XS-3170 APLICACIONES DE DISEÑOS EXPERIMENTALES

MANUAL DE LABORATORIO

ESCUELA DE ESTADISTICA

UNIVERSIDAD DE COSTA RICA

RICARDO ALVARADO BARRANTES

2019

PREFACIO

El presente trabajo es un manual de laboratorios para el curso **XS-3170 Aplicaciones de Diseños Experimentales**, el cual forma parte del programa de Bachillerato en Estadística de la Universidad de Costa Rica. Este curso se ubica en el sexto semestre de la carrera de Estadística y tiene como requisito un curso introductorio llamado **Introducción a los Diseños Experimentales**, el que a su vez tiene como requisito el curso **Modelos de Regresión Aplicados**. Estos tres cursos son una secuencia que pretende desarrollar en los estudiantes las habilidades para planear y conducir adecuadamente experimentos con validez estadística, utilizando los modelos matemáticos apropiados para el análisis de los datos obtenidos en los estudios planteados.

En el primer curso se enseñan las bases de los modelos de regresión, los supuestos y las aplicaciones en estudios observacionales. En el siguiente curso se introducen los diseños experimentales, las técnicas para hacer comparaciones entre tratamientos, los bloques aleatorizados y el uso de covariables que permiten reducir la variabilidad del error experimental. En ambos cursos se concluye con el uso de respuestas binarias, específicamente con el modelo logístico, el cual se usa tanto en estudios observacionales como en estudios experimentales. Todos los modelos que se incluyen en estos cursos asumen que los factores son de efectos fijos.

Programa

El programa del curso que se presenta aquí empieza con un primer capítulo que aborda los diseños para la optimización de una respuesta, los cuales son empleados usualmente en la industria. Estos diseños se conocen como superficies de respuesta. En el segundo capítulo se hace una extensión de los Modelos Lineales Generalizados (glm, por sus siglas en inglés) a variables de conteo, para lo cual se usan modelos con distribución Poisson, quasi-Poisson o Binomial Negativa. Los siguientes cuatro capítulos sirven como introducción a algunos modelos mixtos. El tercer capítulo inicia con los modelos que incorporan uno o varios factores de efectos aleatorios. El cuarto capítulo incorpora factores que tienen una estructura anidada. Se considera tanto el caso en que los factores son de efectos fijos como cuando son de efectos aleatorios. En el quinto capítulo se presentan diseños que tienen dos factores, en los cuales no hay una asignación totalmente aleatoria de las unidades a los tratamientos. El diseño se hace en dos etapas: primero se asignan unidades primarias a los niveles de un factor que se llama parcela, luego cada una de las parcelas se subdivide en unidades secundarias llamadas subparcelas y estas se asignan a los niveles del segundo factor. El último capítulo está dedicado a modelos mixtos con mediciones repetidas sobre sujetos que han sido tomados aleatoriamente de una población. En estos casos, se tiene una dependencia de las observaciones en el tiempo, ya que cada grupo de observaciones pertenece a un mismo sujeto, lo cual hace que esas observaciones estén correlacionadas.

Uso del material

Cada capítulo contiene uno o varios ejercicios basados en un problema, los cuales se desarrollan usando el lenguaje de programación R versión 3.5.1 (R Core Team, 2018). Se espera que el estudiante realice los ejecicios y luego compare sus resultados con las respuestas. Para esto se hace una descripción del problema y luego se da una lista de preguntas con ayudas para resolver los ejercicios. Se indican las funciones de R recomendadas para contestar cada pregunta y la sintaxis apropiada para usar esas funciones. Los ejercicios empiezan con un análisis descriptivo de los datos para que el estudiante pueda visualizarlos sin entrar en la formulación de ningún modelo. Después de cada lista de preguntas se desarrolla cada item con la sintaxis de R y los comentarios pertinentes. Al final de cada problema se da una conclusión.

En la dirección https://drive.google.com/open?id=1hXeDWWtSGGjh3ZZDUk7ygsE8Rg_95Pbk están disponibles todos los datos que se usan en los ejercicios, así como una versión electrónica de este manual. Se pueden enviar sugerencias, preguntas o reportar errores al correo electrónico rab.libros@gmail.com.

Ricardo Alvarado Barrantes

San José, Costa Rica

Índice

1. SUPERFICIES DE RESPUESTA

En este laboratorio se presentan dos ejercicios donde se emplean modelos de superficies de respuesta en los cuales se cuenta con una variable respuesta que se quiere optimizar. En el primer ejercicio llamado **Rendimiento** se ilustran las tres etapas del proceso de optimización: 1) ajuste de un modelo de primer orden para determinar la dirección en que se debe mover, 2) movimiento con máxima pendiente en ascenso y 3) ajuste de un modelo de segundo orden. El segundo ejemplo llamado **Humedad** ilustra una situación con un punto de silla en la que se tiene una meta y se logra resolver la ecuación para alcanzar esa meta.

Se utiliza la librería `rsm` (Lenth, 2009) para el análisis de modelos de superficies de respuesta, estimación de parámetros y graficación de contornos. Además se usa la librería `lattice` (Sarkar, 2008) para visualización de interacciones y la librería `rgl` (Adler et. al, 2017) para visualización en tres dimensiones.

1.1. Rendimiento

Un ingeniero químico quiere encontrar condiciones de operación que maximicen el rendimiento de un proceso. Utiliza dos variables controlables que influyen: tiempo de reacción y temperatura de reacción. Actualmente las condiciones de operación son: 35min y 155^oF, y se obtienen rendimientos de alrededor de 40 %. Dado que hay muchas posibilidades de mejorar ese rendimiento, es muy probable que el óptimo no se encuentre en esta región. Se decide explorar en una región donde el tiempo esté entre 30 y 40 minutos y la temperatura entre 150^oF y 160^oF. Se usa un diseño 2^2 con solo una repetición en cada punto de diseño, y se aumenta con 5 puntos centrales. El diseño está centrado en las condiciones actuales.

Ejercicios

1. Abra el archivo `rendimiento.Rdata`.

- Haga una representación gráfica para ver si existe interacción entre los dos factores de diseño en los 4 puntos del diseño factorial con respecto a la variable respuesta `rend`. Use la función `xyplot` de la librería `lattice`. Tiene que usar solo los primeros cuatro puntos de la base y esto lo puede hacer de esta forma: `with(base[1:4,],xyplot(rend~temp,groups=tiempo,type = " o"))`
- Use la librería `rgl` para visualizar los datos en tres dimensiones. Hay dos variables respuesta `rend` y `rend1`. Mueva el gráfico para darse cuenta en cuál de los dos casos se puede esperar que el modelo de primer orden tenga un buen ajuste. Use: `plot3d(tiempo,temp,rend)`.

---

2. Para la variable respuesta `rend`, estime la variancia del error puro mediante la variancia de los puntos centrales. ¿Cuántos grados de libertad tiene esta estimación?

---

3. Use la librería `rsm` para descartar la posibilidad de interacción entre los dos factores de diseño usando siempre la variable respuesta `rend`. Escriba: `rsm(rend~FO(tiempo,temp)+TWI(tiempo,temp))`, donde **FO** significa First Order (modelo de primer orden) y **TWI** (two way interaction) es la interacción entre esos dos factores. Vea el `anova` y analice solamente la línea de la interacción en el análisis de variancia.

- Asumiendo que no hay interacción, ajuste un modelo de primer orden. Observe en el `summary` la estimación del error puro y la del error cuadrático. Haga la prueba de falta de ajuste (establezca la hipótesis nula adecuada y pruébela). Haga lo mismo con la variable rend1.
- En cada caso observe el coeficiente de determinación y justifique lo que está sucediendo.
- En el caso de **rend** vea la significancia de los coeficientes y explique su significado.
- Escriba la ecuación estimada del modelo de primer orden sin interacción.
- Si se quiere mover hacia la región del óptimo, ¿en qué dirección hay que mover la temperatura (subirla o bajarla) y el tiempo (subirlo o bajarlo)?

---

4. Haga el gráfico de contorno con `contour(mod2a,~tiempo+temp,image=T)`.

- Si se tiene una temperatura de 155^oF, según esta ecuación, ¿cuánto debería ser el tiempo para llegar a un rendimiento de 40 %? ¿Hay otras combinaciones de tiempo y temperatura que produzcan un 40 % de rendimiento?

---

5. Proponga un plan de experimentación secuencial para moverse hacia la región del máximo rendimiento bajo los siguiente escenarios. Recuerde que:

$$\Delta x_j = \frac{\hat{\beta}_j}{\hat{\beta}_i}\Delta x_i$$

- Encuentre cuánto debe moverse la temperatura si:

  i. Se le pide moverse cada 5 minutos a partir del punto central.

  ii. Se le pide moverse cada 2 minutos a partir del punto central.

- Encuentre cuánto debe moverse el tiempo si:

  i. Se le pide moverse 2^oF a partir del punto central.

  ii. Se le pide moverse 1^oF a partir del punto central.

---

6. Se optó por moverse cada 5 minutos y cada 2^oF. Al hacer el experimento secuencial se obtuvieron los datos que se encuentran en `base1`. Haga un gráfico del rendimiento en secuencia del tiempo de ejecución y determine en qué área se puede encontrar el óptimo.

---

7. Basado en los resultados del punto anterior, se decide hacer un nuevo experimento con un modelo de primer orden centrado en tiempo de 85 minutos y temperatura de 175^oF. Se decide explorar en una región donde el tiempo esté entre 80 y 90 minutos y la temperatura entre 170^oF y 180^oF. Los resultados están en `base2`. Analice estos resultados.

---

8. Debido a la falta de ajuste en los resultados del punto anterior se completa el diseño con puntos axiales para obtener un diseño central compuesto. Se llevan a cabo las corridas adicionales que se muestran en `base3`.

   - Estime el modelo de segundo orden usando `rsm(rend~SO(tiempo,temp),base3)`. Analice los términos de segundo orden.
   - Obtenga los valores propios de la matriz B (use `summary(mod4b)$canonical`) para concluir si la superficie alcanza un máximo, un mínimo o tiene un punto de silla.
   - Obtenga los valores de tiempo y temperatura donde se alcanza el óptimo. Búsquelo en el `summary`.
   - Obtenga el rendimiento promedio que se espera obtener con esos valores de tiempo y temperatura.
   - Obtenga el gráfico de contorno y ubique el máximo.
   - Si el ingeniero está satisfecho con un rendimiento de 78 % y no quiere tener un tiempo de reacción muy alto, ¿puede encontrarse una temperatura que sea adecuada sin necesidad de llegar a la que propone el punto óptimo?
   - De qué forma se podría encontrar un intervalo de confianza para el rendimiento esperado en una combinación específica de tiempo y temperatura?
   - ¿Qué le sugiere el hecho de que haya falta de ajuste?

