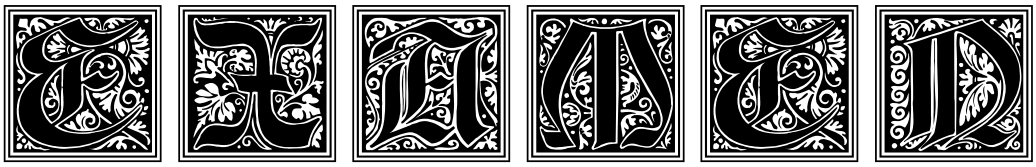
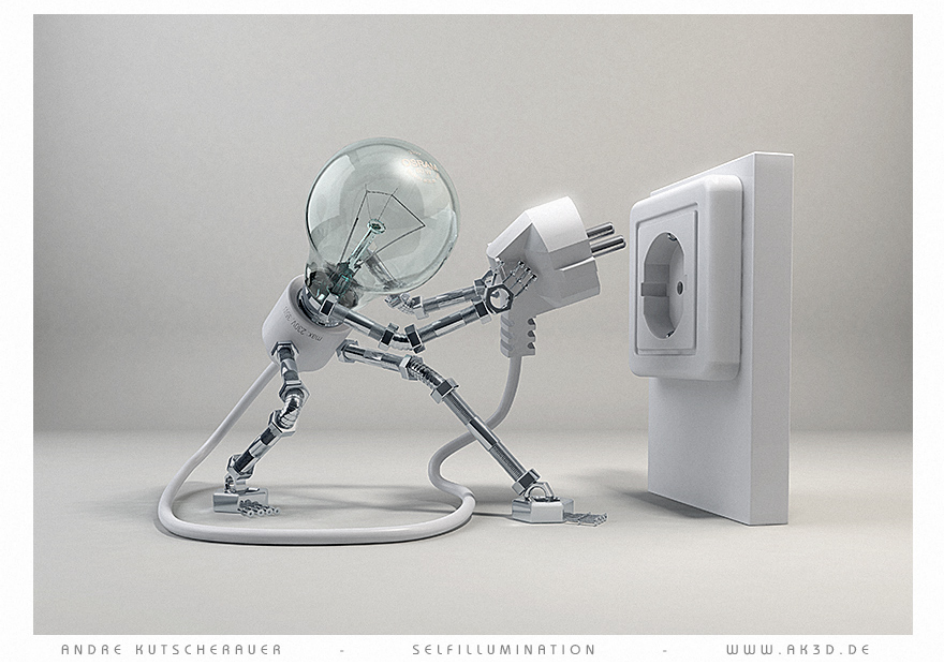
Modelos lineales generales, generalizados y mixtos en ecología:



Roberto Munguía-Steyer
rmunguia.steyer@gmail.com
Departamento de Ecología Evolutiva
Instituto de Ecología, UNAM
04 Octubre-06 Diciembre de 2011



En pos de la iluminación.

Estudiante: _____

1. Preguntas: (2.5 puntos)

- Describe a que modelos pertenecen las siguientes ecuaciones, en las cuales hay un número de observaciones (i), algunas veces pertenecientes a distintos grupos (j):

$$y_i = \alpha + \beta_1 x_i + \epsilon_i,\ \epsilon_i \sim N(0,\, \sigma^2) \tag{1}$$

- Modelo 1: Modelo lineal general, regresión lineal.

$$y_i = \alpha_{j(i)} + \beta_{j(i)} * x_i + \epsilon_i,\ \epsilon_i \sim N(0,\, \sigma^2) \tag{2}$$

- Modelo 2: Modelo lineal general, analisis de covarianza con pend. distintas.

$$y_i = \alpha_{j(i)} * x_i + \epsilon_{j(i)},\ \epsilon_{j(i)} \sim N(0,\, {\sigma_j}^2) \tag{3}$$

- Modelo 3: Modelo de mínimos cuadrados generalizados, het. de varianza por grupo.

$$logit(p_i) = \alpha + \beta_1 x_i,\ y_i \sim Bin(N,\, p) \tag{4}$$

- Modelo 4: Modelo lineal generalizado con distribución binomial.

$$log(\lambda) = \alpha + \beta_1 x_i,\ y_i \sim Pois(\lambda) \tag{5}$$

- Modelo 5: Modelo lineal generalizado con distribución Poisson.

2. Ejercicios

Instrucciones:

Para acceder a las bases de datos de los siguientes ejercicios, coloca el archivo `bdatos.RData` en la carpeta donde tengas tu directorio de trabajo en R. Si desconoces donde se encuentra este directorio, puedes emplear el comando `getwd()` para averiguarlo. Para cargar las bases de datos simplemente tienes que utilizar el siguiente comando: `load("bdatos.RData")`. Cerciórate (`ls()`) que las bases de datos de los ejercicios esten accesibles y procede con el examen.

