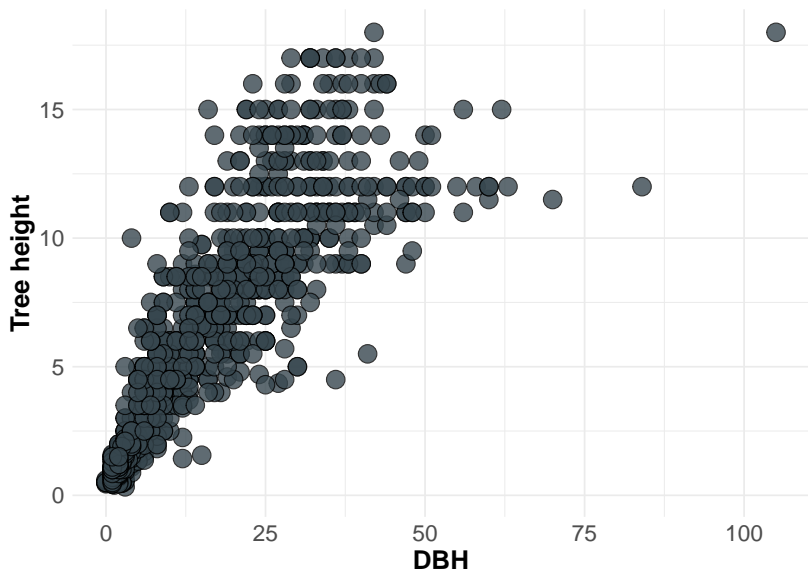


Análisis Exploratorio de Datos

Antonio J. Pérez-Luque

Análisis de datos de seguimiento en la Red de Parques Nacionales

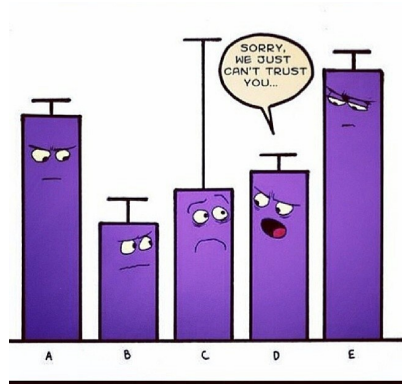
id	date	sp	lat	long	d	id_transect	elev
23812	2012-09-10	Qpy	36.95995	-3.42277	14	P007	1793
23813	2012-09-10	Qpy	36.95995	-3.42277	4	P007	1793
23814	2012-09-10	Qpy	36.95995	-3.42277	18	P007	1793
23815	2012-09-10	Qpy	36.95995	-3.42277	19	P007	1793
23816	2012-09-10	Qpy	36.95995	-3.42277	8	P007	1793
23817	2012-09-10	Qpy	36.95995	-3.42277	4	P007	1793
23818	2012-09-10	Qpy	36.95995	-3.42277	18	P007	1793
23819	2012-09-10	Qpy	36.95995	-3.42277	18	P007	1793



- Es una actitud (**iterativa**) hacia los datos, mas que un conjunto de técnicas
- Detective por un día. No hay una formula mágica
- Encontrar patrones, revelar la estructura, evaluar posibles relaciones...
- Generar preguntas sobre los datos
- Buscar respuestas mediante la visualización, transformación y “modelado” de los datos

Protocolo para explorar los datos

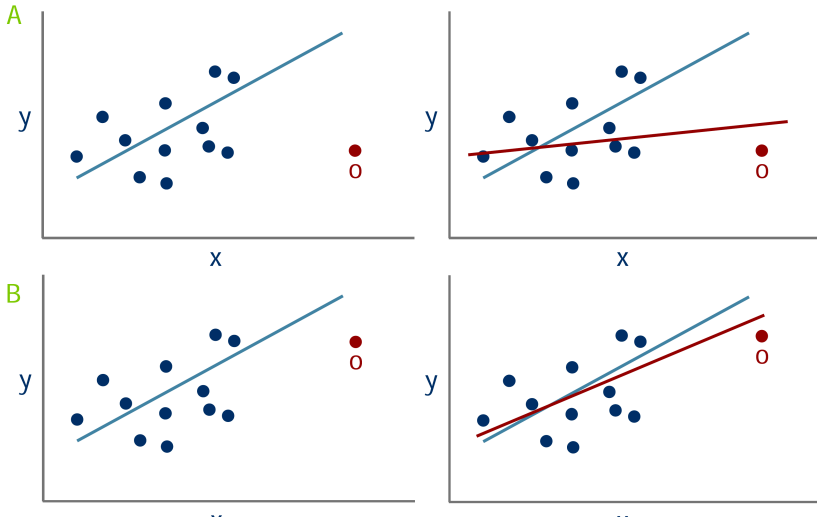
- 1 Detectar valores anómalos (outliers)
- 2 Homogeneidad
- 3 Normalidad
- 4 Relaciones
- 5 Colinealidad
- 6 Interacciones
- 7 Independencia



