Elaboración: Prof. Andrés García

email: andres.garcia.medina@uabc.edu.mx

Fuente: Wood, S. N. (2017). Generalized additive models: an introduction with R. CRC press.

Crecimiento de árboles: un ejemplo usando lme

El paquete `nlme` incluye un marco de datos llamado `Loblolly`, que contiene datos de crecimiento de los pinos Loblolly. La altura (`height`), en pies (los datos son de EE. UU.) y la edad, en años, se registran para 14 árboles individuales. Un factor variable Semilla (`seed`), con 14 niveles, indica la identidad de los árboles individuales. El interés radica en caracterizar la trayectoria de crecimiento medio a nivel poblacional de los pinos Loblolly, pero está claro que esperaríamos una gran variación de árbol a árbol, y probablemente también algún grado de autocorrelación en el componente aleatorio de la altura.

A partir del examen de los gráficos de datos, el siguiente modelo inicial podría ser apropiado para la i-ésima medicion en el j-ésimo árbol:

$$height_{ji} = \beta_0 + \beta_1 age_{ji} + \beta_2 age_{ji}^2 + \beta_3 age_{ji}^3 + b_0 + b_{j1} age_{ji} + b_{j2} age_{ji}^2 + b_{j3} age_{ji}^3 + \epsilon_{j,i}$$

donde ϵ_{ji} son variables aleatorias normales de media cero con correlacion del tipo $\rho(\epsilon_{j,i}, \epsilon_{j,i-k}) = \phi^k$, y ϕ es un parámetro desconocido.

El término ϵ es independiente para diferentes arboles.

Como es usual, β denota los efectos fijos y $b \sim N(0, \Psi)$ denota los efectos aleatorios.

Este modelo se puede estimar utilizando lme, pero para evitar dificultades de convergencia en el siguiente análisis, son útiles dos pasos preparatorios. En primer lugar, conviene centrar la variable edad de la siguiente manera:

In [11]:
```
library(nlme)
data(Loblolly)
Loblolly[1:5,]
```

A nfnGroupedData: 5 × 3

	height	age	Seed
	<dbl>	<dbl>	<ord>
1	4.51	3	301
15	10.89	5	301
29	28.72	10	301
43	41.74	15	301
57	52.70	20	301

In [12]:
```
Loblolly$age <- Loblolly$age - mean(Loblolly$age)
```

En segundo lugar, para este análisis el método de ajuste predeterminado falla sin algún ajuste. Algunos ajustes comienzan utilizando el algoritmo EM para acercarse razonablemente a las estimaciones óptimas de los parámetros y luego cambian al método de Newton, que converge más rápidamente. El número de pasos EM a dar y el número máximo de pasos Newton a permitir son controlables mediante el argumento de control de lme. La función lmeControl ofrece una manera conveniente de producir una lista de control, con algunos elementos modificados respecto de sus valores predeterminados. Por ejemplo

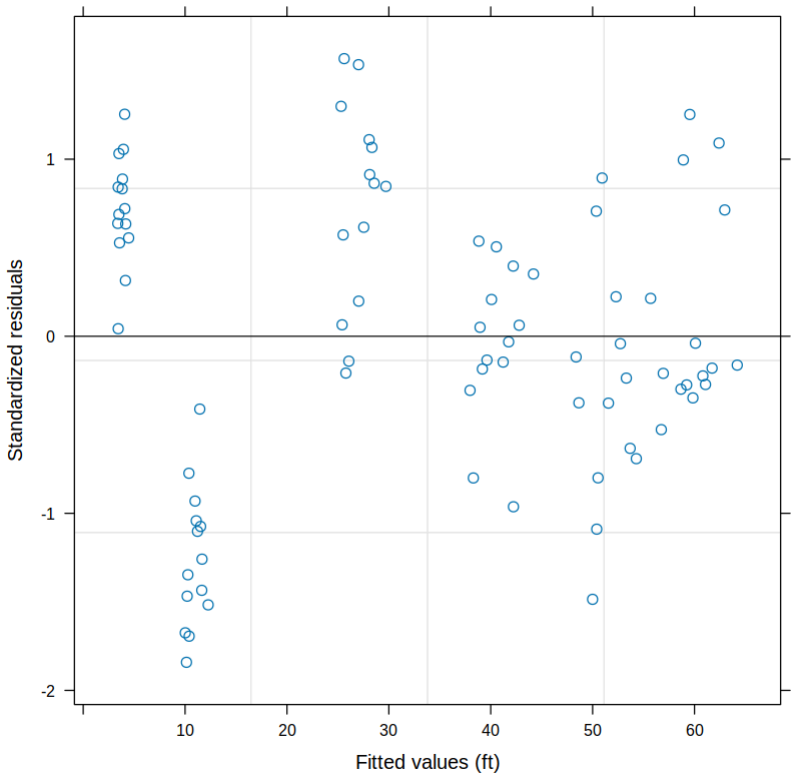
In [13]:
```
lmc <- lmeControl(niterEM=500,msMaxIter=100)
```

produce una lista de control en la que el número de iteraciones EM se establece en 500 y el número máximo de iteraciones Newton se establece en 100.

