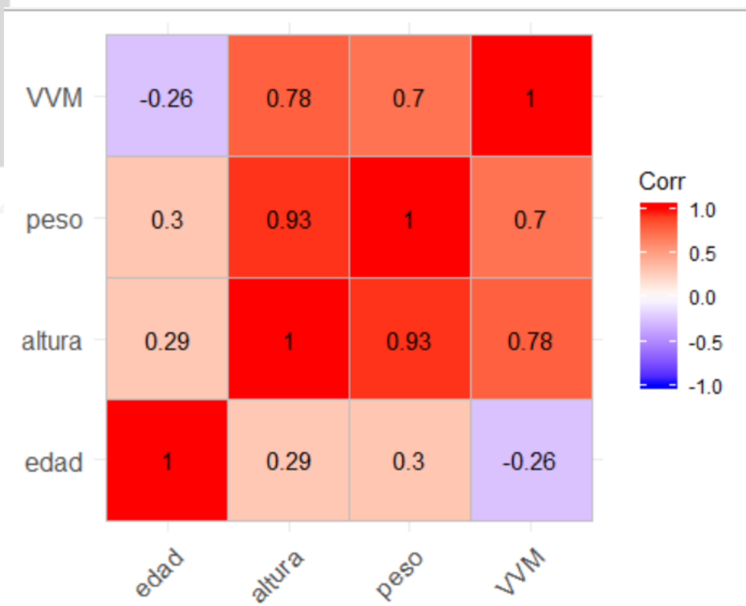
Dosis	Temperatura	Longitud
0	20	89
0	20	88
0	30	66
0	30	59
10	20	93
10	20	73
10	30	82
10	30	77
20	20	100
20	20	67
20	30	57
20	30	68
40	20	69
40	20	59
40	30	62
40	30	59

Coeficiente de correlación lineal entre dosis y temperatura r = 0!
Las VE son ortogonales

En nuestro estudio observacional



heatmap



Colinealidad

¿Qué provoca?

- Las estimaciones de los coeficientes tendrán varianzas muy altas (alto EE), es decir que tendrán poca precisión. Eso puede provocar que las PH individuales sean no significativas aunque el modelo global sea significativo o el R^2 sea alto
- Los coeficientes de regresión pueden presentar signos contrarios a los esperados
- Sin embargo, los estimadores de los valores esperados de la VR seguirán siendo insesgados

$$EE_{\hat{\beta}_j} = \sqrt{\frac{S_e^2}{(1-R_j^2)\cdot\Sigma(x_i-\bar{x})^2}}$$

Menor EE cuanto mayor es la dispersión de X

Se efectúa una regresión de X_i en función de las restantes VE y se calcula el R^2. Si las VE son ortogonales, $R^2_j = 0$

```
m1<-lm(vvm ~ edad + altura + peso)
m2<-lm(vvm ~ edad + peso + altura)
```

SC secuencial o Tipo I

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

```
> anova(m1)
Analysis of Variance Table

Response: VVM
          Df Sum Sq Mean Sq  F value    Pr(>F)
edad       1   2694    2694  23.5461  1.44e-05 ***
altura     1  31678   31678 276.8207 < 2.2e-16 ***
peso       1      3       3   0.0287    0.8662
Residuals 46   5264     114
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
> anova(m2)
Analysis of Variance Table

Response: VVM
          Df  Sum Sq Mean Sq F value    Pr(>F)
edad       1  2694.5  2694.5  23.546 1.440e-05 ***
peso       1 26470.4 26470.4 231.317 < 2.2e-16 ***
altura     1  5210.4  5210.4  45.532 2.179e-08 ***
Residuals 46  5263.9   114.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
```

SC marginal o Tipo III

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

```
Anova Table (Type III tests)

Response: VVM
             Sum Sq Df F value    Pr(>F)
(Intercept)  2570.5  1 22.4628 2.095e-05 ***
edad        10263.9  1 89.6928 2.229e-12 ***
altura       5210.4  1 45.5323 2.179e-08 ***
peso            3.3  1  0.0287    0.8662
Residuals    5263.9 46
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '
> Anova(m2, type="III")
Anova Table (Type III tests)

Response: VVM
             Sum Sq Df F value    Pr(>F)
(Intercept)  2570.5  1 22.4628 2.095e-05 ***
edad        10263.9  1 89.6928 2.229e-12 ***
peso            3.3  1  0.0287    0.8662
altura       5210.4  1 45.5323 2.179e-08 ***
Residuals    5263.9 46
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '
>
```

Obviamente, si las VE son ortogonales, ambas SC coinciden

Es la que calcula lm

Colinealidad

¿Cómo se detecta? Estudiando la asociación entre VE

- Gráficos de dispersión y coeficientes de correlación para todos los pares posibles de X (pero solo detecta colinealidad de a pares)
- Efectuando una regresión de X_i en función de las restantes VE y calculando R^2. El proceso se efectúa para todas las VE. Valores cercanos a 1 indican problemas de colinealidad que pueden involucrar a más de una VE
- VIF (factor de inflación de la varianza): mide para cada X el aumento de la varianza del coeficiente de regresión debido a la correlación entre VE

$$VIF_j = \frac{1}{1 - R_j^2}$$

donde R^2_j es el R^2 que se obtiene al efectuar una RLM con X_j como VD vs las demás X

- Toma valores entre 1 e infinito. Valores superiores a 5 indican colinealidad importante

```
> library(car)
> vif(m1)
    edad   altura     peso 
1.100223 6.967441 6.994053 
```

Colinealidad

¿Cómo se resuelve?