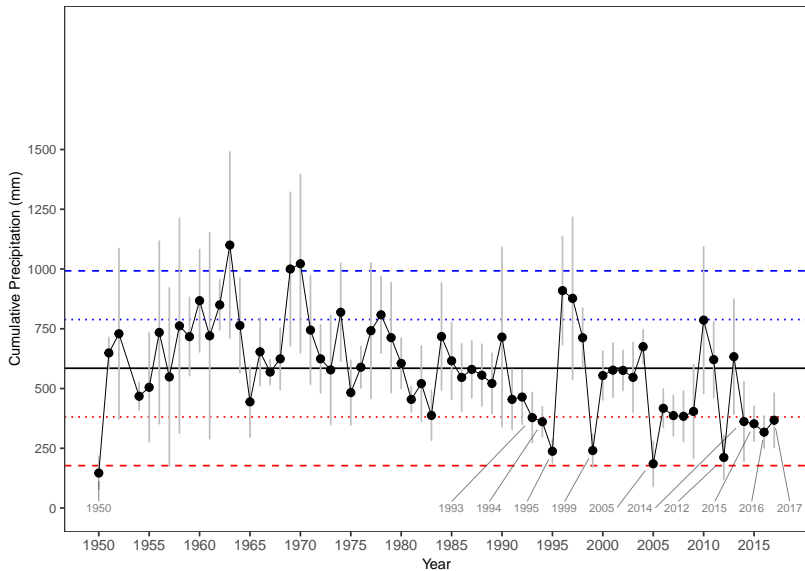
Valores extremos (atípicos) - Outliers

Observaciones extremas, muy alejadas de la mayoría de las observaciones de la variable de interés.

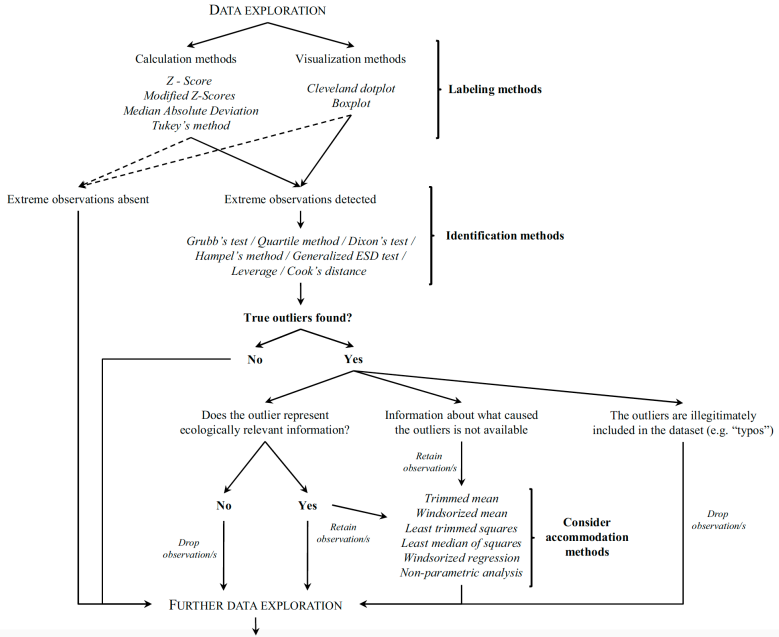
Los **outliers** pueden influir sustancialmente el análisis estadístico.



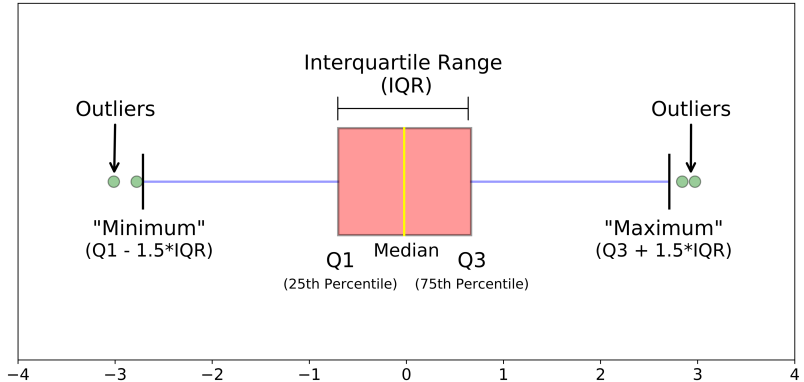
Ojo, los outliers pueden ser el interés de nuestro estudio (epidemiología, eventos extremos)

