Ejemplo de Modelo de efectos aleatorios

El archivo RIKZ.txt, parte del cual se muestra en la tabla 1, contiene datos recogidos por el *Netherlands Institute for Coastal and Marine Management/RIKZ* en verano de 2002, correspondientes a una muestra de 9 áreas intermareales (en realidad playas, por eso se denominan como *Beach* en el archivo) situadas a lo largo de la costa norte del pais. En cada área se fijaron 5 estaciones a distintas alturas sobre el nivel medio de marea (la variable *NAP* recoge el valor de la altura), que permitieron medir la macrofauna y varias variables abióticas. A partir de estos datos se calculó un índice de riqueza biológica (número de especies diferentes) en cada estación. Asimismo también para cada estación se calculó un índice de exposición que combinaba la acción del oleaje, la longitud de la zona de rompientes, la pendiente de la misma, el tamaño del grano y la profundidad de la capa anaeróbica. Este índice toma sólo dos valores, *alto* y *bajo.*

	Sample	Richness	Exposure	NAP	Beach	fBeach
1	1	11	low	0.04	1	1
2	2	10	low	-1.04	1	1
3	3	13	low	-1.34	1	1
4	4	11	low	0.62	1	1
5	5	10	low	-0.68	1	1
6	6	8	low	1.19	2	2
7	7	9	low	0.82	2	2
8	8	8	low	0.64	2	2
9	9	19	low	0.06	2	2
10	10	17	low	-1.33	2	2
11	11	6	high	-0.98	3	3
12	12	1	high	1.49	3	3

Tabla 1: Algunos datos del archivo RIKZ.txt.

La cuestión de fondo consiste en decidir si existe asociación entre la riqueza biológica, la exposición y la altura sobre el nivel medio de la marea. Un primer modelo para el estudio de esta relación es el siguiente:

$$R_{ij} = \beta_0 + \beta_1 \cdot NAP_{ij} + \beta_2 \cdot Exposure_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \approx N\left(0, \sigma_\varepsilon\right)$$

donde R_{ij} representa la riqueza biológica en la estación j de la playa i, NAP_{ij} el correspondiente valor de *NAP*, *Exposure*$_i$ la exposición en la playa i, y ε_{ij} el término residual no explicado.

Este modelo, en principio, es similar al modelo de análisis de la covarianza, por lo que podríamos comenzar realizando una exploración gráfica inicial. En primer lugar leemos los datos:

```
> RIKZ = read.table("http://dl.dropbox.com/u/7610774/Datos/zuur/RIKZ.txt",
     header = T)
> attach(RIKZ)
> Beach = factor(Beach)
```

(nótese la declaración Beach=factor(Beach). Ésta tiene como objetivo indicar a R que nos valores de Beach son simples etiquetas que en este caso sirven para identificar la playa, y que no tienen un valor intrínsecamente numérico).

y a continuación trazamos un gráfico identificando cada nivel de exposición con un símbolo distinto:

```
> require(car)
> scatterplot(Richness ~ NAP | Exposure, reg.line = lm, smooth = F)
```

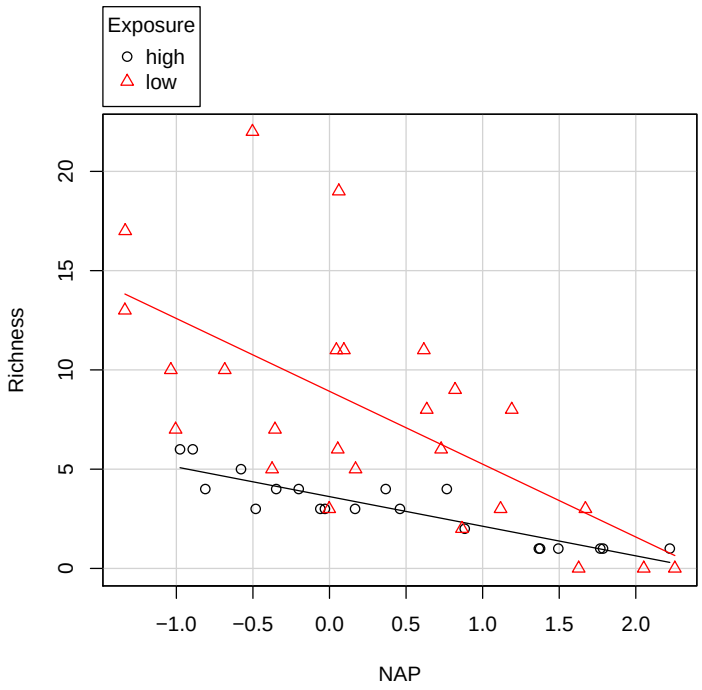


Figura 1: Relación Riqueza-NAP según el nivel de exposición.

En esta gráfica observamos que tanto cuando la exposición es alta como cuando es baja, la riqueza biológica disminuye a medida que aumenta la altura de la estación NAP con respecto al nivel medio de marea; asimismo el gráfico parece indicar que la tasa con que se produce dicha disminución es mayor a niveles de exposición bajos (pendiente negativa más acusada). Para determinar si la diferencia en pendientes es significativa podemos utilizar la estrategia que ya hemos visto para el análisis de la covarianza: ajustar y comparar entre sí los modelos:

1. $R_{ij} = \beta_0 + \beta_1 \cdot NAP_{ij} + \varepsilon_{ij}$ (No hay diferencias en la relación Riqueza-NAP según la exposición).
2. $R_{ij} = \beta_0 + \beta_1 \cdot NAP_{ij} + \beta_2 \cdot Exposure_i + \varepsilon_{ij}$ (La relación entre Riqueza y NAP es lineal, con la misma pendiente cualquiera que sea la exposición, y con ordenada dependiente del nivel de la misma).
3. $R_{ij} = \beta_0 + \beta_1 \cdot NAP_{ij} + \beta_2 \cdot Exposure_i + \beta_4 \cdot NAP_{ij} \cdot Exposure_i + \varepsilon_{ij}$ (La relación Riqueza-NAP es lineal y diferente tanto en pendiente como en ordenada en el origen según el nivel de exposición).