- Eliminando variables: aquellas que proporcionan una información que se obtiene de otras variables ya incluidas en el modelo. Pero ojo, eso no significa que no estén asociadas con la VR
- Combinando las VE asociadas en una nueva variable (técnicas multivariadas) o en índices

¿Mejor modelo?


```
m1<-lm(VVM~ edad+altura+peso)
m2<-lm(VVM~ edad+peso)
m3<-lm(VVM~ edad+altura)
m4<-lm(VVM~ edad)
m5<-lm(VVM~ altura)
```

Selección de modelos:

- Error estándar residual: por ajuste
- R2 ajustado: por ajuste y parsimonia
- AIC: por ajuste y parsimonia
- ECMP: por predicción
- Pruebas de hipótesis: para VE y modelos

	sigma	R2	R2 ajust	df	AIC
m1	10.70	0.867	0.859	5	384.724
m2	14.93	0.736	0.725	4	417.127
m3	10.59	0.867	0.861	4	382.756
m4	27.74	0.068	0.049	3	478.152
m5	18.04	0.606	0.598	3	435.108

	RMSE	Rsquared	MAE	ER
1	11.026	0.847	9.067	7.57
2	15.347	0.703	12.671	10.54
3	10.813	0.853	8.867	7.42
4	28.301	0.013	23.898	19.43
5	18.416	0.573	15.825	12.64

Pruebas de hipótesis para comparar modelos

- Para modelos anidados:

El modelo 2 está anidado en el modelo 1 si todas las VE que se encuentran en el modelo 2 se incluyen en el modelo 1 , es decir, el conjunto de VE en el modelo 2 es un subconjunto del conjunto de VE en el modelo 1. Si dos modelos están anidados, el más complejo puede convertirse en el más simple si algunos coeficientes se hacen nulos

$$Modelo\ 1\ Y_i = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_{ijk} \qquad a\ \ parámetros$$

$$Modelo\ 2\ \ Y_i = \beta_o + \beta_1 X_1 + \varepsilon_{ijk} \qquad b\ parámetros$$

b<*a*, el modelo 1 (más simple, reducido) está anidado en el modelo 2

El criterio para establecer si una o un conjunto de VE deben ser retenidas en un modelo es determinar la significación de la reducción en la SC residual

$$F = \frac{(SCres_2 - SCres_1)/(GL_2 - GL_1)}{SCres_1 / GL_1}$$

```
anova (modelo1, modelo 2)
```

o

```
drop1(modelo1, test="F")
```

¿Mejor modelo?

```
m1<-lm(VVM~ edad+altura+peso)
m2<-lm(VVM~ edad+peso)
m3<-lm(VVM~ edad+altura)
m4<-lm(VVM~ edad)
m5<-lm(VVM~ altura)
```


Comparando modelos por anova (extra SC)

RSS: SCresidual

```
anova(m1,m2)
```

```
Analysis of Variance Table

Model 1: VVM ~ edad + altura + peso
Model 2: VVM ~ edad + peso
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     46  5263.9
2     47 10474.4 -1   -5210.4 45.532 2.179e-08 ***
```

Diferencia en la cant.de parámetros y en la SCres

m2 significativamente **peor** que m1

```
anova(m1,m5)
```

```
Analysis of Variance Table

Model 1: VVM ~ edad + altura + peso
Model 2: VVM ~ altura
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     46  5263.9
2     48 15619.9 -2    -10356 45.249 1.366e-11 ***
```

```
anova(m2,m3)
```

¿Mejor modelo?

```
m1<-lm(VVM~ edad+altura+peso)
m2<-lm(VVM~ edad+peso)
m3<-lm(VVM~ edad+altura)
m4<-lm(VVM~ edad)
m5<-lm(VVM~ altura)
```


`drop1(m1)`

Compara por anova (extra SC) el modelo completo (m1) con modelos anidados, eliminando una variable a la vez. Además informa AIC

```
       Single term deletions

       Model:
       VVM ~ edad + altura + peso
              Df Sum of Sq     RSS    AIC F value    Pr(>F)
m1     <none>                5263.9 240.83
       edad    1   10263.9 15527.8 292.92 89.6928 2.229e-12 ***
m2     altura  1    5210.4 10474.4 273.23 45.5323 2.179e-08 ***
m3     peso    1       3.3  5267.2 238.86  0.0287    0.8662
```

m2 significativamente **peor** que m1; equivale a las pruebas t del summary

Inferencia multimodelo

- Burnham et al (2011). AIC model selection and multimodel inference in behavioral ecology. Behavioral Ecology and Sociobiology, 65(1), 23-35
- Se estiman todos los modelos posibles (anidados o no anidados)
- Se rankean según la teoría de la información (AIC)
- No utiliza PH
- Los modelos tienen distinto “peso” basado en la evidencia muestral