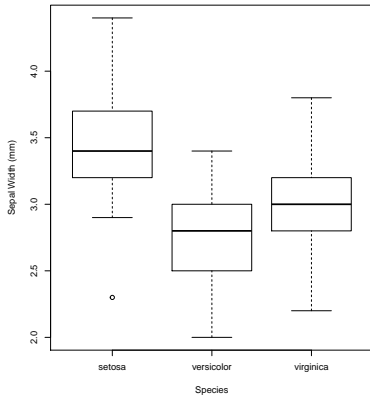
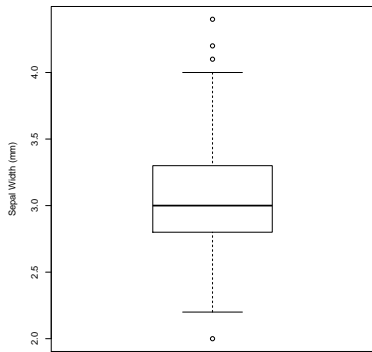


Outliers. ¿Cómo proceder?

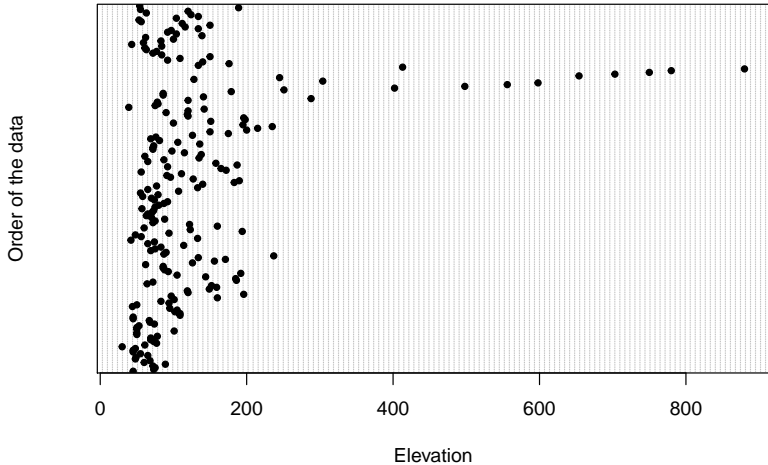


Outliers. Identificación visual





Orden de la observación vs. Valores observados




- Visualización subjetiva
- Aplicación de un test
 - Z-score $Z_i = \frac{Y_i - \bar{y}}{\hat{\sigma}} > 3$
 - Método de Tukey (Boxplot)
 - Test de Grubb. Detecta la presencia de al menos un outlier en el dataset
- package outlier

Biodiversity and Conservation (2018) 27:3295–3300
<https://doi.org/10.1007/s10531-018-1602-2>

COMMENTARY



A conceptual framework to deal with outliers in ecology

Jacinto Benhadi-Marín^{1,2} 

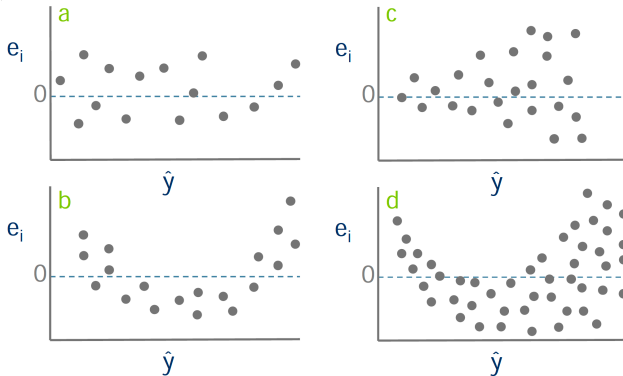
Received: 24 April 2018 / Revised: 18 July 2018 / Accepted: 30 July 2018 / Published online: 1 August 2018
© Springer Nature B.V. 2018

¿Cómo proceder con los Outliers?

- 1 Eliminarlos
- 2 ¿Representan información ecológicamente relevante?
- 3 Construir modelos con/sin outliers
- 4 Aplicar métodos que acomoden la presencia de outliers:
 - Análisis no paramétricos
 - Trimmed means (Anovas Robustas)
 - Aplicar una transformación (cuidado!!)