Recuerda poner a prueba los supuestos de los respectivos modelos, con inspecciones visuales preeliminares, el ajuste del modelo y su posterior validación. En caso de que no se cumplan los supuestos, modifica la parte sistemática o aleatoria del modelo. Asimismo, considera realizar selección de modelos con el criterio de información de Akaike (AIC) para encontrar aquel modelo que presente un buen balance entre el número de parámetros que contiene y su grado de ajuste.

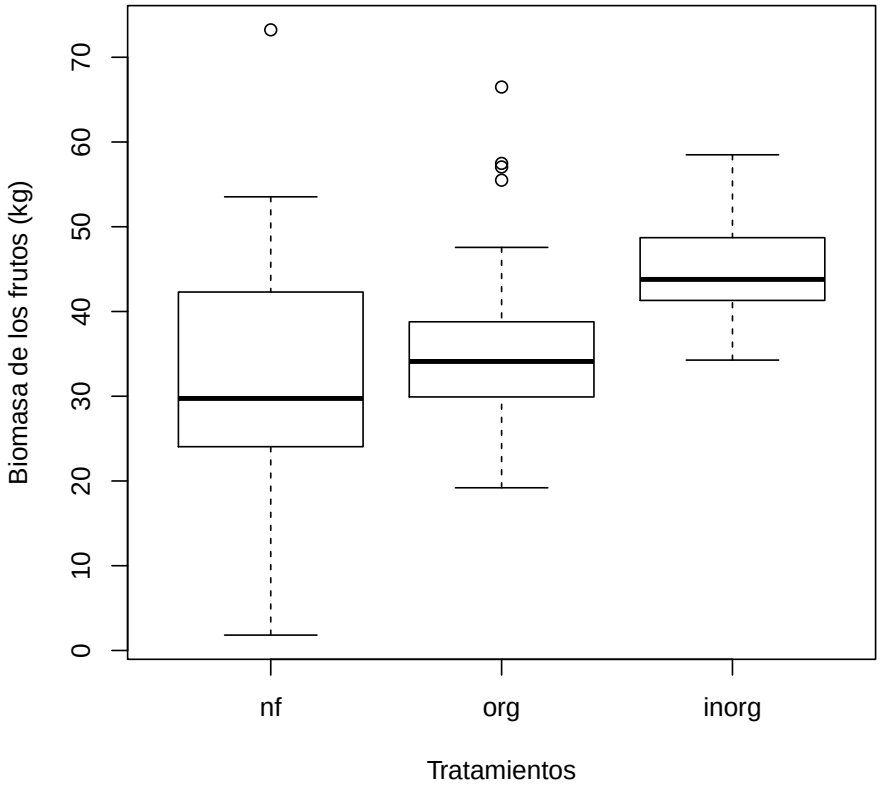
Contesta las preguntas de los ejercicios sobre el papel, pero guarda todos los objetos que generaste en R en un archivo con tu nombre para en caso de duda ver como procediste a resolver los ejercicios. Ejemplo: si yo estuviera realizando el examen, una vez completados los ejercicios, aplicaría el comando `save.image("Roberto.RData")`.