Todos los modelos posibles

AIC corregido para muestras pequeñas


```
library(MuMIn)
dredge(lm(VVM~ edad+peso+altura, na.action = "na.fail"))
```

```
Model selection table
  (Intrc) altur    edad     peso df   logLik  AICc delta weight
4 -155.30 2.075 -1.0280             4 -187.378 383.6  0.00  0.772
8 -159.90 2.124 -1.0260 -0.04902 5 -187.362 386.1  2.44  0.228
7   58.70       -0.9930  1.74300 4 -204.563 418.0 34.37  0.000
2 -150.70 1.727                  3 -214.554 435.6 51.98  0.000
6 -174.70 1.988         -0.25870 4 -214.406 437.7 54.06  0.000
5   30.86                1.42900 3 -220.864 448.3 64.61  0.000
3  172.40       -0.5012          3 -236.076 478.7 95.03  0.000
1  145.70                        2 -237.836 479.9 96.28  0.000
Models ranked by AICc(x)
```

```
dredge(lm(VVM~ edad*peso*altura, na.action = "na.fail"))
```

```
Model selection table
      (Int)   alt     edd      pes    alt:edd  alt:pes   edd:pes alt:edd:pes
4   -155.30 2.075 -1.0280
8   -159.90 2.124 -1.0260 -0.04902
12  -152.60 2.058 -1.0780           0.0003013
24  -299.60 2.907 -1.0240  1.96700            -0.01124
40  -164.70 2.123 -0.9351  0.01390                     -0.001166
16  -158.70 2.117 -1.0460 -0.04843  0.0001177
```

Tabla 1. Algunas características de la inferencia clásica y multimodelo.

	Inferencia clásica	Inferencia multimodelo
Contraste	Contrasta un estadístico obtenido de los datos muestrales con respecto a una hipótesis "nula" del valor de un parámetro en la población. Informa probabilidad de cometer un error de tipo 1 y de tipo 2.	Contrasta varios modelos anidados y no anidados de manera simultánea. Se intenta evitar el sesgo de "enamorarnos" de una sola hipótesis y ver en los datos información que la sustente, cuando en realidad en muchas situaciones hay más de una hipótesis factible de explicar los patrones observados en los datos.
Valor *P* y error de tipo 1	Indica la probabilidad de obtener un valor igual o más extremo al estadístico muestral si la hipótesis "nula" es verdadera. El valor *P* se compara con una probabilidad fijada a priori de cometer un error de tipo 1 (nivel de significancia).	No aplica.
Criterio de información de Akaike (AIC)	No aplica.	Se utiliza para comparar la parsimonia de los distintos modelos planteados. El contraste es relativo; nos indica cuál o cuáles modelos son los que tienen mayor parsimonia. Es deseable, por lo tanto, incorporar en el contraste un modelo nulo (sin predictores). Otros índices como el AICc, QAIC, cAIC, BIC y DIC se utilizan con fines similares.
r^2	Se utiliza como índice de bondad de ajuste en aquellos casos en los que puede ser calculado, como los modelos lineales generales.	Se utiliza como índice de bondad de ajuste en aquellos casos en los que puede ser calculado, como los modelos lineales generales. El r^2 "ajustado" (el cual penaliza por complejidad del modelo) puede ser usado con fines similares a los del AIC.
Potencia = 1-*P*(error de tipo 2)	Se puede calcular dicha probabilidad al fijar una hipótesis alternativa de interés.	No aplica.
Tamaño muestral	Al aumentar el tamaño muestral la potencia aumenta, pero no influye sobre el nivel de significancia (y por lo tanto la probabilidad de cometer un error de tipo 1). Para una diferencia dada, a mayor tamaño muestral más probabilidad de rechazar la hipótesis nula.	Siempre es deseable tener mayor tamaño muestral, pero ello no influye de manera previsible sobre el orden (ranking) de los modelos a ser comparados. Cuando se emplea el AIC, todos los modelos a comparar deben tener el mismo tamaño muestral.

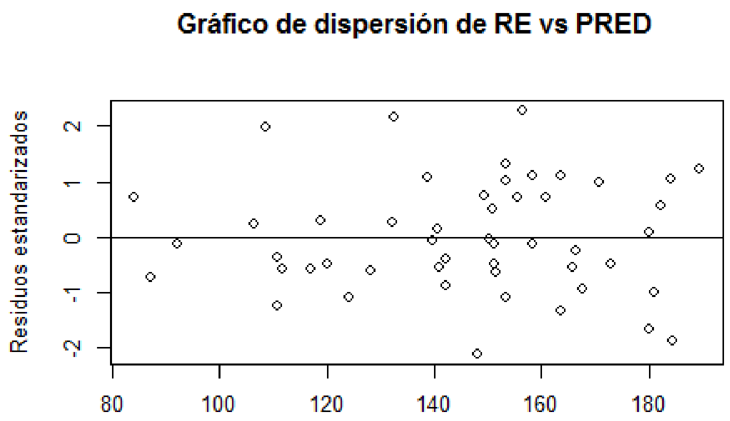
Garibaldi, L. A. et al. (2017). Inferencia multimodelo en ciencias sociales y ambientales. Ecología Austral 27:348-363