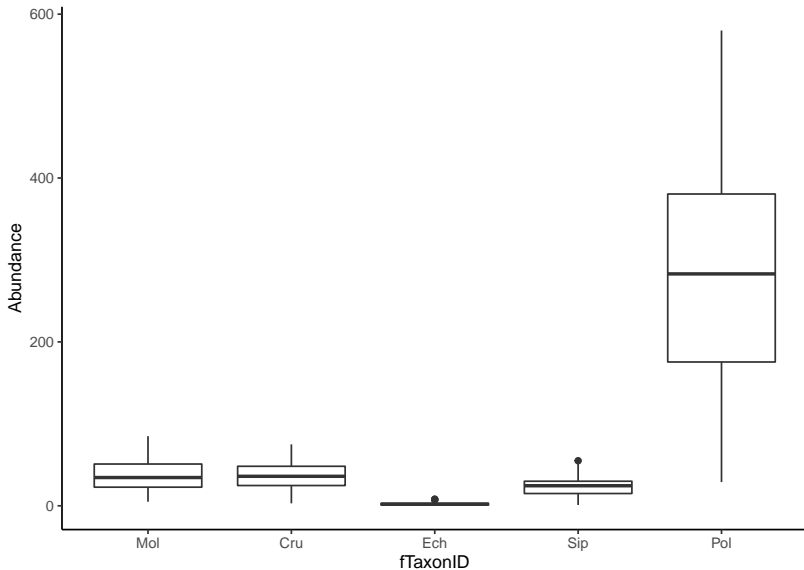
Homogeneidad de la varianza (*homocedasticidad*)

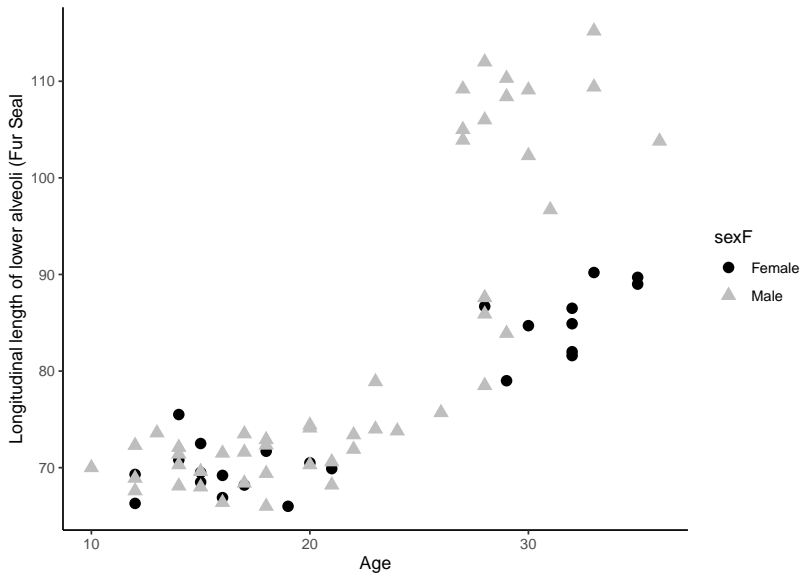
- Asume que la dispersión de todos los posibles valores de la población es la misma para cada valor de la covariable
- Cuando no se cumple, se puede producir una estimación de los errores estándar errónea, lo que implica que los intervalos de confianza que se calculan están sesgados (muy estrechos o muy amplios)
- Mas importante que la normalidad (mod. lineales)



¿Cómo detectar la homogeneidad de las varianzas?

- Boxplot condicionales





Test para detectar la homogeneidad de las varianzas

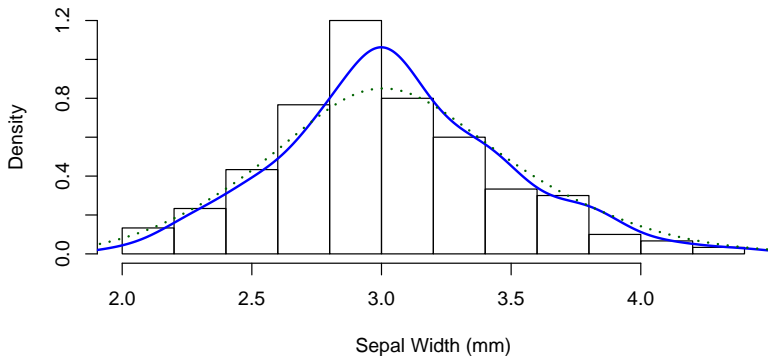
- Test de Barlett (`bartlett.test()`)
 - Asume que la varianza de la muestra o de cada grupo es igual
 - Requiere que los datos sean normales
- Test de Levene
 - Asume igualdad de varianza entre las poblaciones.
 - Menos sensible a la no-normalidad
 - `levene.test()` paquete `lawstat`
- Boxplot de Residuos del modelo vs. Valores ajustados

¿Cómo manejar la heterocedasticidad?