2.1. Ejercicio 1: (2.5 puntos)

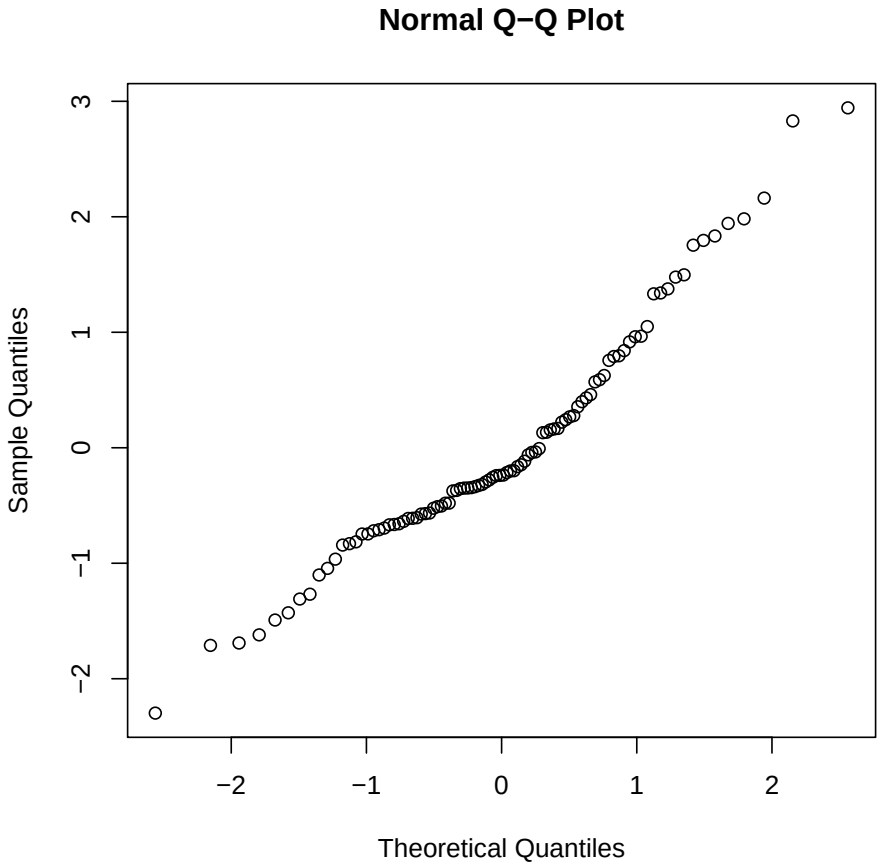
Se realizó un estudio para determinar si la producción de durazno puede ser beneficiada con la utilización de fertilizantes orgánicos utilizando la lombricomposta como abono. La producción de durazno está medida como la biomasa en kg por árbol frutal. Se comparó la biomasa producida en árboles frutales pertenecientes a tres tratamientos: a) sin fertilizante, b) con fertilizante orgánico, c) con el fertilizante inorgánico tradicional. Averigua si existen diferencias entre los tratamientos y de ser así, especifica cuales difieren entre si. La base de datos se llama `duraznos`.

```
> library(nlme)
> library(bbmle)
> #Generando la base de datos: duraznos
> set.seed(1234)
> nobs <- 96
> ntrat <- nobs/3
> fert <- gl(3, ntrat, labels=c("nf", "org", "inorg"))
> pnf <- rnorm(ntrat, 37, sd= 15)
> porg <- rnorm(ntrat, 41, sd=10)
> pinorg <- rnorm(ntrat, 44, sd= 7)
> prod <- c(pnf, porg, pinorg)
> duraznos <- data.frame(prod, fert)
> boxplot(prod~fert, data=duraznos, ylab= "Biomasa de los frutos (kg)",
```

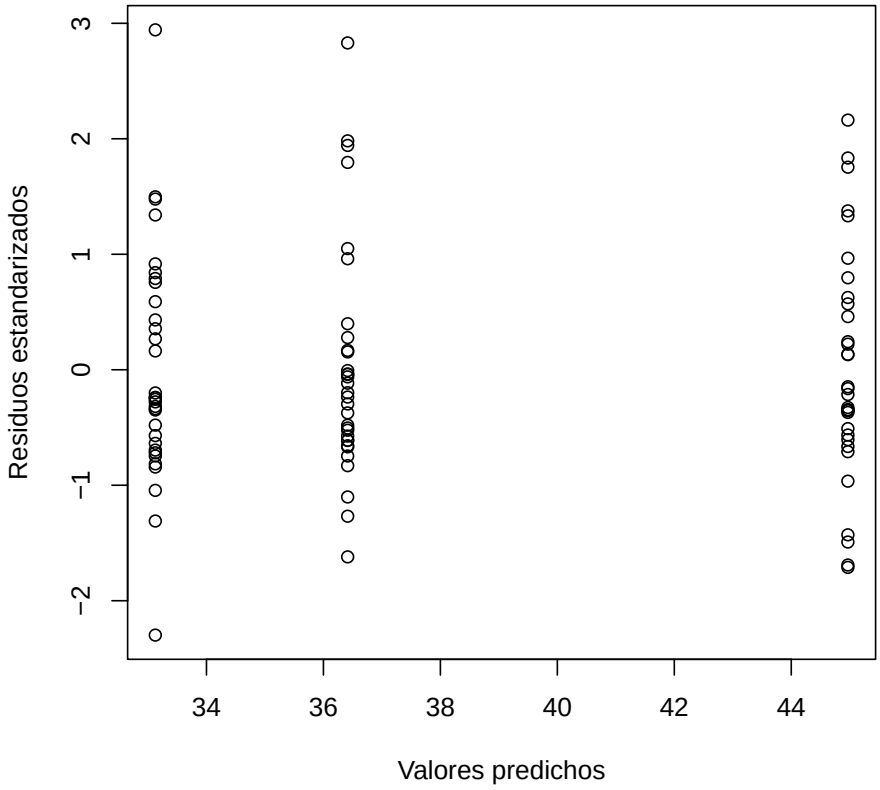
```
+ xlab= "Tratamientos")
> m1 <- lm(prod~fert, data=duraznos)
> anova(m1)

Analysis of Variance Table

Response: prod
          Df Sum Sq Mean Sq F value Pr(>F)
fert       2   2392    1196    10.6  7e-05
Residuals 93  10471     113

> par(mfrow=c(2,2))
> plot(m1)
> m1.gls <- gls(prod~fert, data=duraznos)
> m2.gls <- gls(prod~fert, weights=varIdent(form=~1|fert), data=duraznos)
> AICtab(m1.gls,m2.gls)

       dAIC df
m2.gls  0.0 6
m1.gls 13.2 4

> anova(m2.gls)

Denom. DF: 93
            numDF F-value p-value
(Intercept)     1 2189.25  <.0001
fert            2   14.62  <.0001
```




```
> # Validacion del modelo gls
> residm <- resid(m2.gls, type="normalized")
> qqnorm(residm)
```




```
> plot(fitted(m2.gls), residm, xlab="Valores predichos",
+ ylab= "Residuos estandarizados")
```




```
> #cuales grupos difieren entre si
> summary(m2.gls)

Generalized least squares fit by REML
  Model: prod ~ fert
  Data: duraznos
     AIC    BIC  logLik
  708.47 723.67 -348.24

Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | fert
 Parameter estimates:
     nf     org   inorg
1.00000 0.77977 0.45902

Coefficients:
             Value Std.Error t-value p-value
(Intercept) 33.125    2.4091 13.7501  0.0000
fertorg      3.289    3.0549  1.0766  0.2844
fertinorg   11.843    2.6508  4.4678  0.0000
```

```
 Correlation:
          (Intr) fertrg
fertorg   -0.789
fertinorg -0.909  0.717

Standardized residuals:
     Min       Q1      Med       Q3      Max
-2.29755 -0.61087 -0.23762  0.48817  2.94339

Residual standard error: 13.628
Degrees of freedom: 96 total; 93 residual
```

```
> m3.gls <- gls(prod~fert-1, weights=varIdent(form=~1|fert))
> confint(m3.gls)

           2.5 % 97.5 %
fertnf    28.404 37.847
fertorg   32.732 40.096
fertinorg 42.801 47.136
```