In [14]:
```
m0 <- lme(height ~ age + I(age^2) + I(age^3),Loblolly,
random = list(Seed = ~ age + I(age^2) + I(age^3)),
correlation = corAR1(form = ~ age|Seed),control=lmc)
```

El argumento aleatorio (`random`) especifica que debería haber un término cúbico diferente para cada árbol, mientras que el argumento de correlación (`correlation`) especifica un modelo autorregresivo para los residuos de cada árbol. `form = age|Seed` indica que la edad es la variable que determina el orden de los residuos y que la correlación se aplica dentro de las mediciones realizadas en un árbol, pero no entre mediciones en diferentes árboles.

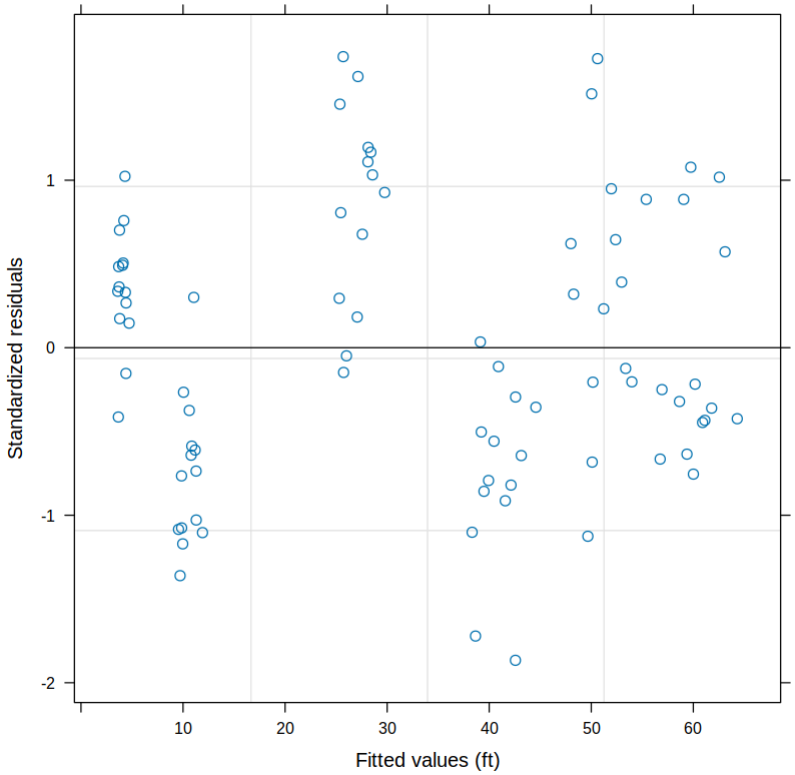
In [15]:
```
plot(m0)
```



El gráfico muestra una clara tendencia en la media de los residuos: el modelo parece subestimar el primer grupo de mediciones, realizado a los 5 años, y luego sobreestimar el siguiente grupo, realizado a los 10 años, antes de subestimar algo el siguiente grupo, que corresponde hasta el año 15. Esto sugiere la necesidad de un modelo más flexible, por lo que también se probaron polinomios de cuarto y quinto orden.