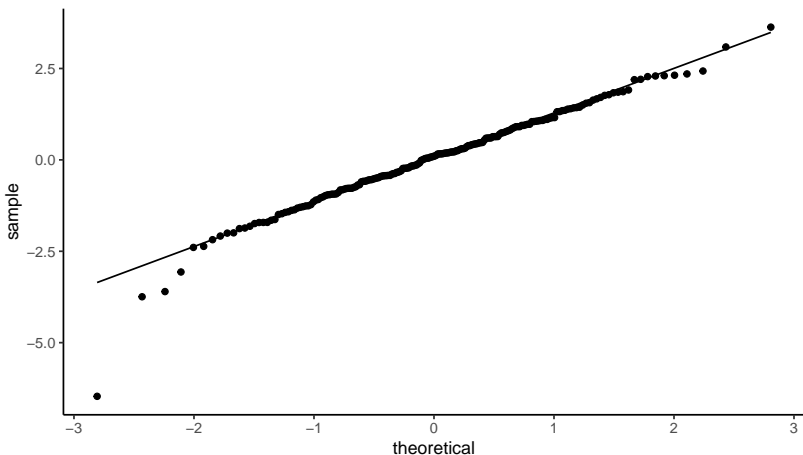
- 1 Transformar la variable respuesta para estabilizar la varianza
- 2 Aplicar técnicas que no requieran heterocedasticidad

Técnicas para detectar normalidad

- Histogramas simples y condicionales
- Gráficos de Densidad (*kernels*)



- Q-Q plots



Test-estadísticos

- Test Shapiro-Wilk (`'shapiro.test()'`) ($n < 30$)
- Test Kolmogorov-Smirnov (`ks.test()`) ($n > 100$)
- Test D'Agostino (sirve además para examinar asimetría y curtosis)

¿Necesitamos que nuestros cumplan el supuesto de normalidad?

No siempre. Depende de la técnica estadística a aplicar:

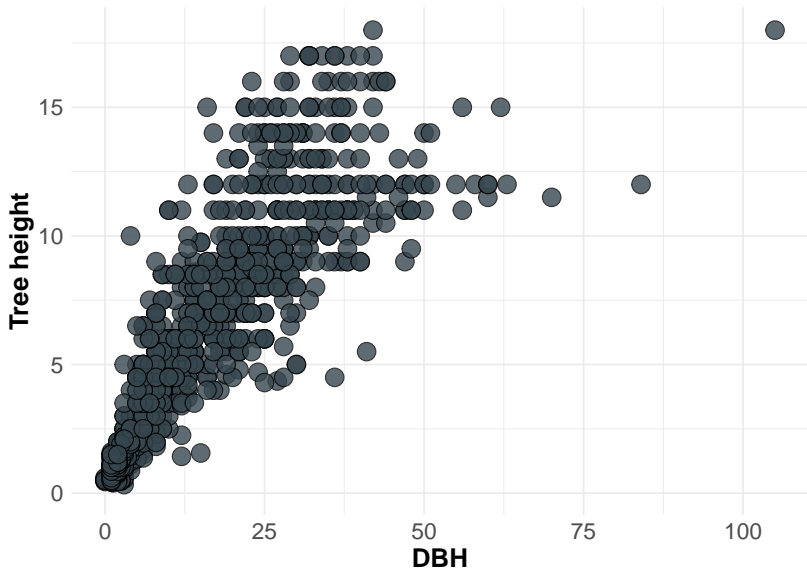
- Análisis de Componentes Principales no requiere normalidad
- Regresión lineal asume normalidad, aunque es razonablemente robusta si no se cumple la normalidad

Soluciones

- Transformar variable respuesta
- ¡Cuidado con la asimetría!

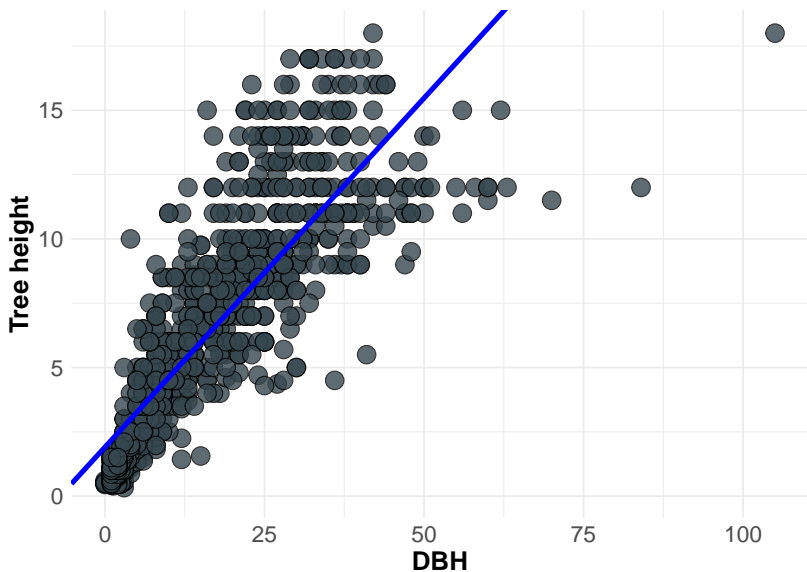
Relación entre variables

- ¿Están las variables asociadas? ¿Cómo están relacionadas?