1. ¿Se cumplieron los supuestos de los modelos lineales ?

   No se cumple el supuesto de homogeneidad de varianza

2. ¿Cuál es el modelo estadístico final?

   Es un modelo de cuadrados mínimos generalizados (GLS), con la estimación de varianzas específicas para cada grupo.

3. ¿Difieren los tratamientos? Reporta la prueba de F.

   Si difieren. F= 14.62, df=2,93, p < 0.001.

4. Si existe un efecto del factor, menciona cuáles son los niveles que difieren entre sí.

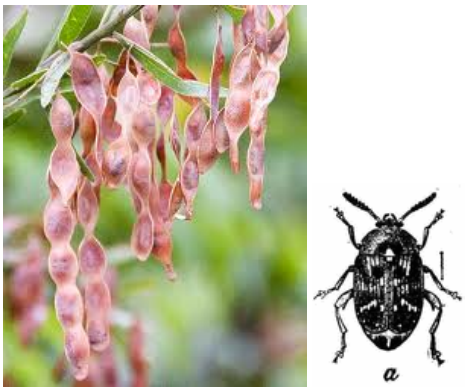
   El fertilizante inorgánico difiere de los dos niveles restantes. El fertilizante orgánico no difiere de la no aplicación del fertilizante.

5. Especifica los valores predichos por tratamiento, con los intervalos de confianza al 95 %.

   Ver el código que antecede a estas preguntas.

2.2. Ejercicio 2: (2.5 puntos)

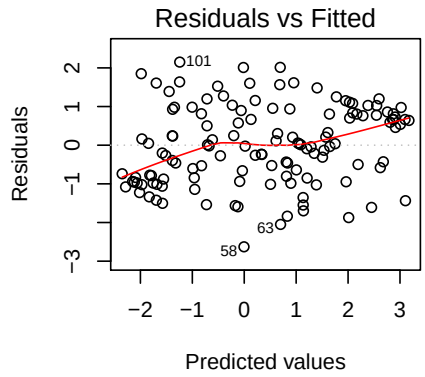
Se realizó un estudio sobre la depredación de semillas por parte de escarabajos de la familia Bruchidae. Se colectaron 130 vainas provenientes de plantas pertenecientes al genero *Acacia*. Se colectó **una** vaina por planta y se determinó cuantas semillas se encontraban intactas y cuantas depredadas. Asimismo, se midió el grosor de cada una de las vainas colectadas. La intención es investigar si existe una relación entre el grosor que cada vaina presenta y la probabilidad de que las semillas sufran depredación. La base de datos se llama `acacia`.