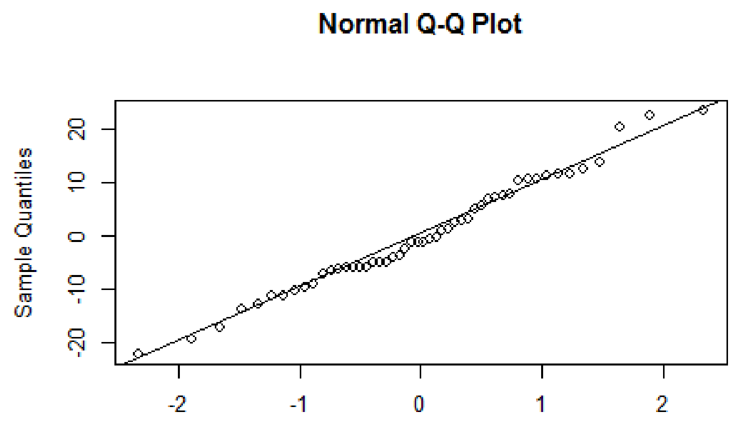
Supuestos y validación m3

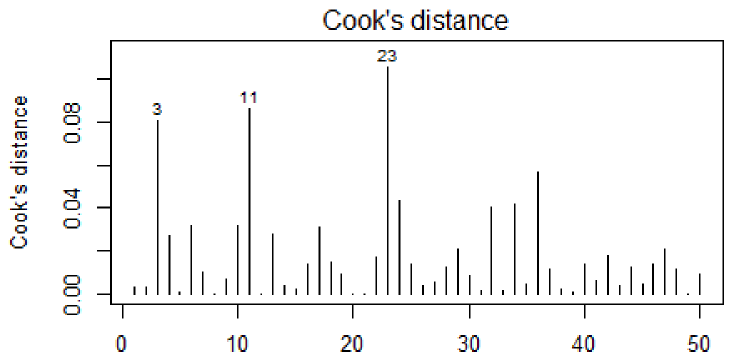
```
vif(m3)
edad      altura
1.093778 1.093778
```