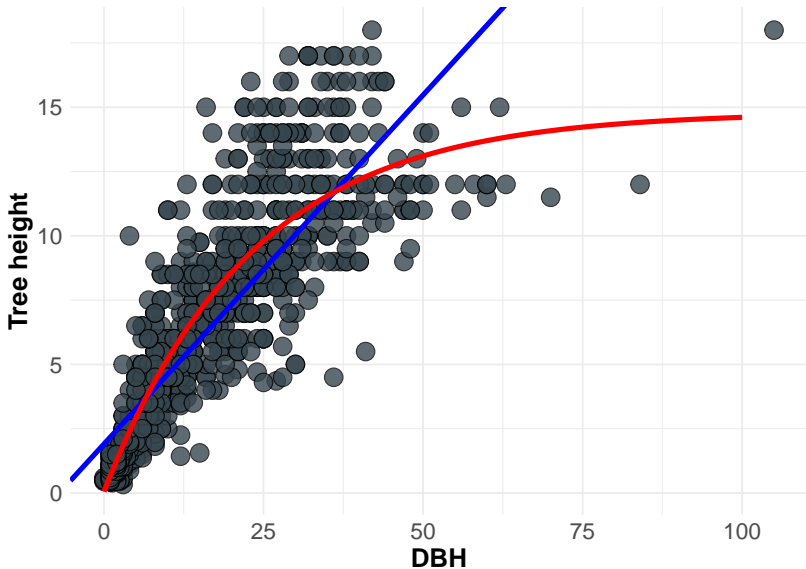
Relación entre variables

- ¿Están las variables asociadas? ¿Cómo están relacionadas?



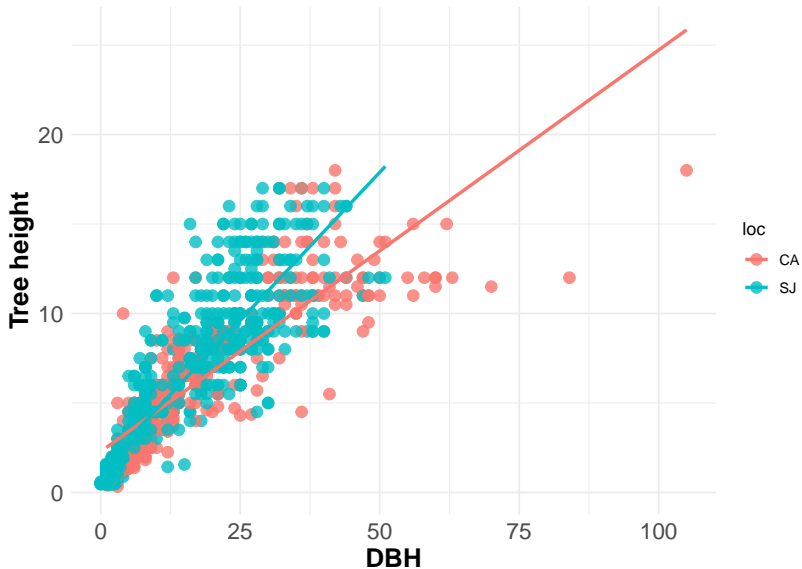
Relación entre variables

- ¿Están las variables asociadas? ¿Cómo están relacionadas?