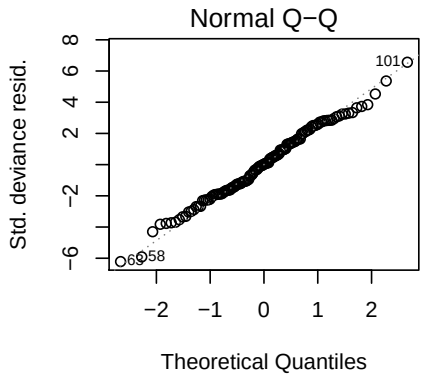


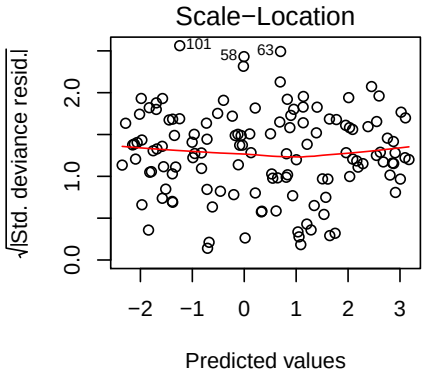

```
> # Vainas con un numero de semillas, variables
> set.seed(3333)
> lambd <- 5
> # numero de vainas colectadas, una por árbol
> nv <- 130
> # numero de semillas por vaina
> nsem <- rpois(nv,5)
> table(nsem)

nsem
 1  2  3  4  5  6  7  8  9 10 12
 1 12 17 26 21 18 13 10  7  4  1

> # Grosor de la vaina
> gv <- runif(nv,0.1, 0.8)
> # valores de los parametros
> a <- 4
> b <- -8
> logistica <- function(x) { exp(x)/(1 + exp(x)) }
> y_det <- logistica(a+b*gv)
> # plot(gv,y_det)
> # Semillas depredadas
> nsemdep <- rbinom(nv, size=nsem,prob=y_det)
> # Base de datos
> acacia <- data.frame(gv, nsem, nsemdep)
> # Ajuste del modelo
> m1 <- glm(cbind(nsemdep, nsem-nsemdep)~gv,binomial)
> summary(m1)
```

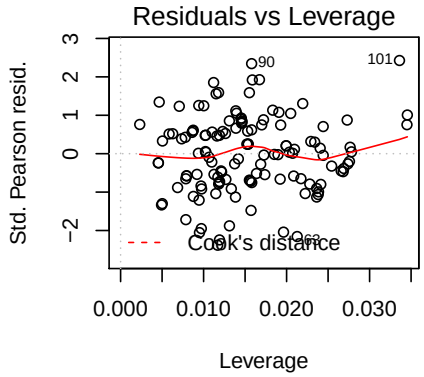
```
Call:
glm(formula = cbind(nsemdep, nsem - nsemdep) ~ gv, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6283  -0.8411   0.0115   0.8134   2.1505

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.987      0.303    13.2   <2e-16
gv            -7.927      0.603   -13.1   <2e-16

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 398.14  on 129  degrees of freedom
Residual deviance: 134.56  on 128  degrees of freedom
AIC: 298.3

Number of Fisher Scoring iterations: 4
```

```
> # Validacion del modelo
> par(mfrow=c(2,2))
> plot(m1)
```



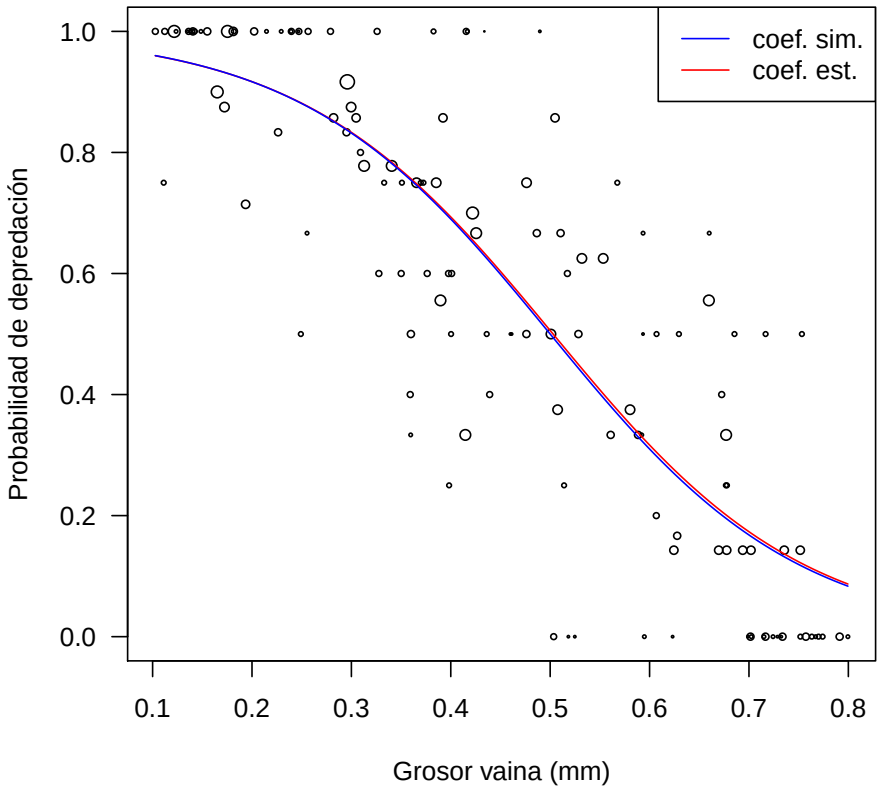







```
> coefi <- coef(m1)
> a.est <- coefi[1]
> b.est <- coefi[2]
> par(las=1)
> plot(gv, nsemdep/nsem, cex = nsem/10, xlab= "Grosor vaina (mm)",
+ ylab= "Probabilidad de depredación")
> curve(plogis(a.est + b.est*x), add=TRUE, col="red")
> curve(plogis(a + b*x), add=TRUE, col="blue")
> legend("topright", c("coef. sim.", "coef. est."),
+ col=c("blue", "red"), lty=c(1,1))
> plogis(a.est + b.est*0.7)

(Intercept)
    0.17335

> vp07 <- list(gv=c(0.3,0.5,0.7))
> predict(m1, vp07, type="response", se=T)

$fit
      1       2       3
0.83322 0.50582 0.17335
```

```
$se.fit
       1        2        3
0.019952 0.024322 0.023741

$residual.scale
[1] 1
```