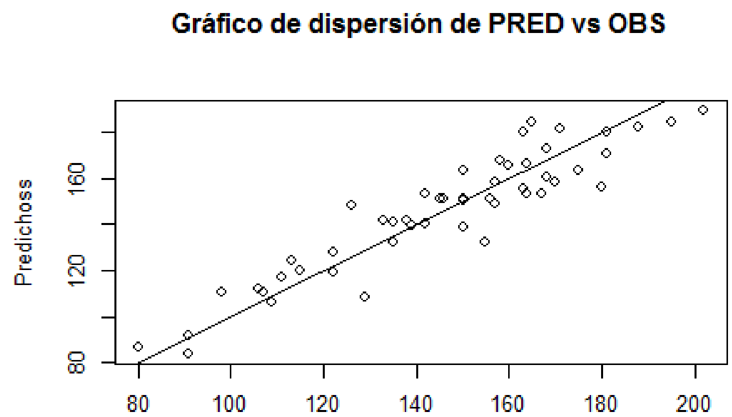


```
m3<-lm(VVM~ edad+altura)
```





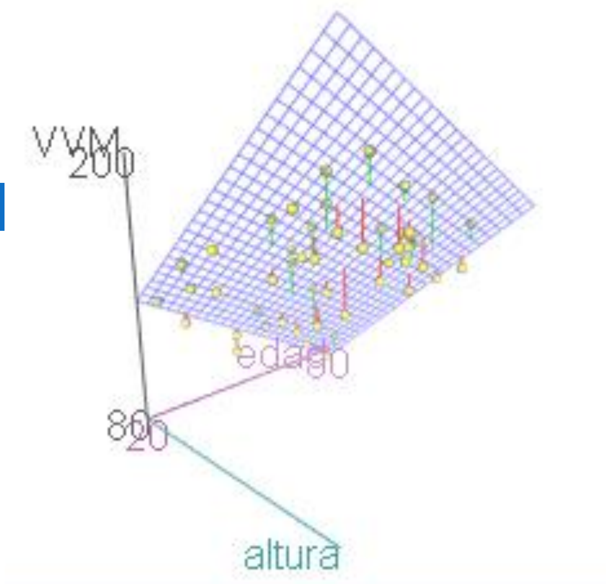




Pruebas de hipótesis

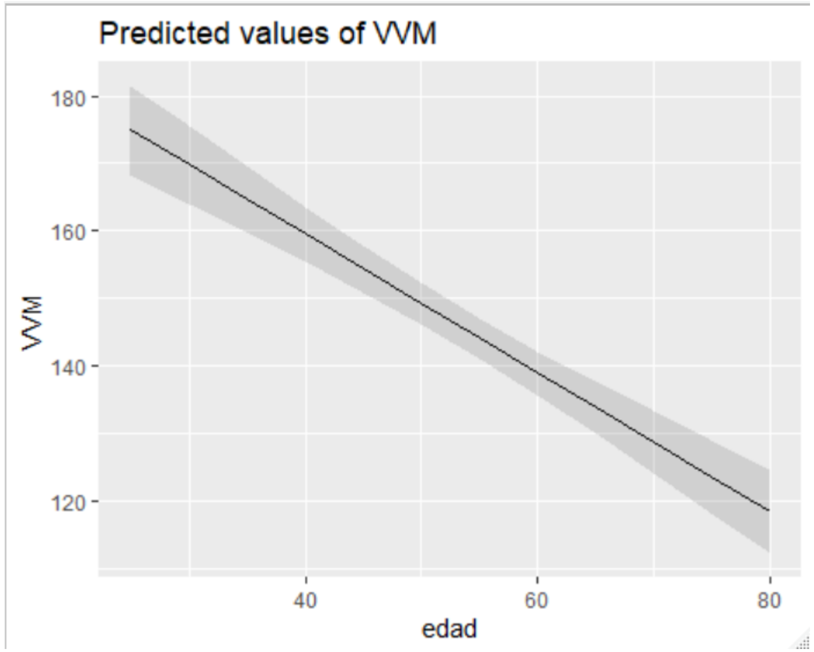

```
m3<-lm(VVM~ edad+altura)
```

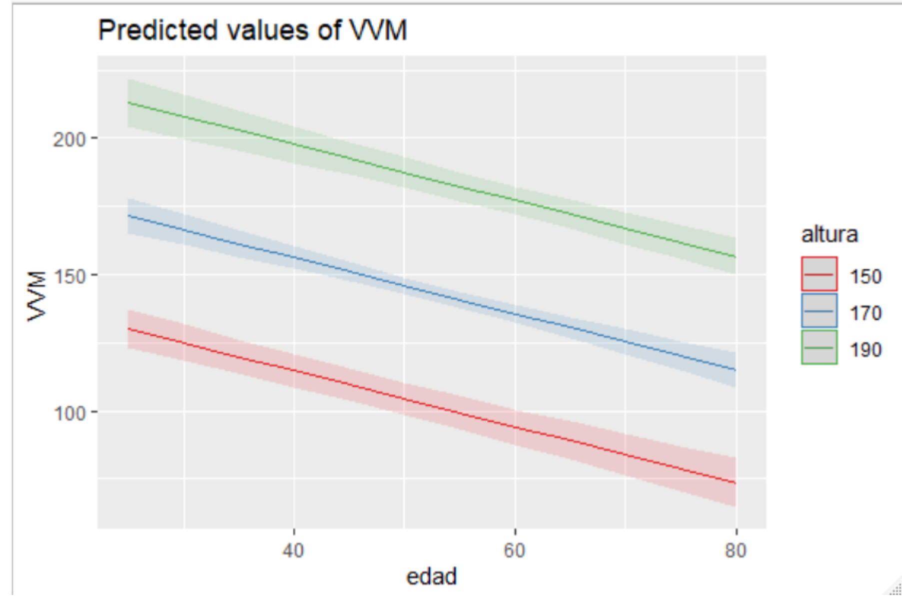
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -155.3453    20.3030  -7.651 8.48e-10 ***
edad          -1.0276     0.1069  -9.611 1.13e-12 ***
altura         2.0746     0.1234  16.813  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.59 on 47 degrees of freedom
Multiple R-squared:  0.8671,    Adjusted R-squared:  0.8615
F-statistic: 153.4 on 2 and 47 DF,  p-value: < 2.2e-16
```

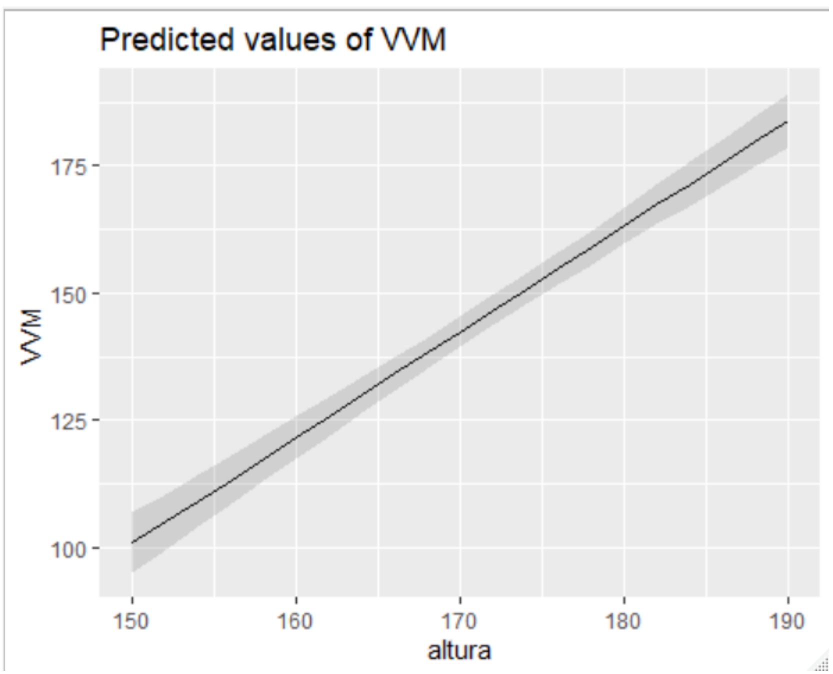


La representación gráfica de modelos múltiples sólo es posible para 2 VE (plano). Para k VE (siendo $k > 2$) ⇨ espacio k-dimensional (hiperplano)