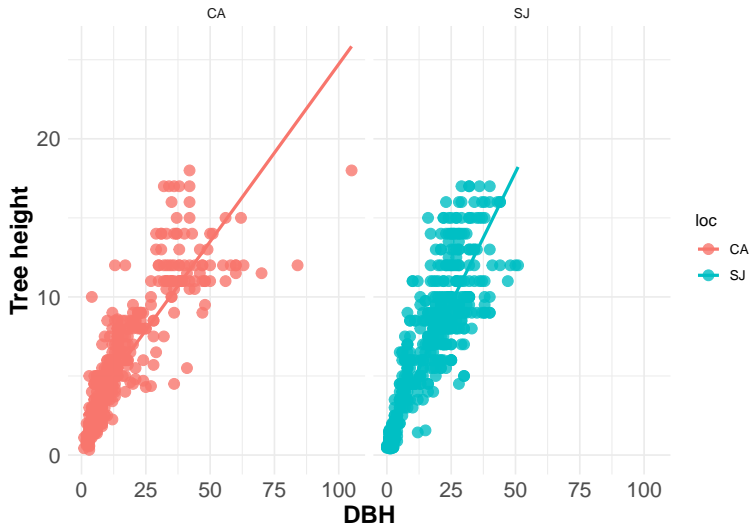
Relación entre variables

- Incluir otras covariables

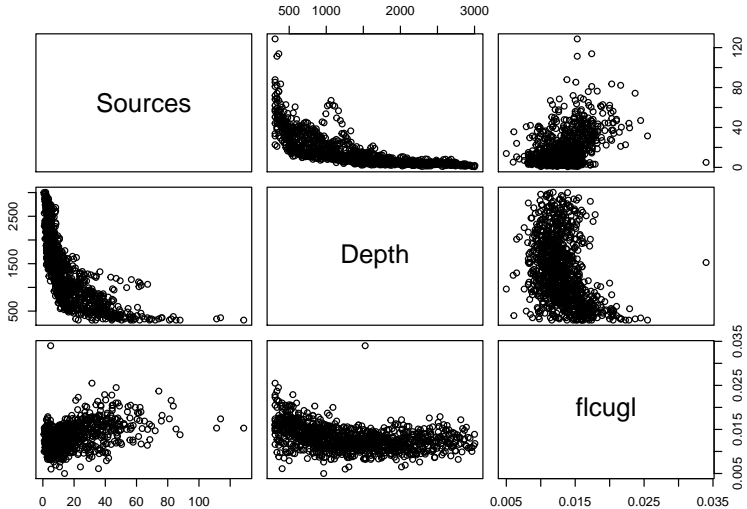


Relación entre variables

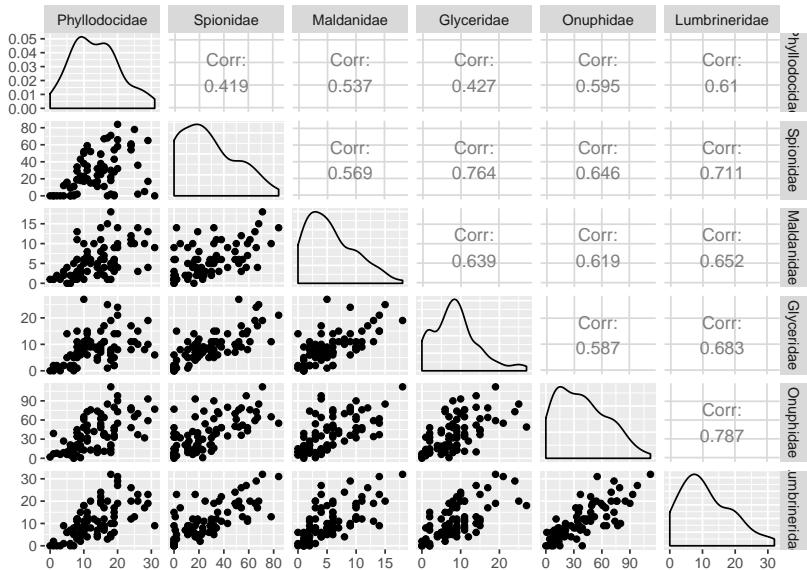
- Incluir otras covariables



Relación entre variables (Pairplots)



Relación entre (muchas) variables



- La colinealidad es la existencia de una correlación entre dos variables explicativas (covariables).
- Es uno de los retos más importantes a la hora de aplicar cualquier técnicas estadística
- Principio de parsimonia (Occam's Razor)
 - Modelo mas simple
 - *"A model should be as simple as possible. But no simpler"* (Einstein)
- Afecta de forma importante a regresiones múltiples, GLMs y a técnicas de análisis multivariante (RDA, CCA, etc).

Métodos para detectarlo

- Matrices de correlación entre variables
- Pairplots
- Factor de Inflación de la Varianza (VIF)

Factor de Inflación de la Varianza (VIF)

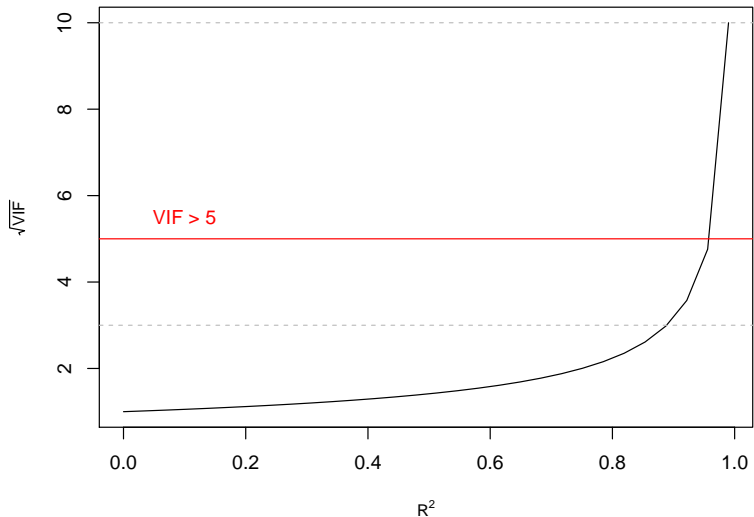
Representa la proporción de la variabilidad (o varianza) de la variable que es explicada por el resto de las variables predictoras del modelo

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

siendo R_j^2 es el coeficiente de determinación de la regresión del j-ésimo regresor sobre el resto.