1. ¿Cuál es el modelo con el que se deben analizar estos datos?

Modelo lineal generalizado con distribución binomial. Si la depredación de las semillas dentro de la misma vaina no fuera independiente (e.g. los bruquidos entraran por el agujero de la vaina que otro brúquido creó, los datos podrían analizarse como un glmm con la vaina como variable aleatoria, y la distribución de tipo Bernoulli.

2. Especifica el valor estimado de los coeficientes del modelo.

Ver código. summary(m1).

3. ¿Cuáles son las probabilidades de que una semilla sea depredada cuando el grosor de la vaina es de 0.3, 0.5 o 0.7 mm respectivamente? Reporta los errores estándar asociados a estos valores predichos.

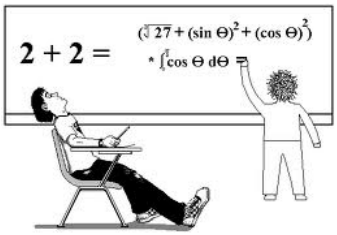
   Ver código, función predict con el argumento:
   type= "response" y se = TRUE.

4. Grafica los valores predichos de la probabilidad de que una semilla de *Acacia* sea depredada por escarabajos brúquidos en función del grosor de la vaina que contiene a las semillas. Toma en cuenta que el número de semillas varía entre las vainas. **Pista:** Emplea el argumento `cex` para modificar el tamaño de los símbolos en la gráfica dependiendo de cuantas observaciones hubo por vaina.

   Ver gráfica.

2.3. Ejercicio 3 (2.5 puntos)

Se realizó una prueba de desempeño psicométrico en 100 estudiantes varones de posgrado de ciencias biológicas con el fin de determinar como la falta de sueño incide en la capacidad de realizar operaciones matemáticas correctas. Se contabilizó cuantas operaciones matemáticas del mismo grado de dificultad podían resolver en un lapso de 10 minutos. Todos los estudiantes fueron sometidos a tres tratamientos, aleatorizando el orden en que fueron asignados: a) el control: 8+ hrs sueño y b) 4 hrs sueño y c) la noche previa en vigilia. La base de datos se llama `vigilia`.



El modelo se describe como:

$$y_i = \alpha_{j(i)} + \beta_1 * vigilia_4h + \beta_2 * vigilia_8h + \epsilon_i \quad (6)$$