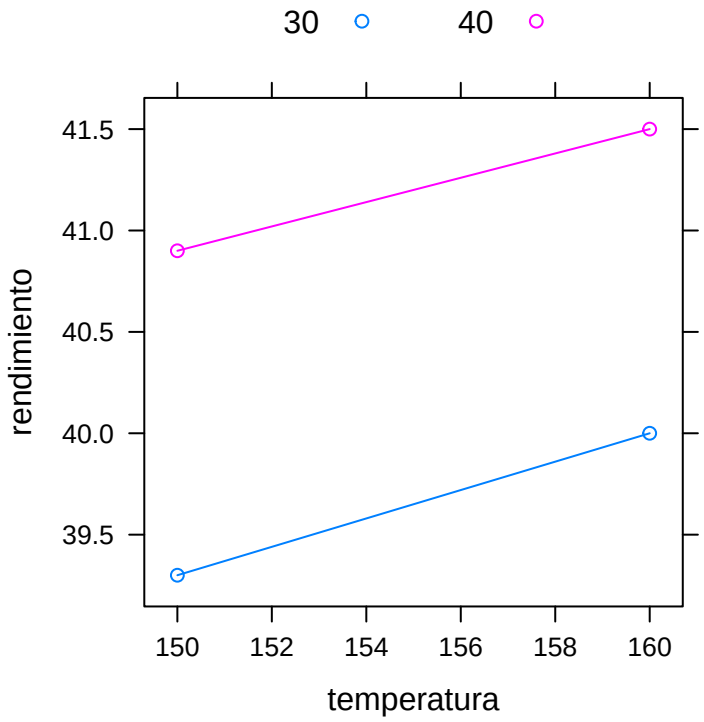
Solución

1. Abra el archivo `rendimiento.Rdata`.

```
load("rendimiento.Rdata")
attach(base)
```

- Haga una representación gráfica para ver si existe interacción entre los dos factores de diseño en los 4 puntos del diseño factorial con respecto a la variable respuesta `rend`. Use la función `xyplot` de la librería `lattice`. Tiene que usar solo los primeros cuatro puntos de la base y esto lo puede hacer de esta forma: `with(base[1:4,],xyplot(rend~temp,groups=tiempo,type= " o"))`

```
library(lattice)
with(base[1:4,],xyplot(rend~temp,groups=tiempo,type="o",
                       auto.key=list(columns=2),
                       xlab="temperatura",ylab="rendimiento"))
```



Definitivamente no hay interacción entre temperatura y tiempo.

- Use la librería `rgl` para visualizar los datos en tres dimensiones. Hay dos variables respuesta `rend` y `rend1`. Mueva el gráfico para darse cuenta en cuál de los dos casos se puede esperar que el modelo de primer orden tenga un buen ajuste. Use: `plot3d(tiempo,temp,rend)`.

```
library(rgl)
plot3d(tiempo,temp,rend)
plot3d(tiempo,temp,rend1)
```

En el primer caso (rend) se ven los puntos centrales muy cercanos al plano que pasa por las esquinas del cubo, mientras que en el otro caso (rend1) los puntos centrales están lejos del plano. Entonces se podría esperar que el modelo de primer orden ajuste mejor en el caso de rend.

Nota: los gráficos en tres dimensiones no aparecen aquí.

---

2. Para la variable respuesta `rend`, estime la variancia del error puro mediante la variancia de los puntos centrales. ¿Cuántos grados de libertad tiene esta estimación?

```
  var(rend[5:9])
```

```
## [1] 0.043
```

La variancia es 0.043 y puesto que se usan 5 puntos, esta estimación tiene 4 grados de libertad.

---

3. Use la librería `rsm` para descartar la posibilidad de interacción entre los dos factores de diseño usando siempre la variable respuesta `rend`. Escriba: `rsm(rend~FO(tiempo,temp)+TWI(tiempo,temp))`, donde **FO** significa First Order (modelo de primer orden) y **TWI** (two way interaction) es la interacción entre esos dos factores. Vea el `anova` y analice solamente la línea de la interacción en el análisis de variancia.

```
library(rsm)
mod1a=rsm(rend~FO(tiempo,temp)+TWI(tiempo,temp))
anova(mod1a)
```

```
## Analysis of Variance Table
##
## Response: rend
##                    Df  Sum Sq Mean Sq F value    Pr(>F)
## FO(tiempo, temp)    2 2.82500 1.41250 40.4213 0.0008188 ***
## TWI(tiempo, temp)   1 0.00250 0.00250  0.0715 0.7997870
## Residuals           5 0.17472 0.03494
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La interacción no resulta significativa ya que la probabilidad asociada es muy alta (p=0.8).

- Asumiendo que no hay interacción, ajuste un modelo de primer orden. Observe en el `summary` la estimación del error puro y la del error cuadrático. Haga la prueba de falta de ajuste (establezca la hipótesis nula adecuada y pruébela). Haga lo mismo con la variable rend1.

```
mod2a=rsm(rend~FO(tiempo,temp))
summary(mod2a)
```

```
##
## Call:
## rsm(formula = rend ~ FO(tiempo, temp))
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 24.944444   2.731553  9.1320 9.697e-05 ***
## tiempo       0.155000   0.017186  9.0188  0.000104 ***
## temp         0.065000   0.017186  3.7821  0.009158 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Multiple R-squared:  0.941,  Adjusted R-squared:  0.9213
## F-statistic: 47.82 on 2 and 6 DF,  p-value: 0.0002057
##
## Analysis of Variance Table
##
## Response: rend
##                   Df  Sum Sq Mean Sq F value    Pr(>F)
## FO(tiempo, temp)   2 2.82500 1.41250 47.8213 0.0002057
## Residuals          6 0.17722 0.02954
## Lack of fit        2 0.00522 0.00261  0.0607 0.9419341
## Pure error         4 0.17200 0.04300
##
## Direction of steepest ascent (at radius 1):
##    tiempo      temp
## 0.9221944 0.3867267
##
## Corresponding increment in original units:
##    tiempo      temp
## 0.9221944 0.3867267
```

```r
mod2b=rsm(rend1~FO(tiempo,temp))
summary(mod2b)
```

```
##
## Call:
## rsm(formula = rend1 ~ FO(tiempo, temp))
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.05556   19.86953  1.3113   0.2377
## tiempo       0.15500    0.12501  1.2399   0.2613
## temp         0.06500    0.12501  0.5199   0.6217
##
## Multiple R-squared:  0.2315, Adjusted R-squared:  -0.02465
## F-statistic: 0.9038 on 2 and 6 DF,  p-value: 0.4538
##
## Analysis of Variance Table
##
## Response: rend1
##                   Df Sum Sq Mean Sq  F value    Pr(>F)
## FO(tiempo, temp)   2 2.8250  1.4125   0.9038 0.4538432
## Residuals          6 9.3772  1.5629
## Lack of fit        2 9.2052  4.6026 107.0375 0.0003364
## Pure error         4 0.1720  0.0430
##
## Direction of steepest ascent (at radius 1):
##    tiempo      temp
## 0.9221944 0.3867267
##
## Corresponding increment in original units:
##    tiempo      temp
## 0.9221944 0.3867267
```

Cuando se analiza falta de ajuste, la hipótesis nula indica que no hay falta de ajuste, es decir, que el modelo propuesto ajusta adecuadamente. En el caso de rend no se rechaza esta hipótesis, pues la probabilidad asociada es muy alta (p=.94). Entonces se puede asumir que el modelo de primer orden si da un buen ajuste, lo que concuerda con lo que se vio en el gráfico. En el caso de rend1, pasa lo contrario pues la probabilidad asociada es muy baja (p=0.0003), entonces se concluye que el modelo de primer orden no da un buen ajuste.

- En cada caso observe el coeficiente de determinación y justifique lo que está sucediendo.

Para rend se tiene un $R^2 = 0{,}94$ lo que favorece el buen ajuste del modelo, en cambio para rend1 $R^2 = 0{,}23$, lo cual concuerda con que existe mal ajuste del modelo.

- En el caso de **rend** vea la significancia de los coeficientes y explique su significado.

```
summary(mod2a)$coef
```

```
##             Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 24.94444 2.73155336 9.131963 9.697124e-05
## tiempo       0.15500 0.01718634 9.018789 1.040409e-04
## temp         0.06500 0.01718634 3.782073 9.158066e-03
```

Las probabilidades asociadas a la hipótesis $H_0 : \beta_j = 0$ son muy bajas, entonces, tanto el coeficiente de tiempo (p=0.0001) como el de temperatura (p=0.009) son significativos. Ambos coeficientes son positivos, lo cual indica que al aumentar cualquiera de esas dos variables va a aumentar la respuesta de forma importante.

- Escriba la ecuación estimada del modelo de primer orden sin interacción.

Si llamamos X_1 al tiempo y X_2 a la temperatura, la ecuación para el rendimiento promedio estimado estaría dada por:

$$\hat{y} = 24{,}94 + 0{,}155X_1 + 0{,}065X_2$$

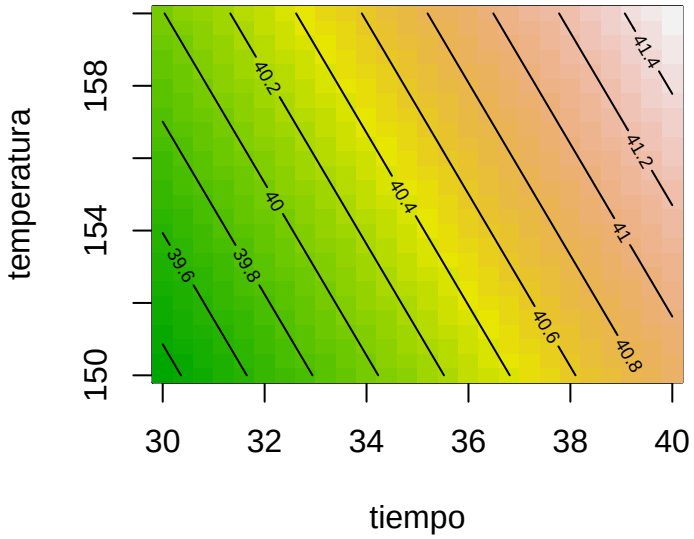
- Si se quiere mover hacia la región del óptimo, ¿en qué dirección hay que mover la temperatura (subirla o bajarla) y el tiempo (subirlo o bajarlo)?

Ya que se quiere aumentar el rendimiento y ambos coeficientes son positivos, hay que subir tanto la temperatura como el tiempo.

---

4. Haga el gráfico de contorno con `contour(mod2a,~tiempo+temp,image=T)`.

```
contour(mod2a,~tiempo+temp,image=T,xlabs=c("temperatura","tiempo"))
```



- Si se tiene una temperatura de 155^oF, según esta ecuación, ¿cuánto debería ser el tiempo para llegar a un rendimiento de 40 %? ¿Hay otras combinaciones de tiempo y temperatura que produzcan un 40 % de rendimiento?

Para responder a esto se fija la temperatura en 155 y se despeja en la ecuación anterior:

$$40 = 24{,}94 + 0{,}155X_1 + 0{,}065 \times 155$$

```
(40-24.94-0.065*155)/0.155
```

```
## [1] 32.16129
```

El tiempo debería ser de 32.2 minutos. Con esto se asegura que el rendimiento promedio sea de 40 %, pero no asegura que una única corrida vaya a tener ese rendimiento ya que puede estar por encima o por debajo y dependerá del error qué tanto se aleje del 40 % . Dado que este modelo tiene un R^2 bastante alto, seguramente los valores van a estar bastante cerca de 40 %.

Como se aprecia en el gráfico producto de la ecuación, hay muchas combinaciones de tiempo y temperatura que producirían un rendimiento promedio de 40 %.

---

5. Proponga un plan de experimentación secuencial para moverse hacia la región del máximo rendimiento bajo los siguiente escenarios. Recuerde que:

$$\Delta x_j = \frac{\hat{\beta}_j}{\hat{\beta}_i}\Delta x_i$$

- Encuentre cuánto debe moverse la temperatura si:

i. Se le pide moverse cada 5 minutos a partir del punto central.