Para ello utilizamos la sintaxis:

```
> lm1 = lm(Richness ~ NAP, data = RIKZ)
> lm2 = lm(Richness ~ NAP + Exposure, data = RIKZ)
> lm3 = lm(Richness ~ NAP * Exposure, data = RIKZ)
> anova(lm1, lm2)

Analysis of Variance Table

Model 1: Richness ~ NAP
Model 2: Richness ~ NAP + Exposure
  Res.Df RSS Df Sum of Sq    F  Pr(>F)
1     43 744
2     42 518  1       226 18.3 0.00011 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> anova(lm2, lm3)

Analysis of Variance Table

Model 1: Richness ~ NAP + Exposure
Model 2: Richness ~ NAP * Exposure
  Res.Df RSS Df Sum of Sq    F Pr(>F)
1     42 518
2     41 468  1      50.4 4.41  0.042 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Al comparar el modelo 1 con el 2 vemos que la diferencia es significativa y el 2 explica mejor la relación entre variables (menor RSS). Asimismo, al comparar el 2 con el 3 nuevamente la diferencia es significativa y el tercer modelo presenta un mejor ajuste. Por tanto, podemos concluir que,

efectivamente, la diferencia en la inclinación de las pendientes es significativa. La estimación del modelo 3 produce el resultado:

```
> summary(lm3)

Call:
lm(formula = Richness ~ NAP * Exposure, data = RIKZ)

Residuals:
   Min     1Q Median     3Q    Max
 -5.93  -1.39  -0.30   0.93  11.23

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.617      0.822    4.40  7.5e-05 ***
NAP              -1.492      0.781   -1.91    0.063 .
Exposurelow       5.304      1.083    4.90  1.5e-05 ***
NAP:Exposurelow  -2.177      1.036   -2.10    0.042 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.38 on 41 degrees of freedom
Multiple R-squared: 0.575,        Adjusted R-squared: 0.544
F-statistic: 18.5 on 3 and 41 DF,  p-value: 9.6e-08
```

No obstante, para validar este modelo hemos de comprobar que los residuos siguen una distribución normal. El qqplot tiene el aspecto que se muestra en la figura 2, en el que se aprecia un desajuste con la normalidad. Este desajuste queda confirmado con el test de Shapiro-Wilk:

```
> shapiro.test(residuals(lm3))

        Shapiro-Wilk normality test

data:  residuals(lm3)
W = 0.8788, p-value = 0.0002206
```

Por tanto no es posible aceptar la normalidad en los residuos.Un vistazo a la figura 1 nos indica que probablemente la causa de la falta de normalidad es la dispersión irregular de los puntos en torno a la recta cuando el nivel de exposición es bajo. La pregunta natural es entonces ¿hay alguna otra variable que no hayamos tenido en cuenta que pudiera dar cuenta de esta dispersión tan irregular?

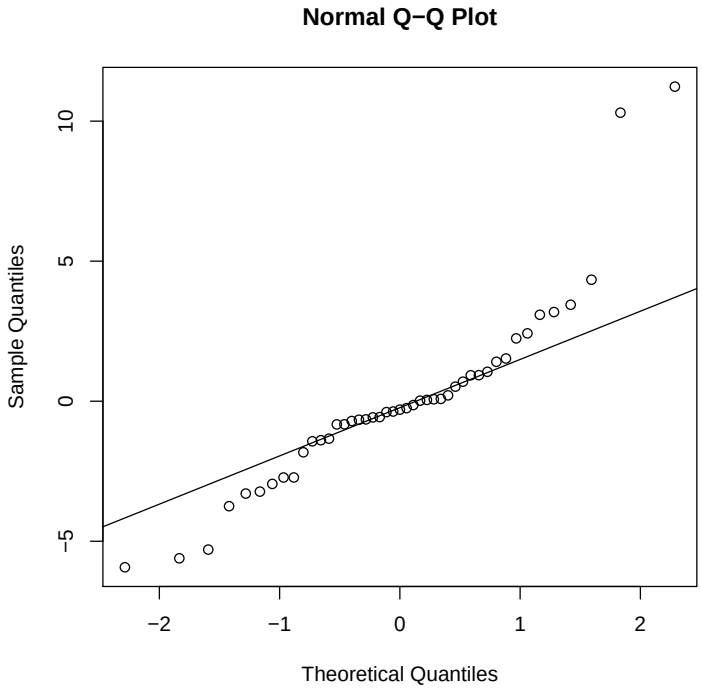


Figura 2: qqplot correspondiente a los residuos del modelo 3.

Una posible causa de tal dispersión podría ser el efecto de la playa. Si volvemos a representar la riqueza biológica frente a NAP, pero considerando la playa como posible factor de agrupamiento de las observaciones, obtenemos la gráfica que se muestra en la figura 3, obtenida mediante la sintaxis:

```
> require(car)
> scatterplot(Richness ~ NAP | Beach, reg.line = lm, smooth = F)
```

Otra forma de visualizar el posible efecto de la playa es utilizar el comando *xyplot*, del paquete *lattice* que produce el gráfico de la figura 4 mediante la sintaxis:

```
> require("lattice")
> xyplot(Richness ~ NAP | Beach, data = RIKZ, type = c("p", "r"))
```

Estas figuras nos indican que en la relación *Riqueza-NAP* existe también un efecto específico de la playa. El comando *lmList()* nos muestra la regresión *Riqueza-NAP* para cada playa:

```
> require(nlme)
> lmRNB = lmList(Richness ~ NAP | Beach, data = RIKZ)
> lmRNB
```

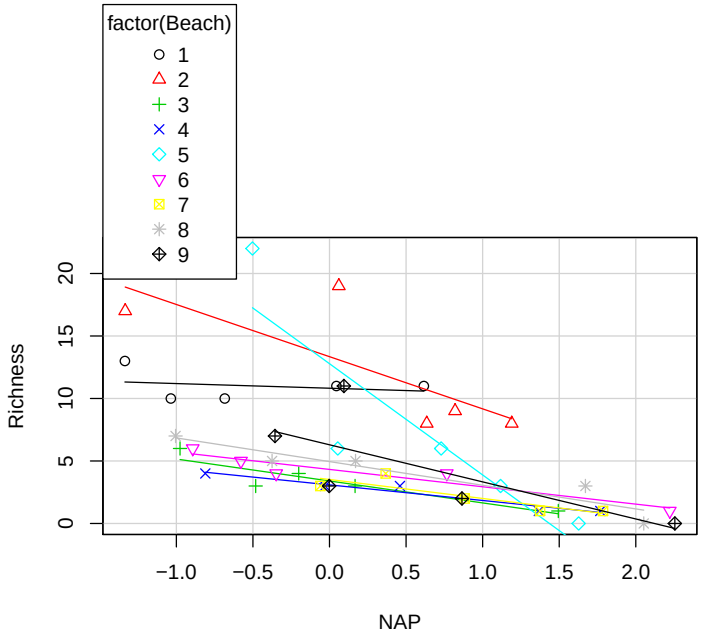


Figura 3: Relación Riqueza-NAP según playa.

```
Call:
  Model: Richness ~ NAP | Beach
   Data: RIKZ

Coefficients:
  (Intercept)     NAP
1      10.822 -0.3718
2      13.346 -4.1753
3       3.401 -1.7554
4       3.088 -1.2486
5      12.783 -8.9002
6       4.325 -1.3885
7       3.521 -1.5176
8       4.951 -1.8931
9       6.295 -2.9675

Degrees of freedom: 45 total; 27 residual
Residual standard error: 2.479
```

Ahora bien, tratar de ajustar un modelo de regresión a cada playa resulta costoso (es preciso estimar 3 parámetros por playa, a saber, los coeficientes de la regresión y el error estándar

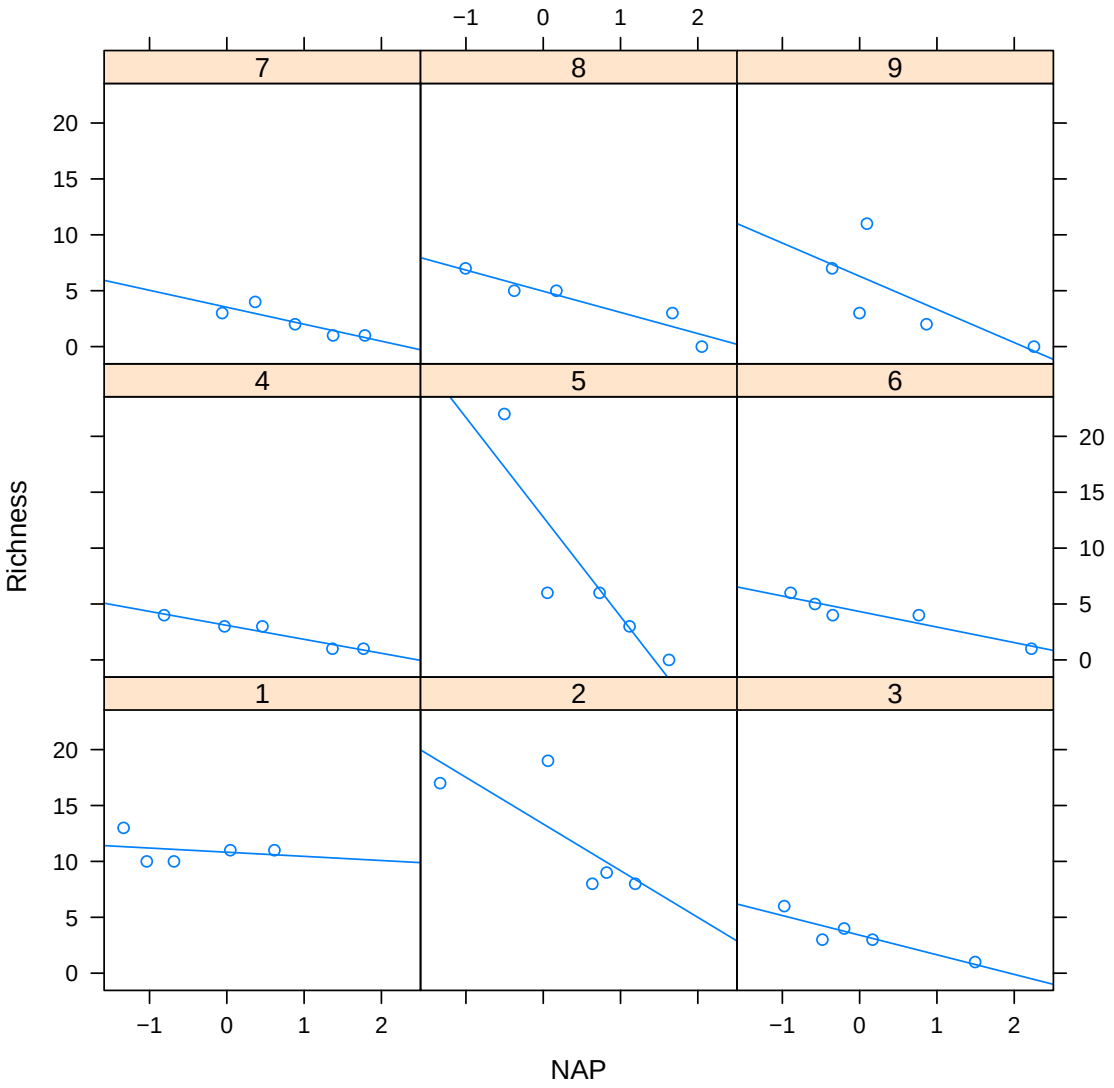


Figura 4: Relación Riqueza-NAP según playa.

de la misma; como hay 9 playas eso significa 27 parámetros). En realidad no hay más que 45 observaciones distintas, por lo que el número de datos resulta insuficiente para llevar a cabo una buena estimación. Además, las playas fueron elegidas al azar entre todas las de la costa norte de Holanda, por lo que en realidad no tiene demasiado interés determinar qué es lo que pasa particularmente en cada una de ellas.

Modelo de efectos aleatorios.

Si no estuviésemos interesados en el efecto de la altura de la estación NAP podríamos estudiar la variación de la riqueza biológica entre playas mediante un *modelo de efectos aleatorios:*

$$R_i = \beta_0 + b_i + \varepsilon_i$$

En este modelo se asume que R_i, la riqueza de la playa *i* es igual a una constante β_0, más el efecto b_i de la playa, más un término aleatorio ε_i que representa otros efectos de menor escala distintos de la playa. Los efectos b_i se suponen con distribución normal $N(0, \sigma_b)$, y el residuo ε_i se supone también $N(0, \sigma_\varepsilon)$; en otras palabras, no interesa tanto aquí estimar el efecto individual de cada playa sino la variabilidad en R_i que puede explicarse por el efecto de la playa. Este modelo puede estimarse del siguiente modo:

```
> require(nlme)
> RIKZ$fBeach = factor(RIKZ$Beach)
> modEfAl = lme(Richness ~ 1, random = ~1 | fBeach, data = RIKZ)
> summary(modEfAl)

Linear mixed-effects model fit by REML
 Data: RIKZ
    AIC   BIC logLik
  267.1 272.5 -130.6

Random effects:
 Formula: ~1 | fBeach
        (Intercept) Residual
StdDev:       3.237    3.938

Fixed effects: Richness ~ 1
            Value Std.Error DF t-value p-value
(Intercept) 5.689     1.228 36   4.631       0

Standardized Within-Group Residuals:
     Min       Q1      Med       Q3      Max
-1.77969 -0.50704 -0.09795  0.25469  3.80632

Number of Observations: 45
Number of Groups: 9
```

En este caso se tiene que $\hat{\beta}_0 = 5{,}689$, $\hat{\sigma}_b = 3{,}237$ y $\hat{\sigma}_\varepsilon = 3{,}938$. El p-valor mostrado corresponde a contrastar si $\beta_0 = 0$, contraste obviamente sin interés ya que $\hat{\beta}_0$ es un estimador de la riqueza media de todas las playas y este valor debe ser obviamente distinto de cero. Más interés tiene la estimación de intervalos de confianza para estos parámetros:

```
> intervals(modEfAl)
```

```
Approximate 95% confidence intervals

 Fixed effects:
            lower  est. upper
(Intercept) 3.198 5.689  8.18
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: fBeach
                lower  est. upper
sd((Intercept)) 1.709 3.237 6.131

 Within-group standard error:
lower  est. upper
3.126 3.938 4.962
```

Modelo de ordenada aleatoria.

Podemos introducir ahora el efecto de la altura de la estación, NAP, considerando la existencia de una asociación lineal entre esta variable y la riqueza biológica; ahora bien, en lugar de estimar una regresión distinta para cada playa como hicimos más arriba –lo que conduce a un exceso de parámetros y a dificultad de interpretación– podemos optar por modelar esta asociación de alguna de las dos formas siguientes:

- Considerando que la pendiente es la misma para todas las playas, y que la ordenada varía aleatoriamente de una playa a otra. Este modelo es similar al modelo de análisis de la covarianza, con la única diferencia de que en lugar de estimar una ordenada para cada regresión, nos limitaremos a estimar la variabilidad de las ordenadas. Formalmente:

  $$R_{ij} = \beta_{0i} + \beta_1 \cdot NAP_{ij} + \varepsilon_{ij}$$

  donde R_{ij} es la riqueza observada en la $j-$ésima estación de la playa i, $\beta_{0i} \approx N\left(\beta_0, \sigma_{\beta_0}\right)$ es la ordenada en la playa i y β_1 es la pendiente común a todas las playas.

- Considerando que tanto la ordenada como la pendiente varían aleatoriamente entre las playas. el modelo es entonces:

  $$R_{ij} = \beta_{0i} + \beta_{1i} \cdot NAP_{ij} + \varepsilon_{ij}$$

  donde $\beta_{0i} \approx N\left(\beta_0, \sigma_{\beta_0}\right)$ y $\beta_{1i} \approx N\left(\beta_1, \sigma_{\beta_1}\right)$.

Veamos la estimación del primer modelo, que podemos llamar *modelo de ordenada aleatoria:*