- ✓ ¿Ecuación estimada?
- ✓ ¿Interpretación de los coeficientes?
- ✓ ¿Cómo darle sentido a la ordenada al origen?
- ✓ ¿Porcentaje de la variabilidad en VVM explicada por la edad y la altura?


Predicted values of VVM
VVM
180
160
140
120
40
60
80
edad

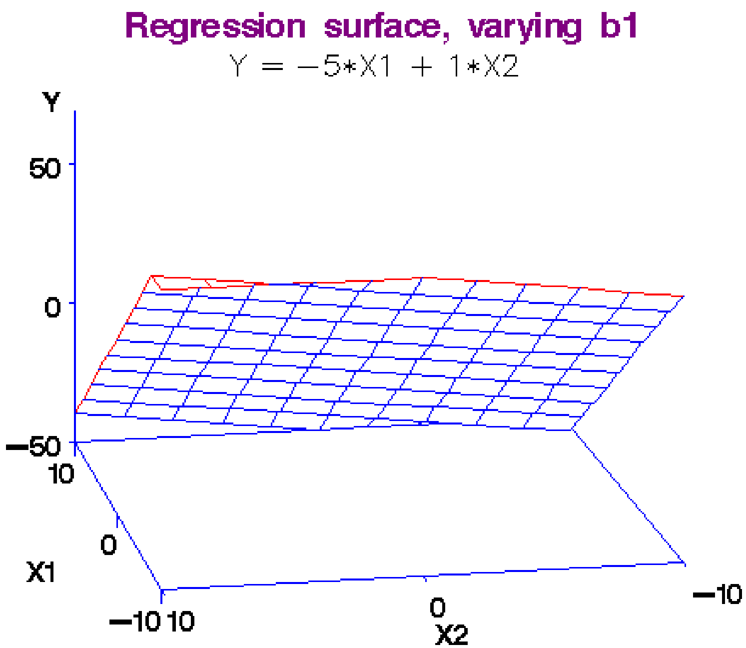

Predicted values of VVM
VVM
200
150
100
40
60
80
edad
altura
150
170
190

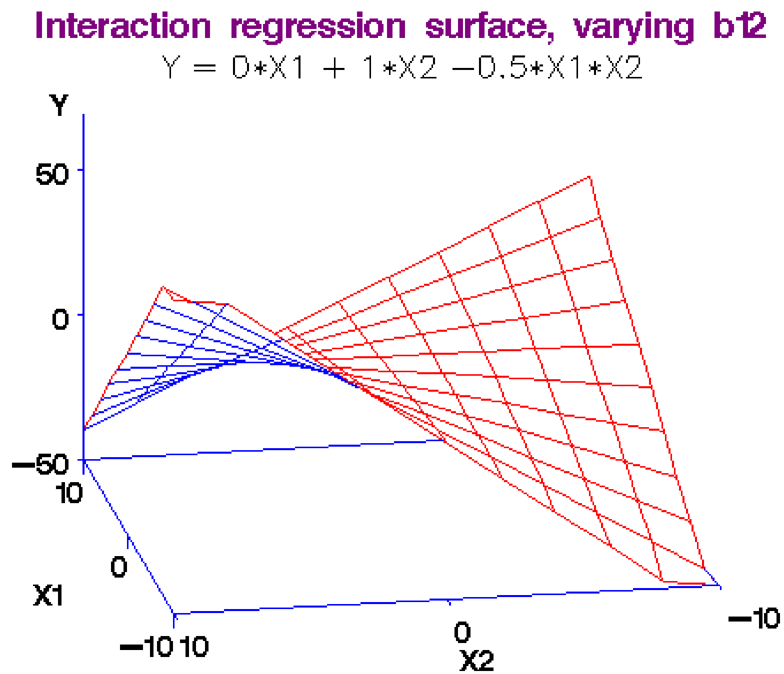

Predicted values of VVM
VVM
175
150
125
100
150
160
170
180
190
altura

Modelo con interacción

$$Y_i = \beta_0 + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \beta_3 X_{1_i} X_{2_i} + \ldots + \varepsilon_i$$

- Con interacción: tanto el efecto de *X1* para un dado nivel de *X2* como el efecto de *X2* para un dado nivel de *X1* dependen del valor de la otra VE





http://www.math.yorku.ca/SCS/spida/lm/visreg.html

¿Interacción significativa?