```
b=mod2a$coef
5*(b[3]/b[2])
```

```
## FO(tiempo, temp)temp
##             2.096774
```

El movimiento debe ser cada 5 minutos y $2{,}1^oF$**.**

ii. Se le pide moverse cada 2 minutos a partir del punto central.

```
2*(b[3]/b[2])
```

```
## FO(tiempo, temp)temp
##            0.8387097
```

El movimiento debe ser cada 2 minutos y $0{,}8^oF$**.**

- Encuentre cuánto debe moverse el tiempo si:

i. Se le pide moverse 2^oF a partir del punto central.

```
2*(b[2]/b[3])
```

```
## FO(tiempo, temp)tiempo
##               4.769231
```

El movimiento debe ser cada 2^oF **y 4.8 minutos.**

ii. Se le pide moverse 1^oF a partir del punto central.

```
1*(b[2]/b[3])
```

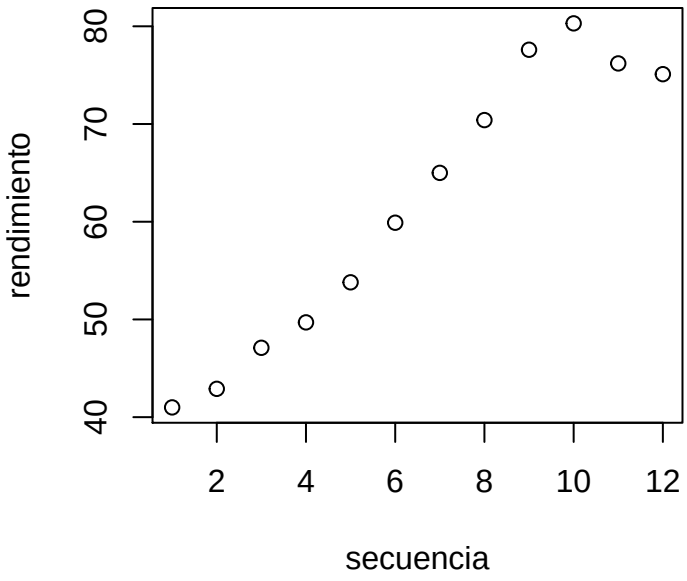
```
## FO(tiempo, temp)tiempo
##               2.384615
```

El movimiento debe ser cada 1^oF **y 2.4 minutos.**

---

6. Se optó por moverse cada 5 minutos y cada 2^oF. Al hacer el experimento secuencial se obtuvieron los datos que se encuentran en `base1`. Haga un gráfico del rendimiento en secuencia del tiempo de ejecución y determine en qué área se puede encontrar el óptimo.

```
plot(base1$rend,xlab="secuencia",ylab="rendimiento")
```



Parece ser que el óptimo se alcanza cerca del décimo punto que tiene un tiempo de 85 minutos y una temperatura de 175^oF.

---

7. Basado en los resultados del punto anterior, se decide hacer un nuevo experimento con un modelo de primer orden centrado en tiempo de 85 minutos y temperatura de 175^oF. Se decide explorar en una región donde el tiempo esté entre 80 y 90 minutos y la temperatura entre 170^oF y 180^oF. Los resultados están en `base2`. Analice estos resultados.

```
mod3a=rsm(rend~FO(tiempo,temp)+TWI(tiempo,temp),base2)
anova(mod3a)
```

```
## Analysis of Variance Table
##
## Response: rend
##                    Df Sum Sq Mean Sq F value Pr(>F)
## FO(tiempo, temp)    2   5.00   2.500   1.150 0.3883
## TWI(tiempo, temp)   1   0.25   0.250   0.115 0.7483
## Residuals           5  10.87   2.174
```

```r
mod3a=rsm(rend~FO(tiempo,temp),base2)
summary(mod3a)
```

```
##
## Call:
## rsm(formula = rend ~ FO(tiempo, temp), data = base2)
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 44.46667   26.48948  1.6787   0.1442
## tiempo       0.20000    0.13614  1.4691   0.1922
## temp         0.10000    0.13614  0.7346   0.4903
##
## Multiple R-squared:  0.3102, Adjusted R-squared:  0.08023
## F-statistic: 1.349 on 2 and 6 DF,  p-value: 0.3283
##
## Analysis of Variance Table
##
## Response: rend
##                   Df Sum Sq Mean Sq  F value    Pr(>F)
## FO(tiempo, temp)  2  5.000  2.5000   1.3489 0.3282610
## Residuals         6 11.120  1.8533
## Lack of fit       2 10.908  5.4540 102.9057 0.0003635
## Pure error        4  0.212  0.0530
##
## Direction of steepest ascent (at radius 1):
##    tiempo      temp
## 0.8944272 0.4472136
##
## Corresponding increment in original units:
##    tiempo      temp
## 0.8944272 0.4472136
```

En el primer modelo se tiene que la interacción entre tiempo y temperatura no es significativa (p=0.75), por lo que se elimina y se hace un segundo modelo sin interacción. En el segundo modelo se observa falta de ajuste (p=0.0003), lo cual es una buena indicación de que se ha llegado a una zona donde hay una curvatura, ya que el modelo de primer orden no ajusta adecuadamente.

---

8. Debido a la falta de ajuste en los resultados del punto anterior se completa el diseño con puntos axiales para obtener un diseño central compuesto. Se llevan a cabo las corridas adicionales que se muestran en `base3`.

- Estime el modelo de segundo orden usando `rsm(rend~SO(tiempo,temp),base3)`. Analice los términos de segundo orden.

```
mod4a=rsm(rend~SO(tiempo,temp),base3)
anova(mod4a)
```

```
## Analysis of Variance Table
##
## Response: rend
##                    Df  Sum Sq Mean Sq F value   Pr(>F)
## FO(tiempo, temp)    2  8.7308  4.3654  9.9240 0.009050 **
## TWI(tiempo, temp)   1  0.2500  0.2500  0.5683 0.475513
## PQ(tiempo, temp)    2 16.6830  8.3415 18.9629 0.001493 **
## Residuals           7  3.0792  0.4399
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Primero se hace el modelo completo de segundo orden incluyendo la interacción. Se observa que ésta no es significativa. Sin embargo, en los modelos de segundo orden siempre se mantiene la interacción.

```
summary(mod4a)$coef
```

```
##                   Estimate   Std. Error    t value    Pr(>|t|)
## (Intercept) -1.309928e+03 3.776060e+02 -3.4690352 0.010419467
## tiempo       9.280666e+00 3.098551e+00  2.9951631 0.020079067
## temp         1.114622e+01 3.674589e+00  3.0333236 0.019025186
## tiempo:temp  1.000000e-02 1.326479e-02  0.7538757 0.475512672
## tiempo^2    -6.336293e-02 1.199106e-02 -5.2841795 0.001142845
## temp^2      -3.398050e-02 9.991479e-03 -3.4009485 0.011427720
```

Aquí se observa la significancia de los términos cuadráticos. Puesto que éstos resultan significativos en conjunto (p=0.0015) se reafirma lo que se había visto anteriormente, que nos encontramos en una zona con curvatura. Tanto el tiempo como la temperatura están contribuyendo a la curvatura, ya que los términos cuadráticos de ambas variables son significativos por separado (p=0.001 para tiempo cuadrático y p=0.011 para temperatura cuadrática).

- Obtenga los valores propios de la matriz B (use `summary(mod4b)$canonical`) para concluir si la superficie alcanza un máximo, un mínimo o tiene un punto de silla.

```
summary(mod4a)$canonical
```

```
## $xs
##    tiempo      temp
##  87.18859 176.83821
##
## $eigen
## eigen() decomposition
## $values
## [1] -0.03315296 -0.06419047
##
## $vectors
##              [,1]       [,2]
## tiempo -0.1632870 -0.9865786
## temp   -0.9865786  0.1632870
```

Los valores propios son -0.03 y -0.06. En este caso el punto estacionario es un máximo puesto que ambos valores son negativos.

- Obtenga los valores de tiempo y temperatura donde se alcanza el óptimo. Búsquelo en el `summary`.

```
summary(mod4a)$canonical$xs
```

```
##    tiempo      temp
##  87.18859 176.83821
```

```
round(mod4a$coef,3)
```

```
##              (Intercept)   FO(tiempo, temp)tiempo     FO(tiempo, temp)temp
##                -1309.928                    9.281                   11.146
##        TWI(tiempo, temp) PQ(tiempo, temp)tiempo^2   PQ(tiempo, temp)temp^2
##                    0.010                   -0.063                   -0.034
```

El máximo se alcanza para tiempo=87.19 y temperatura=176.84.

- Obtenga el rendimiento promedio que se espera obtener con esos valores de tiempo y temperatura.

El rendimiento máximo se obtiene al sustituir estos valores en la siguiente ecuación:

$$\hat{y} = -1309{,}928 + 9{,}281X_1 + 11{,}146X_2 + 0{,}010X_1X_2 - 0{,}063X_1^2 - 0{,}034X_2^2$$

```
x1=87.19; x2=176.84
-1309.928+9.281*x1+11.146*x2+0.01*x1*x2-0.063*x1^2-0.034*x2^2
```

```
## [1] 82.33466
```

Este mismo cálculo se puede obtener usando un vector que multiplica a los coeficientes estimados. Esta forma es más exacta y además permitiría obtener un intervalo de confianza para la media.

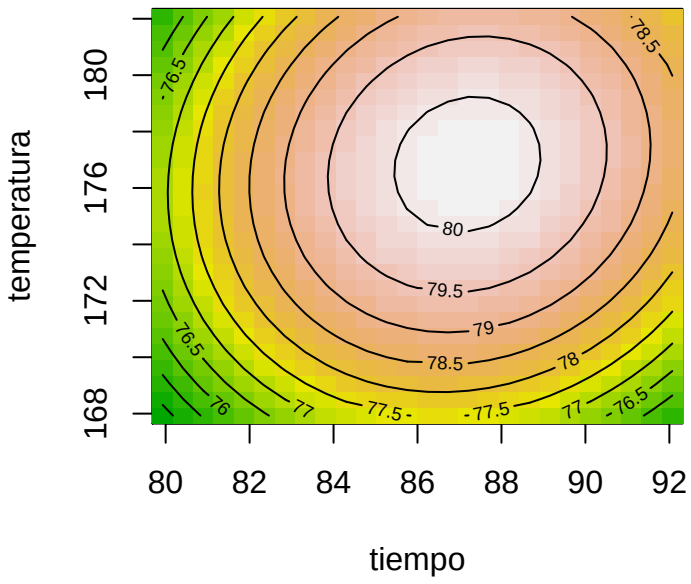
```
xh=c(1,x1,x2,x1*x2,x1^2,x2^2)
b=mod4a$coef
t(xh)%*%b
```

```
##          [,1]
## [1,] 80.1942
```

Con estos valores de tiempo (87.19) y temperatura (176.84) se debe alcanzar un rendimiento promedio de 80.2 %.