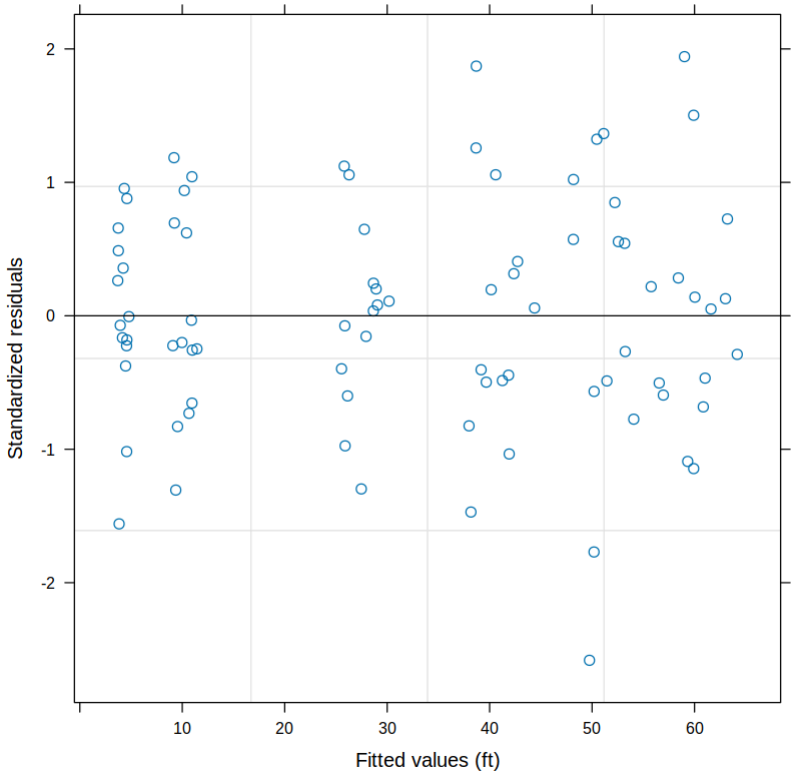
In [16]:
```
m1 <- lme(height ~ age + I(age^2) + I(age^3) + I(age^4),
Loblolly, list(Seed = ~ age + I(age^2) + I(age^3)),
cor = corAR1(form = ~age|Seed), control=lmc)
```

In [17]:
```
plot(m1)
```



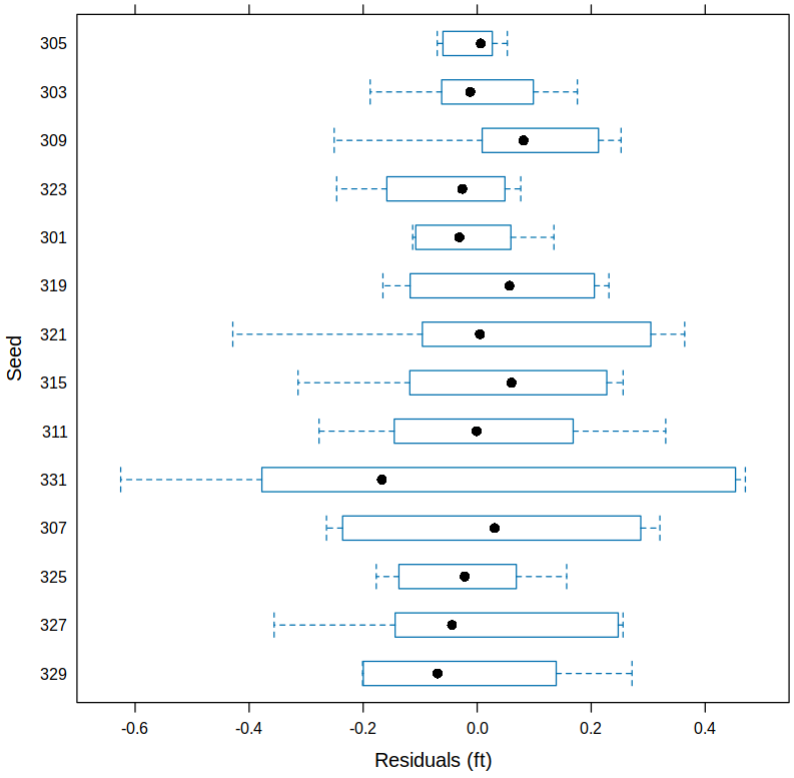
In [19]:
```
m2 <- lme(height ~ age +I(age^2) +I(age^3) +I(age^4) +I(age^5),
Loblolly,list(Seed = ~ age + I(age^2) + I(age^3)),
cor = corAR1(form = ~ age|Seed), control=lmc)
```