```
> modOrdAl = lme(Richness ~ NAP, random = ~1 | fBeach, data = RIKZ)
> summary(modOrdAl)

Linear mixed-effects model fit by REML
 Data: RIKZ
    AIC   BIC logLik
  247.5 254.5 -119.7

Random effects:
 Formula: ~1 | fBeach
        (Intercept) Residual
StdDev:       2.944     3.06

Fixed effects: Richness ~ NAP
             Value Std.Error DF t-value p-value
(Intercept)  6.582    1.0958 35   6.007       0
NAP         -2.568    0.4947 35  -5.192       0
 Correlation:
    (Intr)
NAP -0.157

Standardized Within-Group Residuals:
    Min      Q1     Med      Q3     Max
-1.4227 -0.4848 -0.1576  0.2519  3.9794

Number of Observations: 45
Number of Groups: 9
```

El resultado de esta estimación muestra que $\hat{\beta}_0 = 6{,}582$ y $\hat{\sigma}_{\beta_0} = 2{,}944$, por lo que la ordenada de la regresión Riqueza frente a NAP en las distintas playas β_{0i} sigue aproximadamente una distribución $N\left(6{,}582, 2{,}944\right)$. Asimismo, la pendiente (fija) es $\hat{\beta}_1 = -2{,}568$ y la desviación típica residual se estima como $\hat{\sigma}_\varepsilon = 3{,}06$. Podemos acompañar cada coeficiente de su correspondiente intervalo de confianza:

```
> intervals(modOrdAl)

Approximate 95% confidence intervals

 Fixed effects:
```

```
             lower   est.  upper
(Intercept)  4.357  6.582  8.806
NAP         -3.573 -2.568 -1.564
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: fBeach
                lower  est. upper
sd((Intercept)) 1.616 2.944 5.363

 Within-group standard error:
lower  est. upper
2.421 3.060 3.867
```

Modelo con pendiente y ordenada aleatorias.

La estimación del segundo modelo se llevaría a cabo del siguiente modo:

```
> modPendOrdAl = lme(Richness ~ NAP, random = ~1 + NAP | fBeach,
     data = RIKZ)
> summary(modPendOrdAl)

Linear mixed-effects model fit by REML
 Data: RIKZ
    AIC   BIC logLik
  244.4 255.0 -116.2

Random effects:
 Formula: ~1 + NAP | fBeach
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev Corr
(Intercept) 3.549  (Intr)
NAP         1.715  -0.99
Residual    2.703

Fixed effects: Richness ~ NAP
             Value Std.Error DF t-value p-value
(Intercept)  6.589    1.2648 35   5.209   0e+00
```

```
NAP         -2.830    0.7229 35  -3.915    4e-04
 Correlation:
    (Intr)
NAP -0.819

Standardized Within-Group Residuals:
    Min      Q1     Med      Q3     Max
-1.8213 -0.3411 -0.1675  0.1921  3.0397

Number of Observations: 45
Number of Groups: 9
```

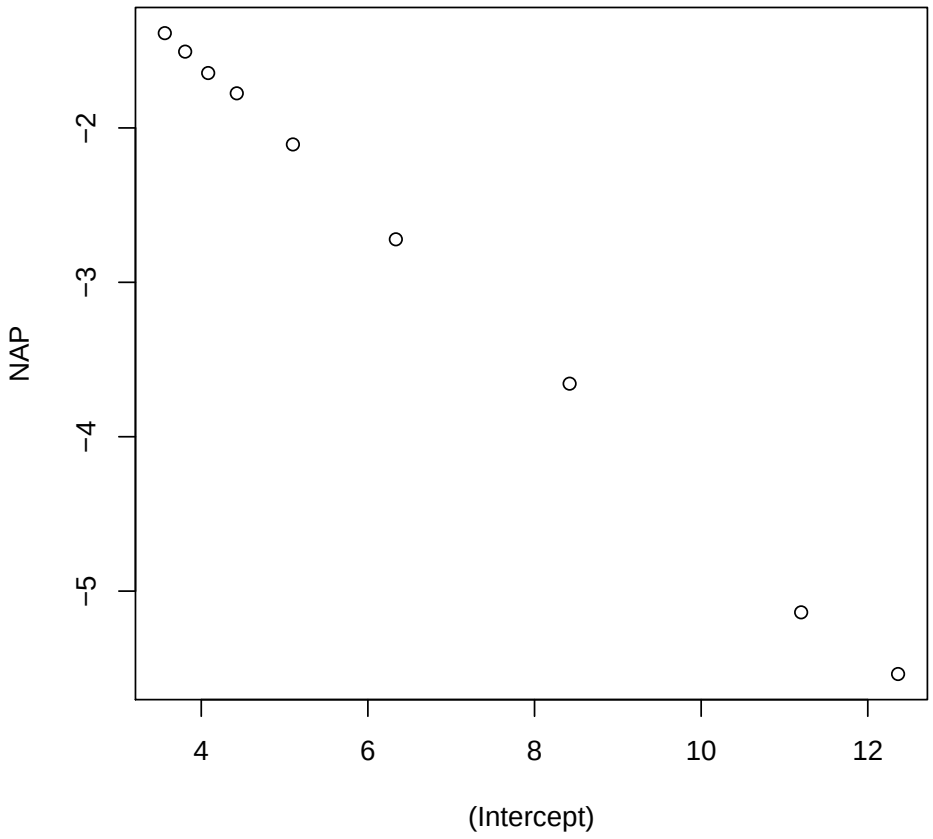
En este caso se tiene para las ordenadas que $\hat{\beta}_0 = 6{,}589$ y $\hat{\sigma}_{\beta_0} = 3{,}549$, por lo que $\beta_{0i} \approx N\left(6{,}589, 3{,}549\right)$; asimismo, para las pendientes resulta $\hat{\beta}_1 = -2{,}83$ con $\hat{\sigma}_{\beta_1} = 1{,}715$, por lo que $\beta_{1i} \approx N\left(-2{,}83, 1{,}715\right)$. La desviación típica residual es $\hat{\sigma}_\varepsilon = 2{,}703$.

Señalemos, por último, que hemos obtenido también la correlación entre las estimaciones de las pendientes y ordenadas en distintas playas; en este caso la correlación vale -0.99. Este valor significa que las pendientes y las ordenadas son muy dependientes entre sí, y que al aumentar la pendiente disminuye la ordenada (lo cual es esperable). Podemos ver estos coeficientes y una representación grafica de los mismos mediante:

```
> coef(modPendOrdAl)

  (Intercept)    NAP
1       8.421 -3.656
2      12.364 -5.537
3       3.807 -1.506
4       3.562 -1.386
5      11.200 -5.137
6       4.426 -1.776
7       4.083 -1.644
8       5.100 -2.107
9       6.335 -2.721

> plot(as.data.frame(coef(modPendOrdAl)))
```



Los correspondientes intervalos de confianza son:

```
> intervals(modPendOrdAl)

Approximate 95% confidence intervals

 Fixed effects:
             lower   est.  upper
(Intercept)  4.021  6.589  9.156
NAP         -4.298 -2.830 -1.362
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: fBeach
                        lower    est. upper
sd((Intercept))        1.9870  3.5491 6.339
sd(NAP)                0.6583  1.7150 4.467
cor((Intercept),NAP) -1.0000 -0.9902 1.000
```

```
 Within-group standard error:
lower  est. upper
1.947 2.703 3.752
```

Selección del mejor modelo.

Podemos plantearnos ahora cuál de los modelos anteriores representa mejor las relaciones entre las variables estudiadas. Recordemos además, que nos interesaba particularmente evaluar el efecto de la variable Exposure, por lo que debemos incluir también esta variable en nuestro análisis. A la hora de elegir entre diversos modelos alternativos existen dos estrategias básicas: la estrategia top-down consistente en comenzar el análisis con el modelo más completo posible e ir procediendo a simplificarlo paso a paso; y la estrategia step-up consistente en comenzar por un modelo simple e ir completándolo paso a paso. Ambas estrategias pueden conducir a resultados diferentes y en general se suele considerar más recomendable utilizar la primera. El protocolo para el ajuste de un modelo mediante la estrategia top-down es el siguiente:

1. Comenzar el modelo más completo posible que sólo contenga efectos fijos. En principio debemos tratar de incluir todas las variables explicativas y tantas interacciones como sea posible (si se dispone de pocos datos es usualmente imposible estimar muchas interacciones). Si no es posible incluir todas las variables explicativas, incluir aquellas que verosímilmente puedan contribuir a una mejor explicación o interpretación del modelo. Dado que este primer modelo sólo tiene parte fija, utilizaremos la función *gls()* que realiza la estimación utilizando el método de máxima verosimilitud restringida (REML), lo que facilita la comparación con otros modelos ajustados con *lme()*

2. Una vez fijada la parte fija del modelo, determinar la estructura óptima de la parte aleatoria. La clave para ello está en no trasladar a la componente aleatoria efectos que razonablemente deben ser fijos. La selección entre diversas formas de especificar la componente aleatoria puede realizarse buscando el menor valor de AIC, o mediante el test de razón de verosimilitudes (que R calcula mediante el comando *anova()* cuando se comparan modelos con distintas componentes aleatorias estimados mediante máxima verosimilitud restringida, REML, que es el método que utiliza por defecto la función *lme()*.

3. Una vez que se ha encontrado la mejor estructura para la componente aleatoria, determinar la mejor estructura para la componente fija. Si hay efectos fijos anidados, con la misma componente aleatoria, debe realizarse la comparación reestimando los modelos mediante máxima verosimilitud, lo que se consigue especificando la opción *method="ML"* en el comando *lme()*.

4. Por último, la estimación del modelo finalmente elegido debe realizarse mediante REML, ya que evita problemas de sesgo en la estimación de las varianzas.

Procedamos ahora a seleccionar el mejor modelo para explicar la riqueza biológica presente en los datos de nuestro ejemplo teniendo en cuenta el factor Exposure, la altura de la estación de muestreo NAP y las diferencias existentes entre playas. De acuerdo con el primer paso del protocolo anterior, debemos comenzar por el modelo más completo posible para los efectos fijos. Por ello elegimos un modelo que contenga como variables explicativas NAP y Exposure, así como sus interacciones:

```
> modelo1 = gls(Richness ~ 1 + NAP * Exposure, data = RIKZ)
> modelo1

Generalized least squares fit by REML
  Model: Richness ~ 1 + NAP * Exposure
  Data: RIKZ
  Log-restricted-likelihood: -114.3