- Obtenga el gráfico de contorno y ubique el máximo.

```
contour(mod4a,~tiempo+temp,image=T,xlabs=c("temperatura","tiempo"))
```



En la figura anterior, la curva de contorno que tiene el mayor rendimiento está en 80 %, pero se puede ver que para los valores de tiempo de 87.19 y temperatura de 176.84, el rendimiento está dentro de ese círculo, lo cual coincide con el cálculo realizado anteriormente.

- Si el ingeniero está satisfecho con un rendimiento de 78 % y no quiere tener un tiempo de reacción muy alto, ¿puede encontrarse una temperatura que sea adecuada sin necesidad de llegar a la que propone el punto óptimo?

En el gráfico se puede ubicar el contorno para 78 % y escoger un tiempo adecuadamente bajo, por ejemplo 82 minutos. Entonces se fija ese tiempo y el rendimiento de 78 % en la ecuación estimada y se despeja para la temperatura:

$$78 = -1309{,}928 + 9{,}281 \times 82 + 11{,}146X_2 + 0{,}010 \times 82X_2 - 0{,}063 \times 82^2 - 0{,}034X_2^2$$

```
beta=mod4a$coef
c=-78+beta[1]+beta[2] * 82+beta[5]* 82^2
c
```

```
## (Intercept)
##   -1052.966
```

```
b=beta[3]+beta[4]*82
b
```

```
## FO(tiempo, temp)temp
##             11.96622
```

Entonces se debe resolver la ecuación de segundo grado:

$$-0{,}034X_2^2 + 11{,}966X_2 - 1052{,}97 = 0$$

```
 a=beta[6]
(-b+c(-1,1)*sqrt(b^2-4*a*c))/(2*a)
```

```
## [1] 179.9419 172.2076
```

Fijando el tiempo en 82 minutos se puede establecer la temperatura cerca de $172{,}2^oF$ o $179{,}9^oF$. Se puede verificar usando la función `predict` que el rendimiento promedio es cerca de 78 % con los valores propuestos:

```
predict(mod4a,data.frame(tiempo=82,temp=172.2))
```

```
##        1
## 77.99801
```

```
predict(mod4a,data.frame(tiempo=82,temp=179.9))
```

```
##        1
## 78.01096
```

- De qué forma se podría encontrar un intervalo de confianza para el rendimiento esperado en una combinación específica de tiempo y temperatura?

Aquí se puede agregar un intervalo de confianza de la media:

```
predict(mod4a,data.frame(tiempo=82,temp=172.2),
        interval="confidence")
```

```
##        fit      lwr      upr
## 1 77.99801 77.26835 78.72767
```

Se espera, con 95 % de confianza, que el rendimiento promedio al fijar el tiempo en 82 y la temperatura en 172.2 se encuentre entre 77.3 % y 78.7 %.

- ¿Qué le sugiere el hecho de que haya falta de ajuste?

Aunque haya falta de ajuste no debe haber preocupación porque el R^2 es bastante alto (0.893), lo cual indica que se tiene poco nivel de error y por eso se llegan a detectar problemas de ajuste que no son relevantes.

Conclusión: el estudio se dividió en tres etapas. En la primera parte se empezó haciendo la experimentación en una zona cercana a las condiciones actuales de operación y se determinó que en esa zona no había una curvatura, por lo tanto, existía un potencial de encontrar otra zona que permitiera maximizar el rendimiento. En una segunda etapa se experimentó sin un diseño fijo sino de una forma secuencial para tratar de encontrar una nueva zona donde la superficie mostrara curvatura. Una vez que se determinó que podría haber una curvatura se volvió a hacer un diseño alrededor de un punto escogido con base en la experimentación previa y se verificó que en esa parte, un plano no tendría un buen ajuste, lo cual daba más indicios de una curvatura, entonces se completaron las mediciones con puntos axiales para tener un diseño central compuesto. Con el último experimento se llegó a determinar una combinación de tiempo y temperaura que permiten un rendimiento promedio de alrededor de 80 %. Además la ecuación obtenida se puede usar para determinar otras combinaciones de tiempo y temperatura que permitan llegar a rendimientos meta que pueden ser menores al máximo.

1.2. Humedad

Don Hermes produce tintes para el cabello y nota que sus ventas están disminuyendo. Hace una encuesta entre los consumidores y encuentra que la razón más importante es que producen resequedad en el cabello. Teóricamente, la resequedad que se produzca en el cabello depende de la combinación de emulsificante y de colorante que se use.
Existe una relación inversa entre el contenido de colorante y/o emulsificante y la humedad del cabello.

Decide estudiar la humedad del cabello (que en realidad se mide por los aceites esenciales) después de la aplicación del tinte en función de estos dos factores para buscar la combinación que le dé la máxima humedad.

Ejercicios

1. Lea los datos del archivo `humedad.Rdata`. Asegúrese que las variables independientes no están como factor.

   - Observe los datos y a partir de ahí describa el diseño que se empleó.

---

2. Estime un modelo de segundo orden.

   - Obtenga los valores propios para determinar qué tipo de punto estacionario se tiene.
   - Obtenga los valores de las variables independientes para los cuales se llega al punto estacionario y comente qué problema tiene este punto.

---

3. Suponga que aunque don Hermes no encuentra una combinación que maximice la humedad, él se pone la meta de llegar al menos a 73.5 %. Obtenga el gráfico de contorno y ubique el valor mínimo del colorante que permita alcanzar esa meta dentro del rango del estudio.

   - Modifique el gráfico de contorno para observar mejor el punto de silla. Agregue en la instrucción del gráfico `bounds=list(colorante=c(40,60),emulsificante=c(10,20))`.

---

4. La decisión de identificar una meta y buscar, con base en el análisis, una combinación de factores con la que se alcance esa meta tiene un problema teórico desde el punto de vista estadístico. Diga cuál es y qué habría tenido que hacerse para evitarlo.

Solución

1. Lea los datos del archivo `humedad.Rdata`. Asegúrese que las variables independientes no están como factor.

```
load("humedad.Rdata")
str(base)
```

```
## 'data.frame':    27 obs. of  3 variables:
##  $ colorante    : int  5 5 5 5 5 5 5 5 5 10 ...
##  $ emulsificante: int  10 10 10 15 15 15 20 20 20 10 ...
##  $ humedad      : int  80 92 94 96 90 84 82 80 75 70 ...
```

- Observe los datos y a partir de ahí describa el diseño que se empleó.

```
base
```

```
##    colorante emulsificante humedad
## 1          5            10      80
## 2          5            10      92
## 3          5            10      94
## 4          5            15      96
## 5          5            15      90
## 6          5            15      84
## 7          5            20      82
## 8          5            20      80
## 9          5            20      75
## 10        10            10      70
## 11        10            10      74
## 12        10            10      79
## 13        10            15      84
## 14        10            15      70
## 15        10            15      69
## 16        10            20      67
## 17        10            20      66
## 18        10            20      60
## 19        15            10      55
## 20        15            10      60
## 21        15            10      62
## 22        15            15      65
## 23        15            15      60
## 24        15            15      60
## 25        15            20      57
## 26        15            20      51
## 27        15            20      50
```

```
attach(base)
```

```
table(colorante,emulsificante)
```

```
##          emulsificante
## colorante 10 15 20
##        5   3  3  3
##        10  3  3  3
##        15  3  3  3
```

El diseño que se empleó en el primer ensayo fue un irrestricto con un arreglo factorial de dos factores: colorante a tres niveles (5, 10 y 15) y emulsificante también a tres niveles (10, 15 y 25).

---

2. Estime un modelo de segundo orden.

```
mod1=rsm(humedad~SO(colorante,emulsificante))
```

- Obtenga los valores propios para determinar qué tipo de punto estacionario se tiene.

```
summary(mod1)$canonical
```

```
## $xs
##     colorante emulsificante
##      51.59338      16.14657
##
## $eigen
## eigen() decomposition
## $values
## [1]  0.03439736 -0.22773069
##
## $vectors
##                      [,1]       [,2]
## colorante     -0.9979684 -0.0637116
## emulsificante -0.0637116  0.9979684
```

Los valores propios son 0.03 y -0.23. Como son de signos opuestos se tiene un punto de silla.

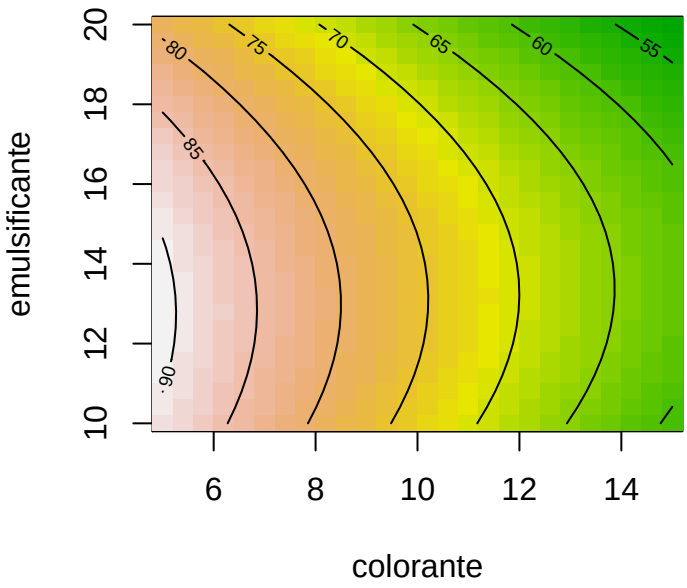
- Obtenga los valores de las variables independientes para los cuales se llega al punto estacionario y comente qué problema tiene este punto.

De la salida anterior se observa que el punto estacionario se alcanza cuado el colorante llega a 51.6 y el emulsificante a 16.1. El problema es que el experimento se realizó en un rango de colorante entre 5 y 15, por lo que se está obteniendo un punto estacionario que se encuentra fuera del rango del análisis.