$$\epsilon_i \sim N(0, \sigma^2) \quad (7)$$

$$a_j \sim N(\mu_a, \sigma_a^2) \quad (8)$$

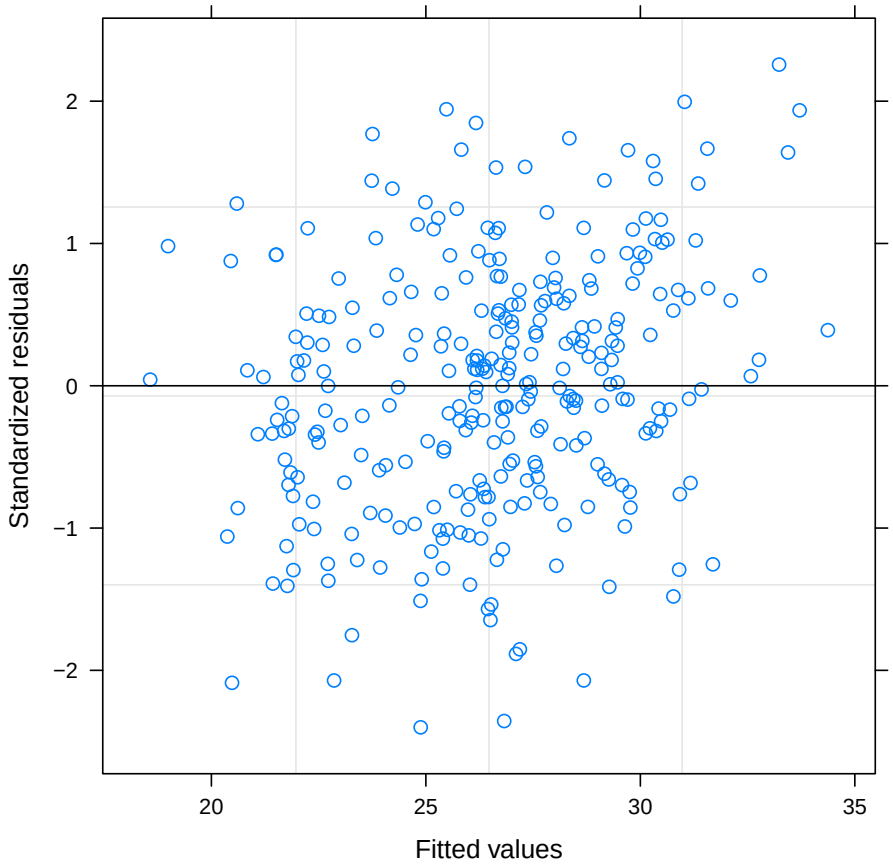
```
> ## linear mixed effects, mucho suenho 8hrs, poco suenho 4hrs, vigilia
> set.seed(1111)
> nind <- 100
> est <- gl(3,1,300, labels=c("ms","ns", "ps"))
> ind <- rep(1:nind, each=3)
> mms <- 28
> dems <- 2
> ind.means <- rnorm(nind,mms,dems)
> eps <- rnorm(300, 0,3)
> x <- rep(1:nind, rep(3,100))
> X <- as.matrix(model.matrix(~as.factor(x)-1))
> yal <- as.numeric( X %*% as.matrix(ind.means) + eps)
> efs <- rep(c(0,-5,-1),100)
> om <- yal + efs
> vigilia <- data.frame(ind,est, om)
> m1 <- lme(om ~ est, random=~1|ind, data=vigilia)
> summary(m1)

Linear mixed-effects model fit by REML
 Data: vigilia
   AIC    BIC  logLik
  1604 1622.5 -797.01

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

```
StdDev:      2.4866   2.8347

Fixed effects: om ~ est
              Value Std.Error  DF t-value p-value
(Intercept) 28.4026   0.37707 198  75.323  0.0000
estns       -4.9446   0.40088 198 -12.334  0.0000
estps       -0.6578   0.40088 198  -1.641  0.1024
 Correlation:
      (Intr) estns
estns -0.532
estps -0.532  0.500

Standardized Within-Group Residuals:
      Min        Q1       Med        Q3       Max
-2.400420 -0.638111  0.018018  0.602374  2.257344

Number of Observations: 300
Number of Groups: 100
```

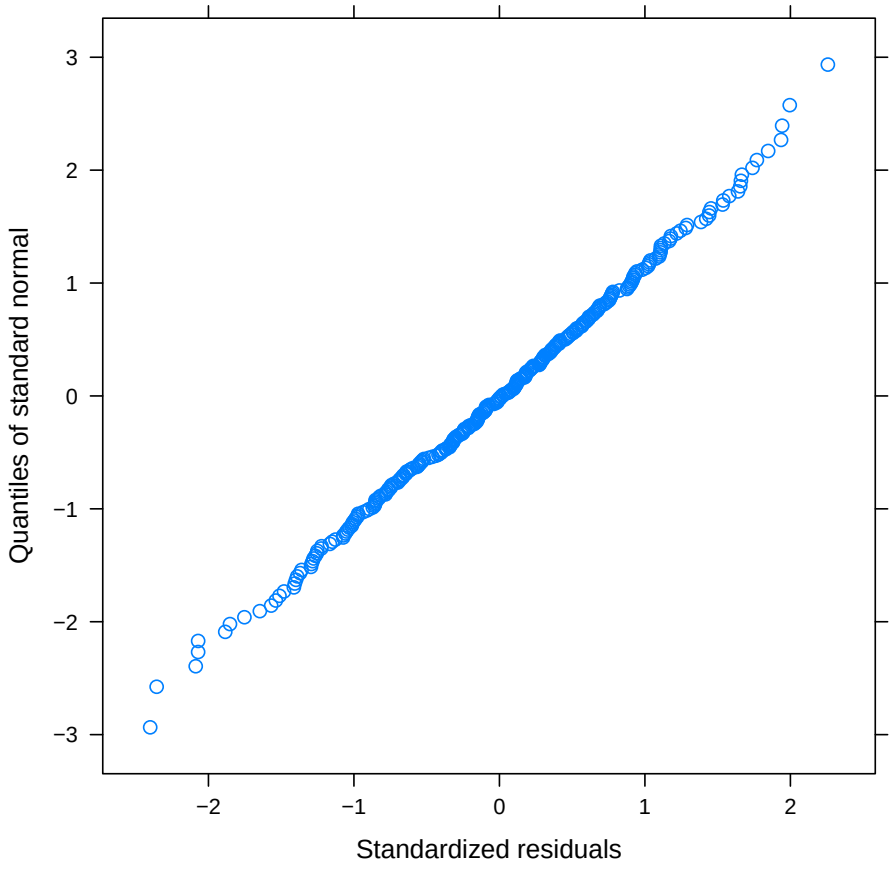
```
> anova(m1)
```