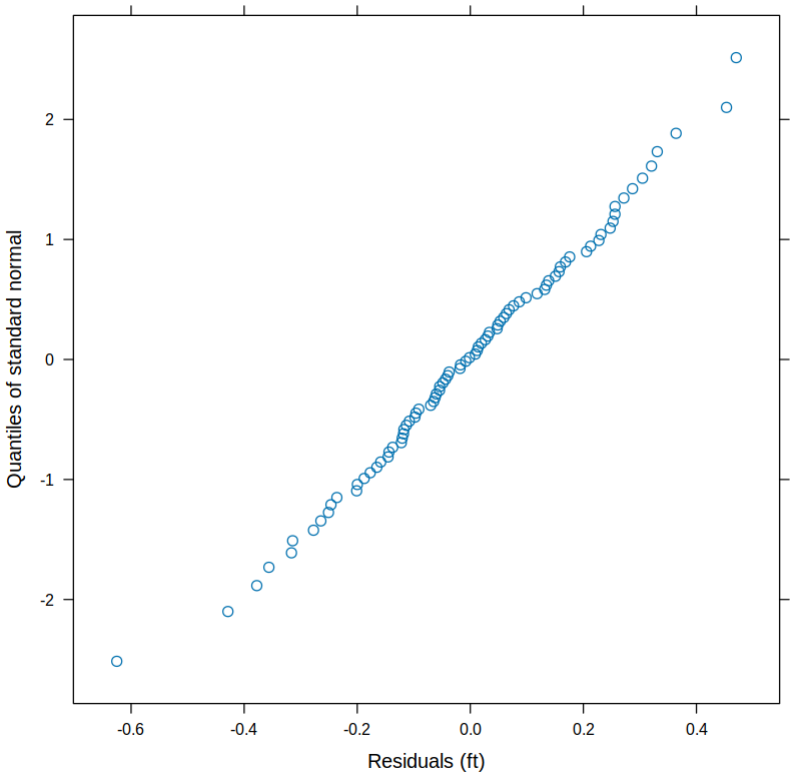
In [20]:
```
plot(m2)
```

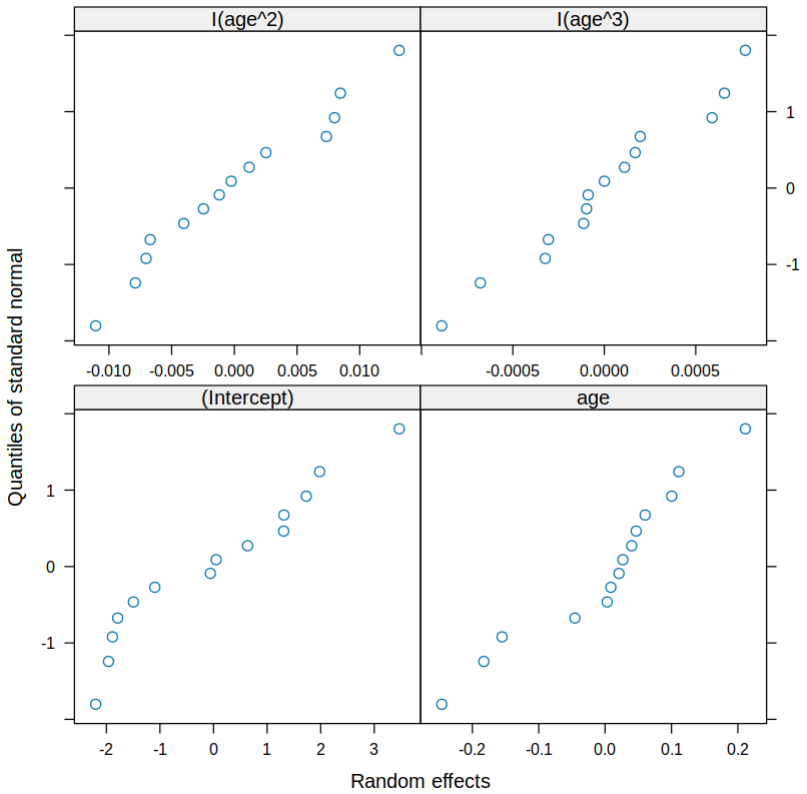


Los gráficos de residuos resultantes se muestran una ligera mejora para m1, pero sólo m2 es realmente satisfactorio. Ahora se pueden crear más gráficos de control para m2.

In [21]:
```
plot(m2,Seed~resid(.))
qqnorm(m2,~resid(.))
qqnorm(m2,~ranef(.))
```


Seed
305
303
309
323
301
319
321
315
311
331
307
325
327
329
-0.6
-0.4
-0.2
0.0
0.2
0.4
Residuals (ft)


Quantiles of standard normal
2
1
0
-1
-2
-0.6
-0.4
-0.2
0.0
0.2
0.4
Residuals (ft)



Una pregunta obvia es si realmente se requiere la elaborada estructura del modelo, con errores cúbicos aleatorios y autocorrelacionados dentro del árbol. Primero intente eliminar el componente de autocorrelación.