---

3. Suponga que aunque don Hermes no encuentra una combinación que maximice la humedad, él se pone la meta de llegar al menos a 73.5 %. Obtenga el gráfico de contorno y ubique el valor mínimo del colorante que permita alcanzar esa meta dentro del rango del estudio.

```
contour(mod1,~colorante+emulsificante,image=T,
        xlabs=c("colorante","emulsificante"))
```



Por la forma que tienen las curvas de contorno, se observa que para un valor de emulsificante de 20 se tendría el menor valor de colorante que permita alcanzar la meta. Entonces se fija el colorante en 20 y la respuesta en 73.5 usando la ecuación que se construye con los coeficientes estimados.

```
beta=mod1$coef
names(beta)=c("intercept","C","E","CE","C2","E2")
round(beta,3)
```

```
## intercept         C         E        CE        C2        E2
##    73.222    -3.978     5.600     0.033     0.033    -0.227
```

$$\hat{\mu_Y} = 73{,}22 - 3{,}98C + 5{,}60E + 0{,}033CE + 0{,}033C^2 - 0{,}227E^2$$

De esta forma la ecuación que hay que resolver es:

$$73{,}5 = 73{,}22 - 3{,}98C + 5{,}60 \times 20 + 0{,}033 \times 20C + 0{,}033C^2 - 0{,}227 \times 20^2$$

```
beta=mod1$coef
a=beta[5]
a
```

```
## PQ(colorante, emulsificante)colorante^2
##                                0.03333333
```

```
b=beta[2]+beta[4]*20
b
```

```
## FO(colorante, emulsificante)colorante
##                              -3.311111
```

```
c=-73.5+beta[1]+beta[3] * 20+beta[6]* 20^2
c
```

```
## (Intercept)
##    21.05556
```

Entonces se debe resolver la ecuación de segundo grado:

$$0{,}033C^2 - 3{,}31C + 21{,}05 = 0$$

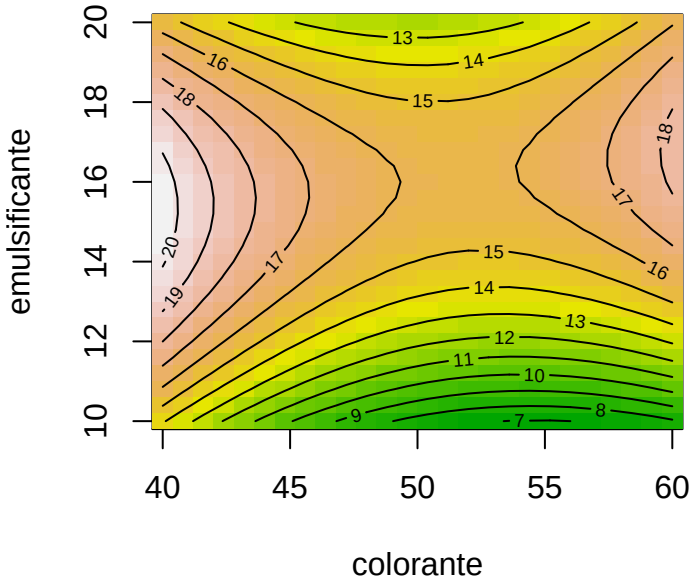
```
(-b+c(-1,1)*sqrt(b^2-4*a*c))/(2*a)
```

```
## [1]  6.82847 92.50486
```

Aunque la ecuación tiene dos soluciones, solo el valor de colorante de 6.8 está dentro del rango del estudio, por lo que la combinación que hace cumplir la meta de don Hermes es colorante en 6.8 y emulsificante en 20.0.

- Modifique el gráfico de contorno para observar mejor el punto de silla. Agregue en la instrucción del gráfico `bounds=list(colorante=c(40,60),emulsificante=c(10,20))`.

```
contour(mod1,~colorante+emulsificante,image=T,
        xlabs=c("colorante","emulsificante"),
        bounds=list(colorante=c(40,60),emulsificante=c(10,20)))
```



---

4. La decisión de identificar una meta y buscar, con base en el análisis, una combinación de factores con la que se alcance esa meta tiene un problema teórico desde el punto de vista estadístico. Diga cuál es y qué habría tenido que hacerse para evitarlo.

El problema teórico que tiene es que el diseño que se utilizó no es rotable, por lo que la precisión de la estimación cambia al moverse en la línea de contorno que produce la meta. Para evitarlo debió usarse un diseño rotable, por ejemplo, un central compuesto.

---

2. MODELOS PARA CONTEOS

En este laboratorio se presenta un ejercicio llamado **Moscas** donde la variable respuesta es un conteo, por lo tanto, sirve para ilustrar el uso de modelos especiales para este tipo de respuestas. Aunque se trata de un diseño experimental con bloques, en un primer análisis se ignora la estructura de bloques con el fin de ilustrar qué se debe hacer cuando existe sobredispersión, para lo cual se hace uso de dos distribuciones: 1) una variante de la distribución Poisson que se va a llamar quasi-Poisson, y 2) la distribución binomial negativa. Al incorporar los bloques la sobredispersión desaparece y se puede analizar usando la distribución Poisson que asume que la variancia condicional es igual a la media condicional.

Se utiliza la librería `AER` (Kleiber & Zeileis, 2008) para verificar el supuesto de equidispersión. Para la estimación de los parámetros se usa la función `glm` cuando se trata de la distribución Poisson o quassi-Poisson, y la función `glm.nb` de la librería `MASS` (Venables & Ripley, 2002) para la binomial negativa.

2.1. Moscas

En un experimento diseñado para ver el tipo de aditivo que se puede poner en una trampa para moscas, se probaron 4 tratamientos: miel, sirope, leche y vinagre.

En diferentes lugares donde había una población importante de moscas se colocaron 4 trampas en cada lugar, una con cada aditivo y se contó el número de moscas atrapadas en cada trampa.

El experimento se repitió durante varios días. Se usaron dos lugares y se repitió durante 4 días, por lo que se tienen 32 observaciones de los conteos para las 4 trampas.

Ejercicios

1. Lea los datos del archivo `moscas.Rdata`. Se han creado dos variables por conveniencia para el factor del tipo de aditivo que se pone en las trampas, una es numérica llamada `trat` y la otra es factor llamada `trat1` que además tiene los nombres de cada nivel del factor (tratamientos).

- Haga el gráfico adecuado para comparar los conteos de moscas por tratamiento. Observe el comportamiento de las medias y de las variancias y compárelas entre tratamientos.
- Justifique por qué en este caso se trata de una distribución de Poisson.

---

2. Utilice un primer modelo de Poisson para el conteo de moscas en función del tipo de aditivo.

- Escriba el modelo que se está proponiendo usando el modelo de suma nula, por lo que debe escribir la restricción del modelo.
- Estime los coeficientes del modelo. Debe usar `family=poisson` en la función `glm`. Observe con `contrasts` cuál es el tratamiento que se está usando como referencia y compruébelo con la matriz de estructura.
- Obtenga las estimaciones de los coeficientes.

---

3. Calcule manualmente los valores ajustados para cada tratamiento. ¿Qué representan estos valores?

- Calcule las medias de la respuesta en cada tratamiento. Compare estos resultados con los promedios estimados con el modelo.
- Obtenga los valores ajustados con la función `predict` usando `type="response"`.

---

4. Obtenga los residuales con la función `residuals` usando `type="response"`.

- Grafique los residuales al cuadrado contra los valores ajustados. Puede poner ambos en logaritmo para visualizar mejor. Agregue la función identidad para ver qué tanto se parecen las medias y las variancias estimadas.

---

5. Obtenga los residuales de Pearson con la función `residuals` usando `type="pearson"` y calcule el parámetro de dispersión. ¿Qué se puede concluir acerca del supuesto de la distribución de Poisson?

- Ajuste de nuevo el modelo tomando en cuenta la sobredispersión. En la función `glm` debe incluir `family=quasipoisson(link=log)`. Observe las estimaciones de los coeficientes y sus errores estándar y compárelos con el del modelo anterior. ¿Qué relación hay entre los errores estándar anteriores y los actuales?
- Obtenga el parámetro de dispersión en ambos modelos usando `summary(mod)$disp`. ¿Tienen sentido estos resultados?
- ¿Está usted de acuerdo con suponer que variancia es proporcional a la media?

---

6. Pruebe que existe un efecto del aditivo sobre el número promedio de moscas. Recuerde que cuando hay sobredispersión y se utiliza la quasi-Poisson, debe usarse la prueba F.

---

7. Usando siempre el modelo de suma nula, escriba los contrastes que se deben usar para definir comparaciones entre pares de promedios

- Usando el modelo del punto 6 que toma en cuenta la sobredispersión, haga los cálculos para obtener las estimaciones de estas comparaciones y explique su significado.
- Haga las pruebas de hipótesis simultáneas para las razones entre todos los pares de tratamientos y determine para cuáles pares de tratamientos se puede concluir que tienen medias diferentes.

---

8. Se sabe que los datos no cumplen el supuesto de una distribución Poisson, donde $E[Y|X] = V[Y|X]$. Se va a utilizar la distribución binomial negativa.

- Escriba el modelo con el factor de diseño y sin usar bloque.
- Estime los parámetros del modelo con binomial negativa. Use la función `glm.nb` en la librería `MASS`.
- Compare los coeficientes y errores estándar de este modelo con el modelo original (`mod1`).

---

9. Haga la prueba formal cuya hipótesis nula es equidispersión ($V[Y|X] = E[Y|X]$). Para esto use la función `dispersiontest` en la librería `AER`. Esta función requiere de un modelo Poisson ajustado con `glm` y una especificación de una hipótesis alternativa mediante el parámetro `trafo`, el cual corresponde a 1 para la quasi-Poisson y 2 para la binomial negativa. Además se usa el parámetro `alternative="greater"` para indicar sobredispersión, sin embargo, no es necesario indicarlo pues este es el default. En el caso de subdispersión se usa `alternative="lower"`.

---

10. Estime la variancia condicional en cada tratamiento con el modelo quasi-Poisson (`mod2`) y con binomial negativa (`mod3`). Compare los resultados de cada modelo con las varianzas observadas y vea cuál las ajusta mejor.

---

11. Hasta ahora se ha ignorado que el experimento se hizo siguiendo una estructura de bloques. Los análisis anteriores sirvieron para comprender el procedimiento, sin embargo, no es correcto hacer el análisis de esa forma. Ahora se hará el análisis correcto y se comparará con lo obtenido anteriormente. Agregue la unidad como un bloque en el modelo inicial (sin sobredispersión). Obtenga el parámetro de dispersión.

    - Haga la prueba del efecto del aditivo pero asumiendo media igual a variancia (sin sobredispersión) por lo que se debe usar la prueba de la razón de las verosimilitudes (LRT).
    - Haga nuevamente las pruebas de hipótesis al comparar pares de razones de medias.
    - Encuentre el límite inferior para las diferencias en aquellos casos donde tiene sentido.

---

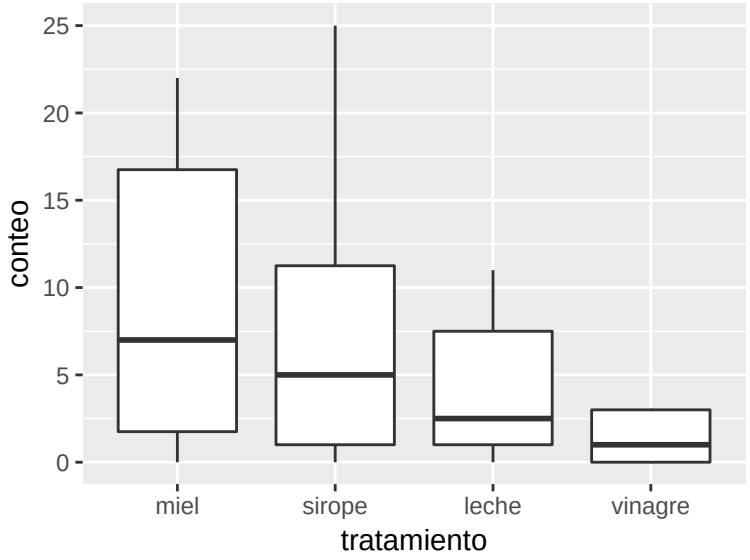
12. Agregue el bloque al modelo con binomial negativa y estime nuevamente los parámetros. Observe el parámetro de dispersión y concluya si es necesario usar la distribución binomial negativa.

Solución

1. Lea los datos del archivo `moscas.Rdata`. Se han creado dos variables por conveniencia para el factor del tipo de aditivo que se pone en las trampas, una es numérica llamada `trat` y la otra es factor llamada `trat1` que además tiene los nombres de cada nivel del factor (tratamientos).