```
            numDF denDF F-value p-value
(Intercept)     1   198  7945.6  <.0001
est             2   198    89.7  <.0001
```

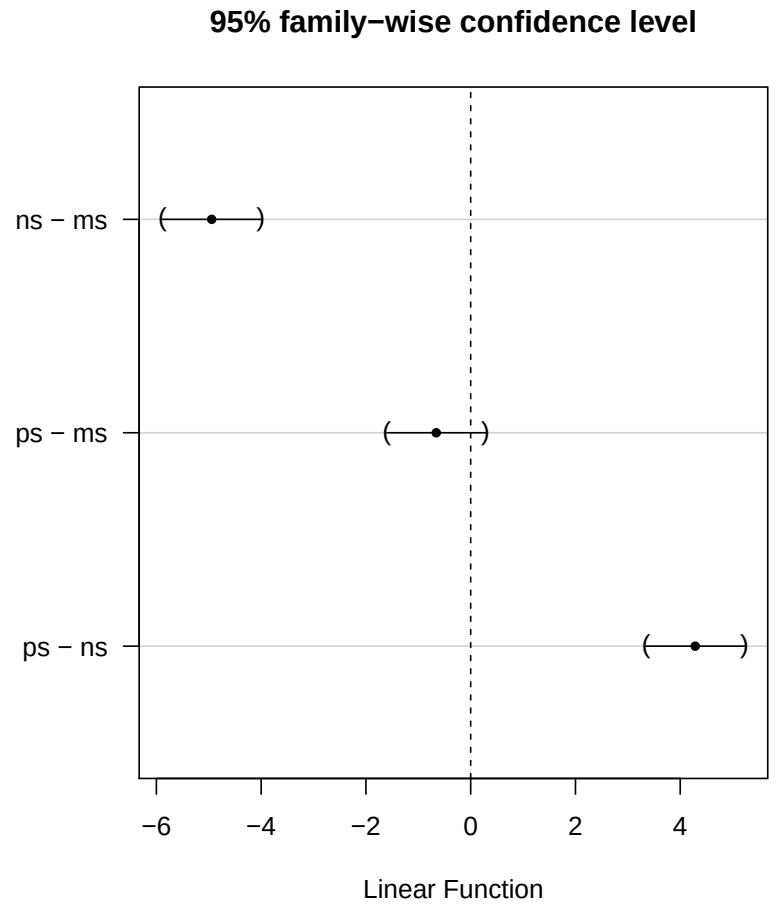
```
> #Validación del modelo
> plot(m1)
```




```
> qqnorm(m1)
```




```
> library(multcomp)
> cm <-glht(m1, linfct = mcp(est= "Tukey"))
> iccm<- confint(cm)
> par(mar=c(5,8,3,2), las=1)
> plot(iccm)
> save(acacia,duraznos,vigilia, file="bdatos.RData")
```



1. ¿Qué tipo de modelo es? ¿Por qué?

   Modelo lineal mixto, con un factor (nivel de sueño) como efecto fijo y los individuos como variable aleatoria. El modelo es mixto porque las mediciones psicométricas tomadas a un mismo individuo no son independientes entre sí.

2. Determina si la falta de sueño incide en la habilidad de realizar operaciones matemáticas. Reporta los coeficientes parciales del modelo y el estadístico de prueba que considera si hay un efecto de las horas de sueño en las operaciones matemáticas.

   Si hay un efecto en el desempeño de las operaciones psicométricas, ver código para observar los coeficientes del modelo y el estadístico de prueba.

3. En caso de que existieran diferencias entre los niveles del factor sueño, reporta cuáles difieren entre si. Obtén los valores predichos y los intervalos de confianza al 95 % para cada nivel.

   Ver la gráfica que antecede a las preguntas, no existen diferencias entre los niveles poco y mucho sueño.

Valores predichos: ms=28.4(IC= 27.67, 29.15),

ns= 23.46 (IC= 22.72, 24.2), ps=27.74(IC = 27.00, 28.49).

4. ¿Cuál es la variabilidad existente en el desempeño entre los estudiantes?

La variabilidad existente entre los estudiantes es de 2.5 desviaciones estándar, no muy alejada de la desviación de 2 unidades utilizada en la creación de la base de datos.