- A mayor valor, mayor probabilidad de que exista colinealidad
¿Cuándo se considera un valor elevado?

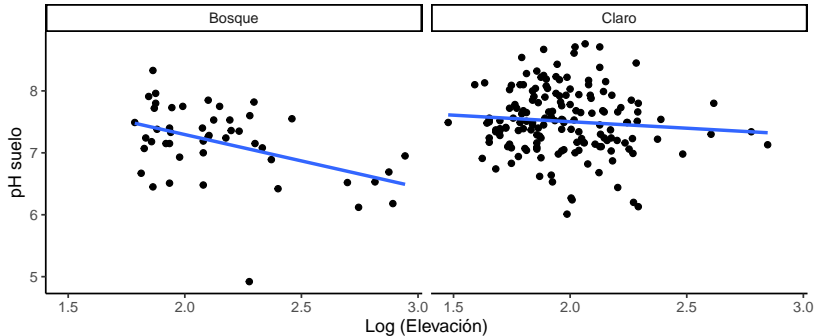
Factor de Inflación de la Varianza (VIF)

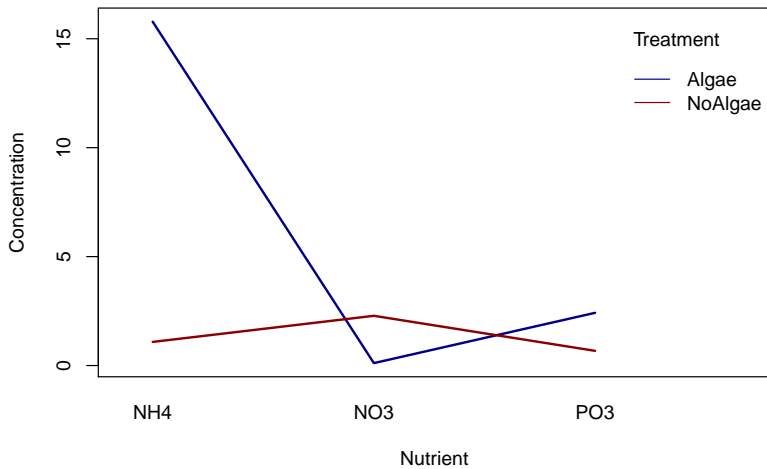


¿Qué hacer cuando existe colinealidad?

- Ir eliminando covariables en función de:
 - VIF
 - sentido ecológico
- Recalcular VIF
- Iterar el proceso
- `vif()` del paquete `car`

- Hablamos de interacción cuando la relación entre las variables x e y depende de otra variable z
- Tipos de interacción:
 - Variable continua - Variable categórica
 - Variables continuas
 - Variables categóricas

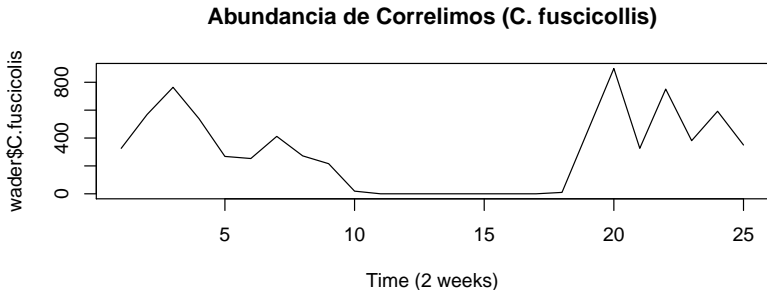




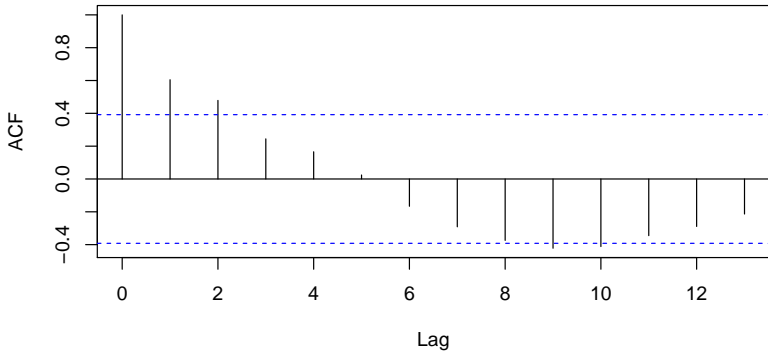
- La mayoría de las técnicas estadísticas asumen independencia de las observaciones

¿Cómo evaluar la independencia?

- Gráfico de la respuesta variable vs. tiempo
- Gráficos de autocorrelación (acf)



C. fuscicollis ACF



- Detectar cualquier patrón indica “dependencia”
- Acomodar la independencia: Modelos mixtos