- Haga el gráfico adecuado para comparar los conteos de moscas por tratamiento. Observe el comportamiento de las medias y de las variancias y compárelas entre tratamientos.

```
load("moscas.Rdata")
attach(base)
library(ggplot2)
qplot(trat1,conteo,geom="boxplot",ylab="conteo",xlab="tratamiento")
```



- Justifique por qué en este caso se trata de una distribución de Poisson.

En este caso se cuenta el número de moscas en un momento y lugar específico bajo un cierto tratamiento. Este número teóricamente puede ir de cero a infinito, siendo en general los conteos obtenidos números no muy grandes, por lo que la distribución es genuinamente Poisson y no convendría aproximarla a la normal por el hecho de no tener conteos muy grandes.

---

2. Utilice un primer modelo de Poisson para el conteo de moscas en función del tipo de aditivo.

- Escriba el modelo que se está proponiendo usando el modelo de suma nula, por lo que debe escribir la restricción del modelo.

El modelo es:

$$\log(\lambda_j) = \beta_0 + \tau_j$$

Se establece la restricción:

$$\tau_1 + \tau_2 + \tau_3 + \tau_4 = 0$$

- Estime los coeficientes del modelo. Debe usar `family=poisson` en la función `glm`. Observe con `contrasts` cuál es el tratamiento que se está usando como referencia y compruébelo con la matriz de estructura.

```
options(contrasts=c("contr.sum","contr.poly"))
mod1=glm(conteo~trat1,family=poisson,data=base)

contrasts(trat1)
```

```
##         [,1] [,2] [,3]
## miel       1    0    0
## sirope     0    1    0
## leche      0    0    1
## vinagre   -1   -1   -1
```

```
data.frame(model.matrix(mod1),trat1)[1:12,]
```

```
##    X.Intercept. trat11 trat12 trat13   trat1
## 1             1      1      0      0    miel
## 2             1      0      1      0  sirope
## 3             1      0      0      1   leche
## 4             1     -1     -1     -1 vinagre
## 5             1      1      0      0    miel
## 6             1      0      1      0  sirope
## 7             1      0      0      1   leche
## 8             1     -1     -1     -1 vinagre
## 9             1      1      0      0    miel
## 10            1      0      1      0  sirope
## 11            1      0      0      1   leche
## 12            1     -1     -1     -1 vinagre
```

Se puede observar que vinagre es la referencia, 1:miel, 2:sirope, 3:leche. Esto se comprueba en la matriz de estrucura porque en todas las columnas se le asignó -1 a los datos que son de vinagre.

- Obtenga las estimaciones de los coeficientes.

```
d=mod1$coef
d1=c(d,-sum(d[2:4]))
round(d1,2)
```

```
## (Intercept)      trat11      trat12      trat13
##        1.51        0.72        0.54       -0.06       -1.19
```

---

3. Calcule manualmente los valores ajustados para cada tratamiento. ¿Qué representan estos valores?

```
d2=d1[-1]
lambda=exp(d[1]+d2)
names(lambda)=levels(trat1)
round(lambda,2)
```

```
##    miel  sirope   leche vinagre
##    9.25    7.75    4.25    1.38
```

Estas cantidades representan el número promedio de mosquitos estimado que se pueden atrapar con cada trampa en orden: miel, sirope, leche y vinagre.

- Calcule las medias de la respuesta en cada tratamiento. Compare estos resultados con los promedios estimados con el modelo.

```
tapply(conteo,trat1,mean)
```

```
##    miel  sirope   leche vinagre
##   9.250   7.750   4.250   1.375
```

Las medias estimadas por el modelo son idénticas a las medias observadas por tratamiento. Esto se debe a que se trata de un modelo saturado.

- Obtenga los valores ajustados con la función `predict` usando `type="response"`.

```
fit=predict(mod1,type="response")
tapply(fit,trat1,mean)
```

```
##    miel  sirope   leche vinagre
##   9.250   7.750   4.250   1.375
```

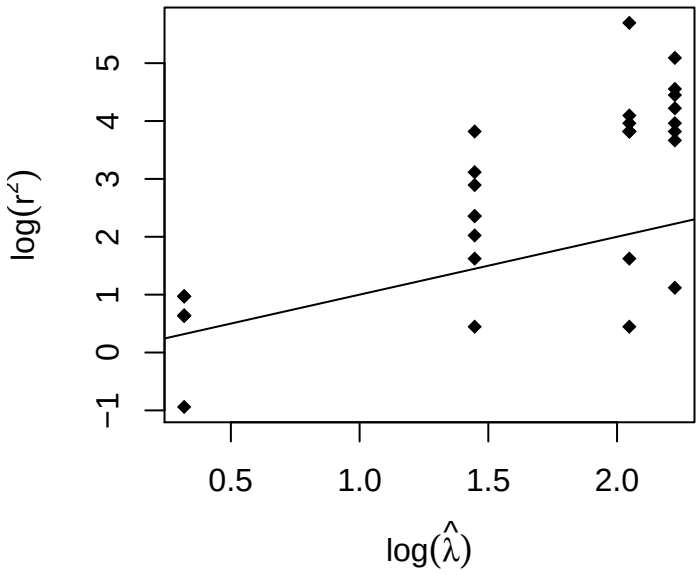
---

4. Obtenga los residuales con la función `residuals` usando `type="response"`.

```
r=residuals(mod1,type="response")
```

- Grafique los residuales al cuadrado contra los valores ajustados. Puede poner ambos en logaritmo para visualizar mejor. Agregue la función identidad para ver qué tanto se parecen las medias y las variancias estimadas.

```
plot(log(r^2)~log(fit),pch=18,xlab=expression(log(hat(lambda))),
     ylab=expression(log(r^2)))
abline(0,1)
```



Se nota que hay sobredispersión porque los valores estimados de la variancia condicional tienden a ser mayores que los de la media condicional (los puntos tienden a estar más arriba de la linea identidad).

---

5. Obtenga los residuales de Pearson con la función `residuals` usando `type="pearson"` y calcule el parámetro de dispersión. ¿Qué se puede concluir acerca del supuesto de la distribución de Poisson?

```
phi=sum(residuals(mod1,type="pearson")^2)/(32-4)
round(phi,1)
```

```
## [1] 6.1
```

Se obtiene un parámetro de dispersión $\Phi = 6{,}1$, el cual es mucho mayor que 1, por lo que se comprueba la sospecha de que existe sobredispersión, por lo tanto, no se cumple el supuesto que dice que la media condicional es igual que la variancia condicional.

- Ajuste de nuevo el modelo tomando en cuenta la sobredispersión. En la función `glm` debe incluir `family=quasipoisson(link=log)`. Observe las estimaciones de los coeficientes y sus errores estándar y compárelos con el del modelo anterior. ¿Qué relación hay entre los errores estándar anteriores y los actuales?

```
mod2=glm(conteo~trat1,family=quasipoisson(link=log),base)
summary(mod2)
```

```
##
## Call:
## glm(formula = conteo ~ trat1, family = quasipoisson(link = log),
##     data = base)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3012  -2.5209  -0.9295   1.2019   4.9050
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.5094     0.2394   6.306 8.07e-07 ***
## trat11        0.7152     0.3140   2.278   0.0306 *
## trat12        0.5383     0.3265   1.649   0.1104
## trat13       -0.0625     0.3837  -0.163   0.8718
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 6.113518)
##
##     Null deviance: 250.72  on 31  degrees of freedom
## Residual deviance: 189.43  on 28  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
ee2=summary(mod2)$coef[,2]
ee1=summary(mod1)$coef[,2]
ee2/ee1
```

```
## (Intercept)      trat11      trat12      trat13
##    2.472553    2.472553    2.472553    2.472553
```

```
sqrt(phi)
```

```
## [1] 2.472525
```

Se puede observar que los coeficientes son los mismos obtenidos anteriormente pero los errores estándar ahora son mayores. Estos son los correctos. El error estándar de cada coeficiente ahora es $\sqrt{\Phi}$ veces el que se había obtenido antes.

- Obtenga el parámetro de dispersión en ambos modelos usando `summary(mod)$disp`. ¿Tienen sentido estos resultados?

```
summary(mod1)$disp
```

```
## [1] 1
```

```
summary(mod2)$disp
```

```
## [1] 6.113518
```

En el primer modelo es 1 porque se asume el modelo Poisson donde la media = variacia. Ese parámetro no se estima, se asume 1. En el segundo modelo se obtiene el valor estimado de Φ.

- ¿Está usted de acuerdo con suponer que variancia es proporcional a la media?

```
med=tapply(conteo,trat1,mean)
v=tapply(conteo,trat1,var)

round(v/med,2)
```

```
##    miel  sirope   leche vinagre
##    8.52   10.20    4.08    1.65
```

En el caso de sirope la variancia es 10.2 veces la media, mientras que en el vinagre la variancia no llega a ser ni el doble de la media. Por lo tanto, no parece que las variancias sean proporcionales a las medias.

---

6. Pruebe que existe un efecto del aditivo sobre el número promedio de moscas. Recuerde que cuando hay sobredispersión y se utiliza la quasi-Poisson, debe usarse la prueba F.

```
drop1(mod2,test="F")
```

```
## Single term deletions
##
## Model:
## conteo ~ trat1
##        Df Deviance F value  Pr(>F)
## <none>      189.43
## trat1   3   250.72  3.0201 0.04633 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cuando hay sobredispersión se usa la prueba F. En este caso se obtine una probabilidad de error tipo I de 0.046$<$ α, por lo que se rechaza la hipótesis nula que dice que la media del número de mosquitos es igual para todos los tratamientos, lo cual se expresa simbólicamente como $H_0 : \lambda_{miel} = \lambda_{sirope} = \lambda_{leche} = \lambda_{vinagre}$. Por lo tanto, se concluye que alguno de los tratamientos está produciendo una media diferente a los demás, es decir, hay un efecto del tratamiento.

---

7. Usando siempre el modelo de suma nula, escriba los contrastes que se deben usar para definir comparaciones entre pares de promedios.

```
med
```

```
##    miel  sirope   leche vinagre
##   9.250   7.750   4.250   1.375
```

```
contrasts(trat1)
```

```
##         [,1] [,2] [,3]
## miel       1    0    0
## sirope     0    1    0
## leche      0    0    1
## vinagre   -1   -1   -1
```

```
cmie =c(1,1,0,0)
csir =c(1,0,1,0)
clec =c(1,0,0,1)
cvin =c(1,-1,-1,-1)
c.ms =cmie-csir
c.ml =cmie-clec
c.mv =cmie-cvin
c.sl =csir-clec
c.sv =csir-cvin
c.lv =clec-cvin
contr =cbind(c.ms,c.ml,c.mv,c.sl,c.sv,c.lv)
contr
```

```
##      c.ms c.ml c.mv c.sl c.sv c.lv
## [1,]    0    0    0    0    0    0
## [2,]    1    1    2    0    1    1
## [3,]   -1    0    1    1    2    1
## [4,]    0   -1    1   -1    1    2
```

- Usando el modelo del punto 6 que toma en cuenta la sobredispersión, haga los cálculos para obtener las estimaciones de estas comparaciones y explique su significado.