In [24]:
```
m3 <- lme(height ~ age+I(age^2)+I(age^3)+I(age^4)+I(age^5),Loblolly,list(See
```

In [25]:
```
anova(m3,m2)
```

A anova.lme: 2 × 9

	call	Model	df	AIC	BIC	logLik	Test	L.Ratio	p
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<fct>	<dbl>	
m3	lme.formula(fixed = height ~ age + I(age^2) + I(age^3) + I(age^4) + I(age^5), data = Loblolly, random = list(Seed = ~age + I(age^2) + I(age^3)), control = lmc)	1	17	250.4616	290.5257	-108.2308		NA	
m2	lme.formula(fixed = height ~ age + I(age^2) + I(age^3) + I(age^4) + I(age^5), data = Loblolly, random = list(Seed = ~age + I(age^2) + I(age^3)), correlation = corAR1(form = ~age | Seed), control = lmc)	2	18	239.3575	281.7783	-101.6788	1 vs 2	13.10408	0.000

El comando anova en realidad está realizando aquí una prueba de razón de verosimilitud generalizada (GLRT), que rechaza m3 a favor de m2. anova también informa el AIC para los modelos, lo que también sugiere que es preferible m2. Parece haber pruebas sólidas de autocorrelación en los residuos dentro de los árboles. Quizás el modelo de efectos aleatorios podría simplificarse eliminando la dependencia del crecimiento específico de los árboles del cubo de la edad.

In [26]:
```
m4 <- lme(height~age+I(age^2)+I(age^3)+I(age^4)+I(age^5),Loblolly,list(Seed=
```

In [27]:
```
anova(m4,m2)
```

A anova.lme: 2 × 9

	call	Model	df	AIC	BIC	logLik	Test	L.Ratio	p
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<fct>	<dbl>	
m4	lme.formula(fixed = height ~ age + I(age^2) + I(age^3) + I(age^4) + I(age^5), data = Loblolly, random = list(Seed = ~age + I(age^2)), correlation = corAR1(form = ~age | Seed), control = lmc)	1	14	253.7579	286.7519	-112.8790		NA	
m2	lme.formula(fixed = height ~ age + I(age^2) + I(age^3) + I(age^4) + I(age^5), data = Loblolly, random = list(Seed = ~age + I(age^2) + I(age^3)), correlation = corAR1(form = ~age | Seed), control = lmc)	2	18	239.3575	281.7783	-101.6788	1 vs 2	22.40041	0.000

Comparación de puntuaciones AIC (que también podrían haberse obtenido utilizando AIC(m4,m2)) sugiere enfáticamente que m2 es el mejor modelo.

Otro modelo obvio que se puede probar es uno con una estructura de efectos aleatorios menos general. Los modelos hasta ahora han permitido correlacionar los efectos aleatorios de cualquier árbol de una forma general muy restrictiva. Simplemente se ha asumido que $b_j \sim N(0, \Psi_\theta)$, donde la única restricción en la matriz Ψ_θ es que debe ser positiva definida. Posiblemente un modelo menos flexible podría ser suficiente. Por ejemplo, Ψ_θ diagonal con elementos positivos. Esta estructura se puede definir con ayuda de la función `lme` como:

In [28]:
```
m5 <- lme(height~age+I(age^2)+I(age^3)+I(age^4)+I(age^5),Loblolly,list(Seed=
```

Aquí la función `pdDiag` indica que la matriz de covarianza para los efectos aleatorios en cada nivel de `Seed` debe tener una estructura diagonal (definida positiva). `m5` se puede comparar con `m2`.

In [29]:
```
anova(m2,m5)
```

A anova.lme: 2 × 9

	call	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-v
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<fct>	<dbl>	
m2	lme.formula(fixed = height ~ age + I(age^2) + I(age^3) + I(age^4) + I(age^5), data = Loblolly, random = list(Seed = ~age + I(age^2) + I(age^3)), correlation = corAR1(form = ~age | Seed), control = lmc)	1	18	239.3575	281.7783	-101.6788		NA	
m5	lme.formula(fixed = height ~ age + I(age^2) + I(age^3) + I(age^4) + I(age^5), data = Loblolly, random = list(Seed = pdDiag(~age + I(age^2) + I(age^3))), correlation = corAR1(form = ~age | Seed), control = lmc)	2	12	293.7080	321.9886	-134.8540	1 vs 2	66.35051	2.285

Nuevamente, tanto la prueba GLRT como la comparación AIC favorecen el modelo más general m2. En este caso se cumplen los supuestos GLRT: m5 equivale a establecer las covarianzas de los efectos aleatorios en m2 en cero, pero como las covarianzas pueden ser positivas o negativas, esto no está en el límite del espacio de parámetros y los supuestos GLRT se cumplen.