```
m6<-lm(VVM~ edad*altura)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.526e+02  7.625e+01  -2.001   0.0513 .
edad        -1.078e+00  1.348e+00  -0.800   0.4279
altura       2.058e+00  4.550e-01   4.523 4.27e-05 ***
edad:altura  3.013e-04  7.965e-03   0.038   0.9700
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 10.7 on 46 degrees of freedom
Multiple R-squared:  0.8671,    Adjusted R-squared:  0.8585
F-statistic: 100.1 on 3 and 46 DF,  p-value: < 2.2e-16
```

	sigma	R2	R2 ajust	df	AIC
m1	10.70	0.867	0.859	5	384.724
m2	14.93	0.736	0.725	4	417.127
m3	10.59	0.867	0.861	4	382.756
m4	27.74	0.068	0.049	3	478.152
m5	18.04	0.606	0.598	3	435.108
m6	10.70	0.867	0.867	5	384.754

```
> anova(m4,m7)
Analysis of Variance Table

Model 1: VVM ~ edad + altura
Model 2: VVM ~ edad * altura
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     47 5267.2
2     46 5267.1  1   0.16388 0.0014   0.97
```

Interacción entre variables cuantitativas

- La inclusión de un término de interacción X1*X2 provoca colinealidad entre las variables involucradas y X1*X2, dando lugar a valores altos de VIF. Las estimaciones de los coeficientes son insesgadas pero los EE de X1 y de X2 (pero no de X1*X2) están inflados. Puede solucionarse centrando las variables antes de generar la interacción
- La inclusión de potencias (X^2, X^3, etc) genera el mismo efecto

Ver script clase 7

Interpretación de la interacción

$$Y_i = \beta_0 + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \beta_3 X_{1_i} X_{2_i} + \varepsilon_i$$

Se puede reescribir como:

$$Y_i = \beta_0 + \beta_2 X_{2_i} + (\beta_1 + \beta_3 X_{2_i}) X_{1_i} + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{1_i} + (\beta_2 + \beta_3 X_{1_i}) X_{2_i} + \varepsilon_i$$

El coeficiente de X1 cambia según el valor de X2

El cambio en la respuesta media por el incremento de una unidad en *X1* cuando *X2* se mantiene constante es $\beta_1 + \beta_3 X_2$

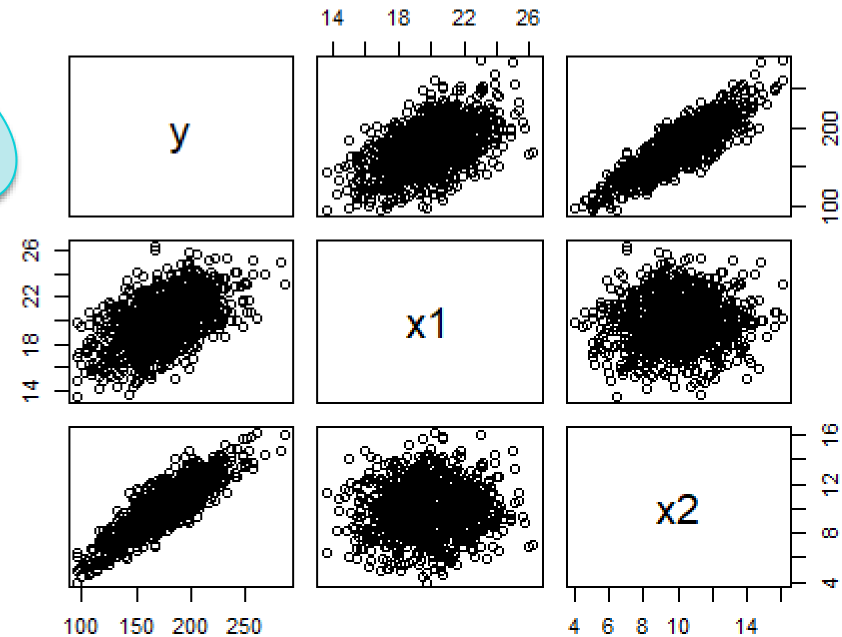
Se pueden elegir valores de X2 (por ej: $\bar{X}, \bar{X} - S, \bar{X} + S$ y calcular el coeficiente para X1

Similarmente, el cambio en la respuesta media con un incremento de una unidad en *X2* cuando *X1* se mantiene constante es $\beta_2 + \beta_3 X_1$ y se pueden elegir valores de X1 y estimar el coeficiente de X2

Simulamos modelo con interacción

Ver script

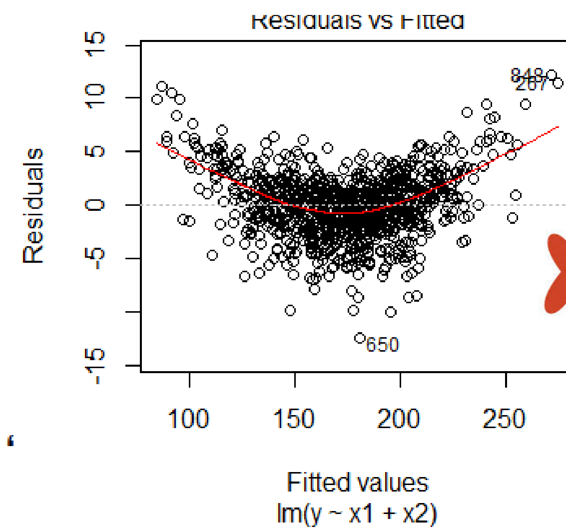
```
x1 = rnorm(1000,20,2)
x2 = rnorm(1000,10,2)
beta0 <-5
beta1 <-2
beta2<-3
beta3<-0.5
e = rnorm(1000,mean=0,sd=2)
y=
beta0+beta1*x1+beta2*x2+beta3*x1*x2+e
bd1<-cbind.data.frame(y,x1,x2)
```



Ajustamos un modelo aditivo

```
m1= lm(y ~ x1+x2, data=bd1)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -92.6117     0.9925   -93.3   <2e-16 ***
x1            6.9572     0.0448   155.4   <2e-16 ***
x2           12.8564     0.0463   278.0   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 3.02 on 997 degrees of freedom
Multiple R-squared:  0.991,     Adjusted R-squared:  0.991
F-statistic: 5.23e+04 on 2 and 997 DF,  p-value: <2e-16
```



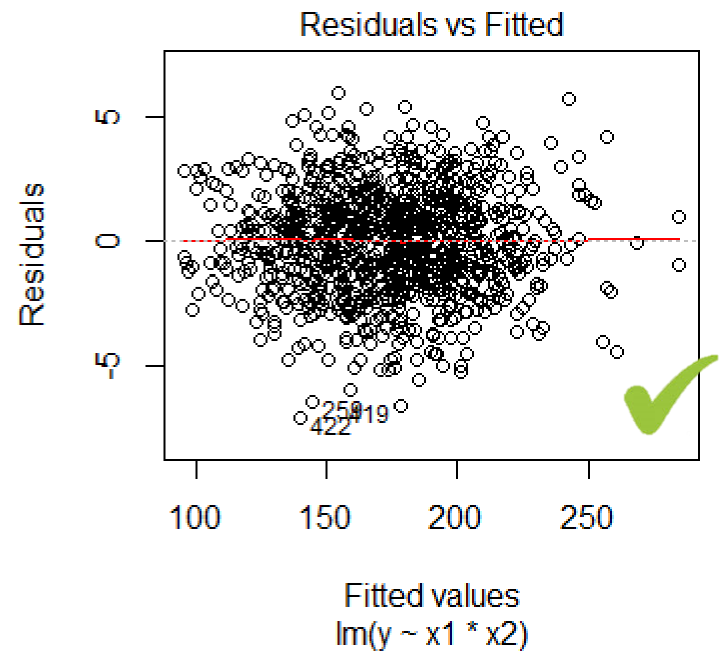