```
b=mod2$coef
L=exp(t(contr)%*%b)
row.names(L)=c("miel-sirope","miel-leche","miel-vinagre",
"sirope-leche","sirope-vinagre","leche-vinagre")
round(L,2)
```

```
##                [,1]
## miel-sirope    1.19
## miel-leche     2.18
## miel-vinagre   6.73
## sirope-leche   1.82
## sirope-vinagre 5.64
## leche-vinagre  3.09
```

Estas cantidades representan la razón entre cada par de medias, por ejemplo, la media de miel es 19 % mayor que la media de sirope, la media de miel es 6.7 veces la media de vinagre, etc. Se nota que las que más se diferencian son miel y vinagre, así como sirope y vinagre.

- Haga las pruebas de hipótesis simultáneas para las razones entre todos los pares de tratamientos y determine para cuáles pares de tratamientos se puede concluir que tienen medias diferentes.

Debido a que se usa el modelo quasi-Poisson, se debe utilizar la distribución t en las pruebas. Esta probabilidad debe usar los grados de libertad residuales del modelo. El resultado de la probabilidad debe compararse contra $\alpha/6$.

```
cov=vcov(mod2)
ee=sqrt(diag(t(contr)%*%cov%*%contr))
eta=t(contr)%*%b
qt=eta/ee
p=pt(qt,28,lower.tail = F)
row.names(p)=row.names(L)
round(p,3)
```

```
##                  [,1]
## miel-sirope    0.340
## miel-leche     0.070
## miel-vinagre   0.012
## sirope-leche   0.132
## sirope-vinagre 0.021
## leche-vinagre  0.099
```

Todas las probabilidades son superiores a $\alpha/6 = 0{,}008$ por lo que no se puede decir en cual de las comparaciones hay diferencias.

---

8. Se sabe que los datos no cumplen el supuesto de una distribución Poisson, donde $E[Y|X] = V[Y|X]$. Se va a utilizar la distribución binomial negativa.

- Escriba el modelo con el factor de diseño y sin usar bloque.

$$\log(\lambda_j) = \beta_0 + \tau_j$$

$$\sigma_j^2 = \lambda_j + \frac{\lambda_j^2}{\theta} = \lambda_j + \alpha\lambda_j^2$$

- Estime los parámetros del modelo con binomial negativa. Use la función `glm.nb` en la librería `MASS`.

```
library(MASS)
mod3=glm.nb(conteo~trat1)
```

- Compare los coeficientes y errores estándar de este modelo con el modelo original (`mod1`).

```
round(cbind(summary(mod1)$coef[,1:2],summary(mod3)$coef[,1:2]),3)
```

```
##             Estimate Std. Error Estimate Std. Error
## (Intercept)    1.509      0.097    1.509      0.214
## trat11         0.715      0.127    0.715      0.354
## trat12         0.538      0.132    0.538      0.356
## trat13        -0.063      0.155   -0.063      0.365
```

Los coeficientes estimados en ambos modelos son iguales pero los errores estándar son más grandes en el modelo con binomial negativa.

---

9. Haga la prueba formal cuya hipótesis nula es equidispersión ($V[Y|X] = E[Y|X]$). Para esto use la función `dispersiontest` en la librería `AER`. Esta función requiere de un modelo Poisson ajustado con `glm` y una especificación de una hipótesis alternativa mediante el parámetro `trafo`, el cual corresponde a 1 para la quasi-Poisson y 2 para la binomial negativa. Además se usa el parámetro `alternative="greater"` para indicar sobredispersión, sin embargo, no es necesario indicarlo pues este es el default. En el caso de subdispersión se usa `alternative="lower"`.

```
library(AER)
```

```
dispersiontest(mod1,trafo=1)
```

```
##
##  Overdispersion test
##
## data:  mod1
## z = 3.6318, p-value = 0.0001407
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##    alpha
## 4.349208
```

```
dispersiontest(mod1,trafo=2)
```

```
##
##  Overdispersion test
##
## data:  mod1
## z = 4.7318, p-value = 1.113e-06
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##     alpha
## 0.8012645
```

En el caso de la quasi-Poisson, el parámetro α es lo que usualmente se conoce como Φ, mientras que en el caso de la binomial negativa, α es el coeficiente del término cuadrático de la expresión para la variancia condicional. En ambos casos se rechaza la hipótesis nula $H_0 : \alpha = 0$ por lo que no se puede asumir que la media condicional sea igual a la variancia condicional. Cualquiera de las dos especificaciones alternativas funcionan mejor que la nula.

---

10. Estime la variancia condicional en cada tratamiento con el modelo quasi-Poisson (`mod2`) y con binomial negativa (`mod3`). Compare los resultados de cada modelo con las varianzas observadas y vea cuál las ajusta mejor.

Para quasi-Poisson se tiene que $\sigma_j^2 = \Phi\lambda_j$ y para binomial negativa $\sigma_j^2 = \lambda_j + \alpha\lambda_j^2$.

```
vobs=tapply(conteo,trat1,var)
med=tapply(predict(mod2,type="response"),trat1,mean)
phi=summary(mod2)$disp
vquasi=phi*med
alpha=1/summary(mod3)$theta
vnegbin=med+alpha*med^2
round(cbind(vobs,vquasi,vnegbin),2)
```

```
##          vobs vquasi vnegbin
## miel    78.79  56.55  109.03
## sirope  79.07  47.38   77.79
## leche   17.36  25.98   25.31
## vinagre  2.27   8.41    3.58
```

La variancia de miel es subestimada por el modelo quasi-Poisson y sobrestimada por la binomial negativa, la de sirope está muy cerca la binomial negativa y lejos el quasi-Poisson, la de leche está parecida con las dos y la de vinagre está más cerca con la binomial negativa. No hay ninguna de las dos que realmente aproxime todas las variancias pero es un poco mejor con la binomial negativa.

---

11. Hasta ahora se ha ignorado que el experimento se hizo siguiendo una estructura de bloques. Los análisis anteriores sirvieron para comprender el procedimiento, sin embargo, no es correcto hacer el análisis de esa forma. Ahora se hará el análisis correcto y se comparará con lo obtenido anteriormente. Agregue la unidad como un bloque en el modelo inicial (sin sobredispersión). Obtenga el parámetro de dispersión.

```
mod4=glm(conteo~trat1+as.factor(unid),family=poisson)
sum(residuals(mod4,type="pearson")^2)/(32-11)
```

```
## [1] 0.7951179
```

En este caso el parámetro de dispersión es un poco menor a uno. Ahora el problema de sobredispersión desapareció pero podría estarse dando la situación opuesta: la variancia condicional menor que la media condicional (subdispersión). Para verificar si se está dando subdispersión sería recomendable realizar la prueba de la hipótesis de equidispersión, sin embargo, en este punto no se realizará.

- Haga la prueba del efecto del aditivo pero asumiendo media igual a variancia (sin sobredispersión) por lo que se debe usar la prueba de la razón de las verosimilitudes (LRT).

```
drop1(mod4,test="LRT")
```

```
## Single term deletions
##
## Model:
## conteo ~ trat1 + as.factor(unid)
##                 Df Deviance    AIC     LRT  Pr(>Chi)
## <none>               17.333 122.68
## trat1            3   78.628 177.97  61.295 3.108e-13 ***
## as.factor(unid)  7  189.426 280.77 172.093 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ahora el tratamiento resulta mucho más significativo con una probabilidad de error tipo I bajísima (<0.0001).

- Haga nuevamente las pruebas de hipótesis al hacer razones de medias.

Sólo se toman las covariancias de la parte que interesa. En este caso se usa la distribución normal en las pruebas ya que el modelo es Poisson.

```
b=mod4$coef[1:4]
cov=vcov(mod4)[1:4,1:4]
ee=sqrt(diag(t(contr)%*%cov%*%contr))
eta=t(contr)%*%b
qt=eta/ee
p=pnorm(qt,lower.tail = F)
row.names(p)=row.names(L)
round(p,3)
```

```
##                  [,1]
## miel-sirope     0.152
## miel-leche      0.000
## miel-vinagre    0.000
## sirope-leche    0.002
## sirope-vinagre 0.000
## leche-vinagre   0.001
```

Ahora sí se ven diferencias entre la mayoría de tratamientos. De hecho el único par en el que no se detectan diferencias es entre miel y sirope.

- Encuentre el límite inferior para las diferencias en aquellos casos donde tiene sentido.

Se hacen 5 límites simultáneos y se realiza la corrección de Bonferroni.

```
z=qnorm(1-0.05/5)
LIM=exp(eta[-1]-z*ee[-1])
round(LIM,2)
```

```
## c.ml c.mv c.sl c.sv c.lv
## 1.34 3.17 1.11 2.63 1.38
```

Los tratamientos en los que se ven las mayores diferencias son miel y vinagre (la media de miel es al menos 3.17 veces la de vinagre), así como sirope y vinagre (la media de sirope es al menos 2.63 veces la de vinagre).

---

12. Agregue el bloque al modelo con binomial negativa y estime nuevamente los parámetros. Observe el parámetro de dispersión y concluya si es necesario usar la distribución binomial negativa.

```
mod5=glm.nb(conteo~trat1+factor(unid))
summary(mod5)$theta
```

```
## [1] 35838.56
```

El parámetro de dispersión es 35839, el cual es suficientemente alto como para pensar que $\alpha = 1/\theta \to 0$. Por lo tanto, la distribución binomial negativa tiende a la Poisson y se puede asumir que no hay sobredispersión.

También se puede ver el modelo Poisson con equidispersión y bloque, y verificar que no se rechaza la hipótesis de equidispersión:

```
mod6=glm(conteo~trat1+factor(unid),family=poisson)
dispersiontest(mod6,trafo=2)
```

```
##
##  Overdispersion test
##
## data:  mod6
## z = -2.0088, p-value = 0.9777
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##       alpha
## -0.03712368
```

Conclusión: el análisis correcto de este caso es el que se realizó en el punto 11, ya que se incluyeron los bloques en el modelo, tal como fue diseñado el estudio. La inclusión de los bloques hace que el modelo Poisson se ajuste correctamente puesto que se cumple el supuesto de equidispersión. En este caso no es necesario usar el ajuste quasi-Poisson o la distribución binomial negativa. En este análisis se encotró que hay gran diferencia entre aplicar miel o sirope comparado con aplicar vinagre. Al usar miel o sirope se atrapa una cantidad de moscas mucho mayor en promedio que cuando se aplica vinagre.

---

3. MODELOS DE EFECTOS ALEATORIOS

Los modelos mixtos tienen una amplia gama de aplicaciones. En este laboratorio se presentan 3 ejercicios para introducir los modelos de efectos aleatorios. En el primer ejercicio llamado **Colorantes** se tiene un solo factor aleatorio que son lotes de producción, y se quiere verificar si la variabiliad existente entre los promedios de los lotes es mucho mayor que la variabilidad del error. En el segundo ejercicio llamado **Espectofotómetro** se incluyen dos factores aleatorios para analizar las fuentes de variabilidad presentes. Finalmente se presenta un tecer ejercicio llamado **Escarabajos** donde la respuesta es binaria y se incluyen bloques aleatorios.

Se utiliza la librería `lme4` (Bates et al., 2015) para la estimación de modelos mixtos. Cuando la respuesta es normal se usa la función `lmer`, mientras que cuando la respueta es binaria se usa la función `glmer`. Además se usa la librería `lattice` (Sarkar, 2008) para visualización de los datos.