Coefficients:
    (Intercept)             NAP     Exposurelow NAP:Exposurelow
          3.617          -1.492           5.304          -2.177

Degrees of freedom: 45 total; 41 residual
Residual standard error: 3.379
```

Ahora, siguiendo el segundo paso del protocolo, procedemos a buscar la estructura óptima para la componente aleatoria. En principio, en este caso cabe pensar en dos posibilidades: considerar como aleatoria la variación entre las ordenadas en cada playa, manteniendo fija la pendiente; o considerar que tanto las ordenadas como las pendientes se encuentran sujetas a efectos aleatorios. En cualquier caso, antes de estimar los modelos, recodificamos el factor Exposure de forma que la primera categoría sea “low” y la segunda “high”, lo que permite interpretar mejor el significado de los parámetros al tomar la categoría “low” como de referencia (de no hacerlo así, por defecto R tomaría como categoría de referencia aquella que va primero en orden alfabético, esto es, “high”)

```
> RIKZ$Exposure = factor(RIKZ$Exposure, levels = c("low", "high"))
> modelo2a = lme(Richness ~ 1 + NAP * Exposure, random = ~1 | fBeach,
     data = RIKZ)
> modelo2b = lme(Richness ~ 1 + NAP * Exposure, random = ~1 + NAP |
     fBeach, data = RIKZ)
```

Los valores de AIC para estos modelos pueden obtenerse mediante:

```
> AIC(modelo1, modelo2a, modelo2b)

         df   AIC
modelo1   5 238.5
modelo2a  6 236.5
modelo2b  8 237.1
```

Como vemos, el modelo de ordenada aleatoria (modelo2a) es el que presenta menor valor de AIC, por lo que constituye la opción preferida. Podemos, no obstante, comparar los modelos mediante el test de razón de verosimilitudes, y comprobar que no existen diferencias significativas entre el 2a y el 2b, por lo que resulta preferible el más simple (que es el de menor AIC), que a su vez es significativamente mejor que el modelo 1:

```
> anova(modelo1, modelo2a)

         Model df   AIC   BIC logLik   Test L.Ratio p-value
modelo1      1  5 238.5 247.1 -114.3
modelo2a     2  6 236.5 246.8 -112.2 1 vs 2    4.04  0.0444

> anova(modelo2a, modelo2b)

         Model df   AIC   BIC logLik   Test L.Ratio p-value
modelo2a     1  6 236.5 246.8 -112.2
modelo2b     2  8 237.1 250.8 -110.6 1 vs 2   3.359  0.1864
```

En el paso 3 del protocolo para el ajuste del mejor modelo, comprobamos ahora la mejor estructura para la componente fija del modelo. Para ello aplicamos la función *summary()* al modelo elegido en el paso anterior y evaluamos la significación de cada uno de los términos de la parte fija:

```
> summary(modelo2a)

Linear mixed-effects model fit by REML
 Data: RIKZ
    AIC   BIC logLik
  236.5 246.8 -112.2

Random effects:
 Formula: ~1 | fBeach
        (Intercept) Residual
```

```
StdDev:       1.819    2.943

Fixed effects: Richness ~ 1 + NAP * Exposure
                  Value Std.Error DF t-value p-value
(Intercept)       8.861    1.0208 34   8.680  0.0000
NAP              -3.464    0.6279 34  -5.517  0.0000
Exposurehigh     -5.256    1.5452  7  -3.401  0.0114
NAP:Exposurehigh  2.000    0.9461 34   2.114  0.0419
 Correlation:
                 (Intr) NAP    Expsrh
NAP              -0.181
Exposurehigh     -0.661  0.120
NAP:Exposurehigh  0.120 -0.664 -0.221

Standardized Within-Group Residuals:
     Min       Q1      Med       Q3      Max
-1.48493 -0.41609 -0.07704  0.15207  3.73132

Number of Observations: 45
Number of Groups: 9
```

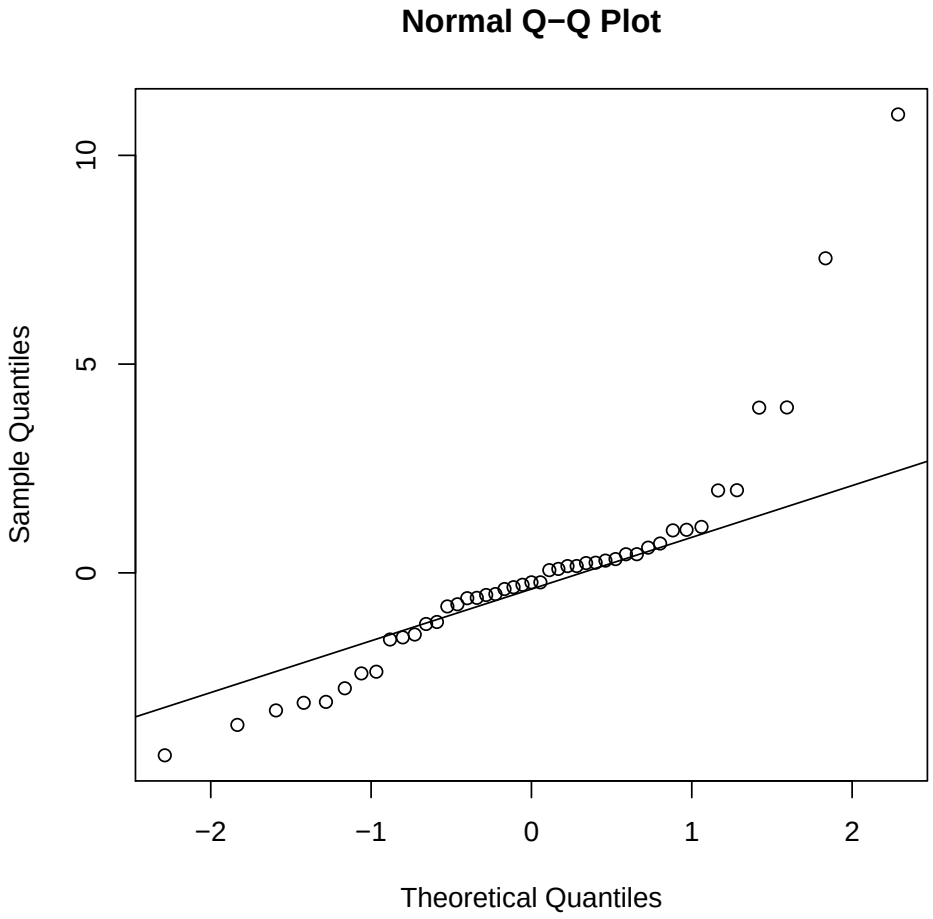
Como vemos, todos los términos del modelo son significativos; no obstante, el p-valor correspondiente a la interacción de NAP con Exposure es relativamente alto, por lo que podemos probar a simplificar el modelo, eliminando dicha interacción. Para comparar las componentes fijas debemos realizar la estimación por el método de máxima verosimilitud ML:

```
> modelo3 = lme(Richness ~ 1 + NAP + Exposure, random = ~1 | fBeach,
     data = RIKZ, method = "ML")
> modelo2a.ml = lme(Richness ~ 1 + NAP * Exposure, random = ~1 |
     fBeach, data = RIKZ, method = "ML")
> anova(modelo2a.ml, modelo3)

            Model df   AIC   BIC logLik   Test L.Ratio p-value
modelo2a.ml     1  6 242.1 252.9 -115.1
modelo3         2  5 244.8 253.8 -117.4 1 vs 2   4.645  0.0311
```

Como vemos, la diferencia entre ambos modelos es significativa, y el modelo 2a continúa teniendo menor valor de AIC. Así pues, acabamos el proceso eligiendo el modelo2a como el que produce mejor ajuste a los datos. Podemos verificar la normalidad de sus residuos mediante un qqplot:

```
> qqnorm(residuals(modelo2a))
> qqline(residuals(modelo2a))
```



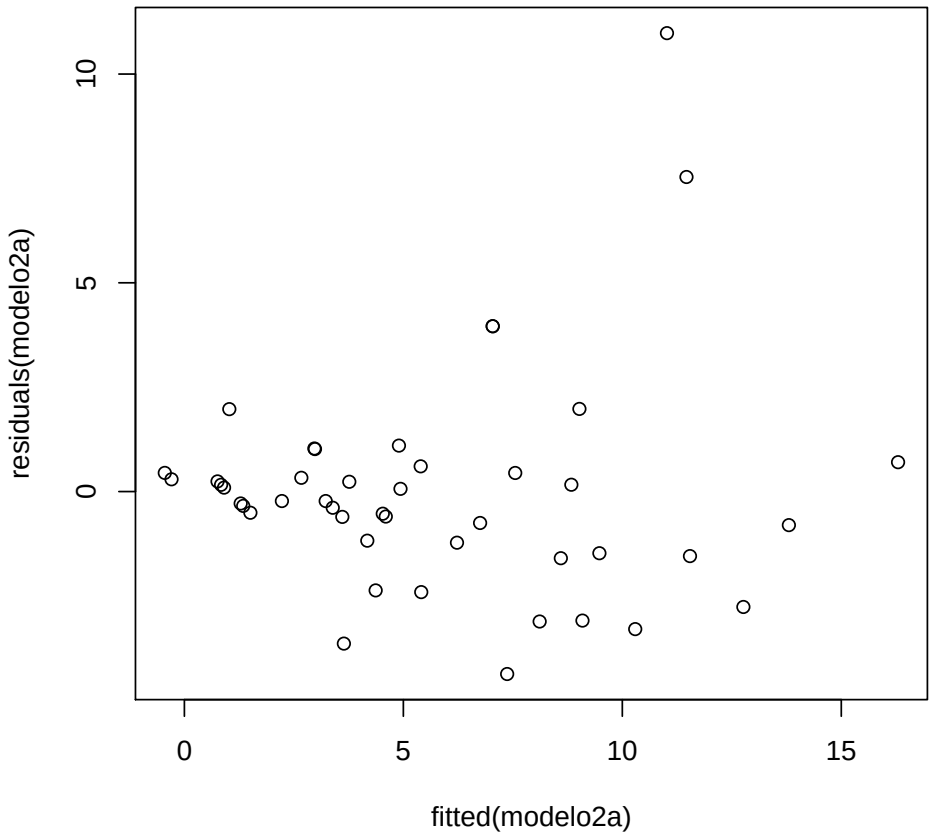
o utilizando el test de Shapiro-Wilk:

```
> shapiro.test(residuals(modelo2a))

        Shapiro-Wilk normality test

data:  residuals(modelo2a)
W = 0.8239, p-value = 8.329e-06
```

```
> plot(fitted(modelo2a), residuals(modelo2a))
```


residuals(modelo2a)
10
5
0
0
5
10
15
fitted(modelo2a)