```
vif(m1)
x1 x2 1 1
```

Ajustamos un modelo multiplicativo

```
m1= lm(y ~ x1*x2, data=bd1)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.630      3.003    2.21    0.027 *
x1             1.925      0.151   12.71   <2e-16 ***
x2             2.780      0.299    9.31   <2e-16 ***
x1:x2          0.510      0.015   33.92   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 2.06 on 996 degrees of freedom
Multiple R-squared:  0.996,     Adjusted R-squared:  0.996
F-statistic: 7.55e+04 on 3 and 996 DF,  p-value: <2e-16
```

```
> vif(m2)
  x1    x2 x1:x2
24.7  89.9 116.9
```

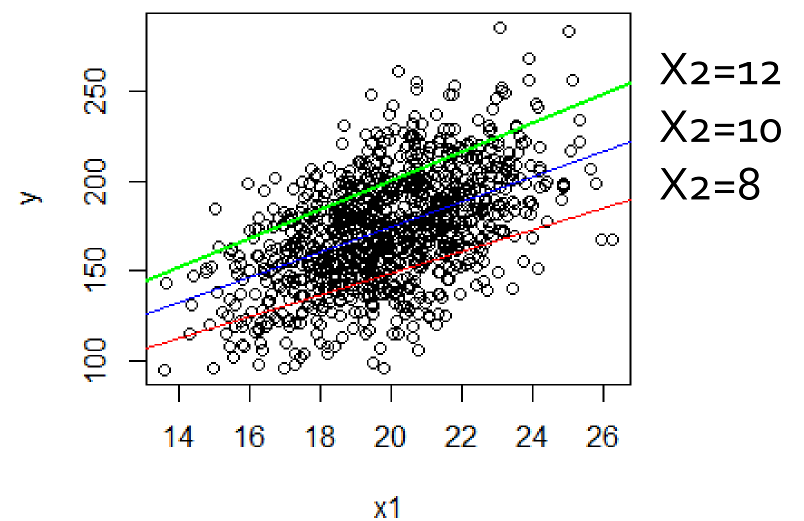


$$Y_i = \beta_0 + \beta_2 X_{2_i} + (\beta_1 + \beta_3 X_{2_i}) X_{1_i} + \varepsilon_i$$

$$\hat{Y}_i = 6{,}63 + 2{,}78 X_{2_i} + (1{,}93 + 0{,}51 X_{2_i}) X_{1_i}$$

X2	Y
$\bar{X} - S$ =10-2=8	28,87+6X1
$\bar{X}$ =10	34,4+7,03X1
$\bar{X} + S$ =10+2=12	40+8,05X1

Idem para X1



Estrategia de análisis

1. Estadística descriptiva
2. Estudiar supuestos: linealidad, igualdad de varianzas, normalidad, outliers, observaciones influyentes, colinealidad
3. Estimación y selección de modelos; eventualmente incluir interacciones (siempre en los experimentos diseñados / según las hipótesis en estudios observacionales)
4. Estudiar el desempeño del modelo final: observados vs predichos; R^2; validación cruzada
5. Magnitud del efecto: a través de los coeficientes de regresión. Algunos sugieren utilizar los coeficientes estandarizados (llamados beta), sin unidades:

$$beta_i = b_i \frac{S_X}{S_Y} = r$$

Construcción de modelos

- Asegurarse de incluir a todas las VE relevantes, basándose en la pregunta de investigación, teoría y conocimiento de la temática
- Ojo con la colinealidad si el objetivo es explicativo. Se pueden combinar las VE que tienden a medir la misma dimensión del fenómeno (por ejemplo mediante un índice o técnicas multivariadas) o seleccionar la más relevante del conjunto
- Considerar la posibilidad de incluir interacciones (principalmente entre variables con mayores efectos)
- Estrategias para retener o eliminar VEs:
  - VE NS pero con el signo esperado: mantener
  - VE NS y sin el signo esperado: eliminar
  - VE S y con el signo esperado: mantener
  - VE S y sin el signo esperado: revisar

Gelman, Andrew, Jennifer Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, 2007

Recomendaciones

- La inclusión de una interacción o un término cuadrático suele inducir colinealidad, que solo afecta los EE de los términos de menor orden
- La no inclusión de un término de interacción relevante puede dar residuos con patrón
- Centrar las X si se desea ganar interpretación en la ordenada al origen
- Se puede ajustar el modelo máximo, registrar el porcentaje de variabilidad explicado e ir descartando términos , o al revés
- También pueden estimarse todos los modelos y rankearlos por algún criterio, por ej AIC (MuMin)
- Considerar el objetivo: ¿predicción o explicación? Ver Shmueli, G. (2010). To explain or to predict?. Statistical science, 25(3), 289-310.
- Evitar ajustar modelos complejos con pocos datos. Algunos sugieren 10 observaciones por cada VE
- En la selección de modelos se debe respetar el principio de marginalidad (ver diapo siguiente)

Principio de marginalidad

Implica que los términos de menor orden no deberían ser removidos antes que los de mayor orden.

- Polinomios:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^{\,2}$$ ✓

$$E(Y) = \beta_0 + \beta_2 X_1^{\,2}$$ ✗ aunque β1 sea NS

- Modelos con interacciones: si el modelo incluye la interacción de X1 * X2, el efecto principal de cada variable debe ser incluido en el modelo

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$ ✓

$$E(Y) = \beta_0 + \beta_3 X_1 X_2$$ ✗

$$E(Y) = \beta_0 + \beta_1 X_1$$ ✓

Aunque si la interacción es significativa no tiene sentido evaluar los coeficientes ni la significación de los efectos principales de las VE involucradas, se sobreentiende que ambas variables afectan a la VR.