3.1. Colorantes

En una investigación sobre la elaboración de un colorante se sospecha que la variación del producto final puede estar siendo determinada por variaciones en la calidad de un producto intermedio (ácido H). Se quiere encontrar cuánta de la variación entre lote y lote del ácido H contribuye a la variación en la producción del colorante hecho con este producto. Se analiza la variabilidad total y se separa la variabilidad que introducen los lotes que son materia prima, de tal forma que la variabilidad restante es atribuida al proceso productivo. Por lo tanto, si la variabilidad total está fuertemente determinada por la variabilidad de los lotes, el productor de colorantes debe llamar la atención al proveedor del ácido H para lograr que su producto sea más estable.

Se toman seis lotes de producto intermedio (ácido H) y de cada uno se hacen cinco preparaciones de colorante en el laboratorio. Se determina la producción de cada preparación mediante gramos de color estándar.

Ejercicios

1. Abra el archivo `colorantes.Rdata`.

---

2. Haga un gráfico para ver el comportamiento de la respuesta de lote a lote. Haga un análisis descriptivo. En la librería `ggplot2` puede usar la función `qplot` de la siguiente forma: `qplot(X,Y,geom="boxplot")`.

---

3. Ajuste el modelo mixto con la función `lmer` de la librería `lme4`. El modelo tiene en la parte de efectos fijos solo el intercepto (ponga un uno pero no es indispensable) y en la parte de efectos aleatorios el lote `(1|lote)` – en esta parte el 1 representa el promedio: `mod1=lmer(prod~1+(1|lote))` es equivalente a `mod1=lmer(prod~(1|lote))`.

- Obtenga del `summary` del `mod1` las estimaciones de las variancias correspondientes a este modelo.
- Asegúrese que puede obtener manualmente esas estimaciones a partir del anova de un modelo lineal.

---

4. Ajuste de nuevo el modelo pero con máxima verosimilitud (R lo hace por default usando la máxima verosimilitud restringida). En este caso es necesario hacerlo con máxima verosimilitud para hacer los perfiles que son pruebas de máxima verosimilitud. Basta agregar la instrucción `REML=F` dentro del `lmer`: `mod1a=lmer(prod~1+(1|lote),base,REML=F)`.

- Obtenga los intervalos de confianza para los parámetros del modelo (por default R usa un nivel de confianza de 0.95): `confint(profile(mod1a),level=0.95)`. En el resultado se debe interpretar `sigma` como la desviación estándar del error y `sigma01` la desviación estándar de lote.

---

5. En los modelos mixtos, los efectos aleatorios no son de interés en sí mismos, sino que lo interesa es su variancia. Sin embargo, es posible Obtener una estimación de estos efectos para tener una idea de su comportamiento. Obtenga las estimaciones de los efectos aleatorios con `ranef(mod1)`.

Solución

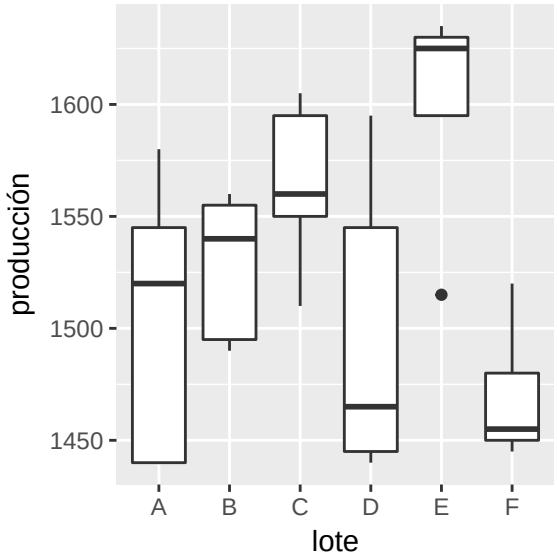
1. Abra el archivo `colorantes.Rdata`.

```
load("colorantes.Rdata")
attach(base)
```

---

2. Haga un gráfico para ver el comportamiento de la respuesta de lote a lote. Haga un análisis descriptivo. En la librería `ggplot2` puede usar la función `qplot` de la siguiente forma: `qplot(X,Y,geom="boxplot")`.

```
library(ggplot2)
qplot(lote,prod,geom="boxplot",ylab="producción")
```



Hay algunos lotes que tienen producciones muy bajas mientras que otros las tienen muy altas, aunque no interesa identificar cuáles son los lotes con producción más baja o más alta puesto que los lotes son aleatorios.

---

3. Ajuste el modelo mixto con la función `lmer` de la librería `lme4`. El modelo tiene en la parte de efectos fijos solo el intercepto (ponga un uno pero no es indispensable) y en la parte de efectos aleatorios el lote `(1|lote)` – en esta parte el 1 representa el promedio: `mod1=lmer(prod~1+(1|lote))` es equivalente a `mod1=lmer(prod~(1|lote))`.

```
library(lme4)
mod1=lmer(prod~(1|lote))
```

- Obtenga del `summary` del **mod1** las estimaciones de las variancias correspondientes a este modelo.

```
summary(mod1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: prod ~ (1 | lote)
##
## REML criterion at convergence: 319.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.4117 -0.7634  0.1418  0.7792  1.8296
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  lote     (Intercept) 1764     42.00
##  Residual             2451     49.51
## Number of obs: 30, groups:  lote, 6
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  1527.50      19.38    78.8
```

A partir de la descomposición de la variancia se obtiene el porcentaje de la variabilidad debida a las diferencias producidas por los lotes. La variancia de las medias de lote a lote es 1764, mientras que la la variancia residual es 2451. De esta forma se obtiene el porcentaje:

```
1764/(1764+2451)*100
```

```
## [1] 41.85053
```

El 41.8 % de la variabilidad es debida a los lotes.

- Asegúrese que puede obtener manualmente esas estimaciones a partir del anova de un modelo lineal.

```
mod2=lm(prod~lote)
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: prod
##           Df Sum Sq Mean Sq F value   Pr(>F)
## lote       5  56358 11271.5  4.5983 0.004398 **
## Residuals 24  58830  2451.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se parte de las esperanzas de los cuadrados medios:

$E[CMLote] = \sigma^2_\epsilon + n\sigma^2_L$

$E[CMRes] = \sigma^2_\epsilon$

Se empieza estimando σ^2_ϵ con el CMRes y se sustituye en la primera ecuación:

$CMLote = CMRes + n\sigma^2_L$

Por lo tanto:

$\hat{\sigma^2_L} = \frac{CMLote - CMRes}{n}$

```
(11271.5-2451.2)/5
```

```
## [1] 1764.06
```

---

4. Ajuste de nuevo el modelo pero con máxima verosimilitud (R lo hace por default usando la máxima verosimilitud restringida). En este caso es necesario hacerlo con máxima verosimilitud para hacer los perfiles que son pruebas de máxima verosimilitud. Basta agregar la instrucción `REML=F` dentro del `lmer`: `mod1a=lmer(prod~1+(1|lote),base,REML=F)`.

```
mod1a=lmer(prod~1+(1|lote),REML=FALSE)
```

- Obtenga los intervalos de confianza para los parámetros del modelo (por default R usa un nivel de confianza de 0.95): `confint(profile(mod1a),level=0.95)`. En el resultado se debe interpretar `sigma` como la desviación estándar del error y `sigma01` la desviación estándar de lote.

```
confint(profile(mod1a))
```

```
##                   2.5 %     97.5 %
## .sig01         12.19854   84.06305
## .sigma         38.22998   67.65770
## (Intercept) 1486.45150 1568.54849
```

Puesto que el intervalo de 95 % de confianza va de 12.2 a 84.1 (no llega al extremo de cero), se confirma que la fuente de variabilidad de lote a lote no es nula.

---

5. En los modelos mixtos, los efectos aleatorios no son de interés en sí mismos, sino que lo interesa es su variancia. Sin embargo, es posible Obtener una estimación de estos efectos para tener una idea de su comportamiento. Obtenga las estimaciones de los efectos aleatorios con `ranef(mod1)`.

```
ranef(mod1)
```

```
## $lote
##   (Intercept)
## A -17.6068514
## B   0.3912634
## C  28.5622255
## D -23.0845384
## E  56.7331877
## F -44.9952868
```

Hay algunos lotes que tienen promedios inferiores a la media general (A, D, F), mientras que otros tienen promedios que están bastante por encima (C, E). Esto va a favor del resultado anterior que indica que hay una variabilidad importante de lote a lote, es decir, que los promedios de los diferentes lotes no son iguales en general.

Conclusión: se tiene una estimación de la variancia de los lotes que representa el 41.8 % de la variabilidad total, lo cual es una indicación bastante fuerte de que la estabilidad en la producción del colorante está siendo bastante afectada por la materia prima. Si los lotes de ácido H fueran más estables se tendría un colorante más estable también.

3.2. Espectofotómetro

Se está desarrollando un nuevo modelo de espectrofotómetro para uso en laboratorios clínicos con el objetivo de cuantificar sustancias y microorganismos. Se quiere evaluar el funcionamiento de estos instrumentos sabiendo que un componente crítico del desempeño es la consistencia de las mediciones de un día a otro, y de una máquina a otra. Se quiere saber si la variabilidad de las mediciones entre las máquinas operadas durante varios días está dentro de los estándares aceptables.

Se seleccionan aleatoriamente 4 máquinas etiquetadas como A, B, C y D. Cada día se preparan 8 replicaciones de muestras de suero en sangre con el mismo lote de reactivos. Dos muestras de suero se asignan aleatoriamente a cada una de las cuatro máquinas en cada uno de los 4 días para un diseño completamente al azar con dos repeticiones de cada tratamiento.

Se miden los niveles de triglicéridos (mg/dl) en las muestras de suero. Los datos se muestran a continuación:

Día	**A**	**A**	**B**	**B**	**C**	**C**	**D**	**D**
1	142.3	144.0	148.6	146.9	142.9	147.4	133.8	133.2
2	134.9	146.3	145.2	146.3	125.9	127.6	108.9	107.5
3	148.6	156.5	148.6	153.1	135.5	138.9	132.1	149.7
4	152.0	151.4	149.7	152.0	142.9	142.3	141.7	141.2

Ejercicios

1. Introduzca los datos. Defina `dia` y `maquina` como factor.

---

2. Haga gráficos para ver el comportamiento de la respuesta de día a día y de máquina a máquina. Haga un análisis descriptivo.

---

3. Ajuste el modelo que tiene en la parte de efectos fijos solo el intercepto y en la parte de efectos aleatorios el día `(1|dia)` y la máquina `(1|maq)`. Se podría incluir la interacción `(1|dia:maq)`, sin embargo, en este caso no se hace pues esta interacción no tiene sentido desde el punto de vista teórico.

- Obtenga las estimaciones de las variancias correspondientes a este modelo tanto en R como manualmente.
- Obtenga e interprete los intervalos de 95 % de confianza para las desviaciones estándar. La estimaciones van en el orden en que aparecen en el summary, es decir, desviación estándar de máquinas (`sig01`), desviación estándar de días (`sig03`) y desviación estándar de error (`sigma`).

Solución

1. Introduzca los datos. Defina `dia` y `maquina` como factor.

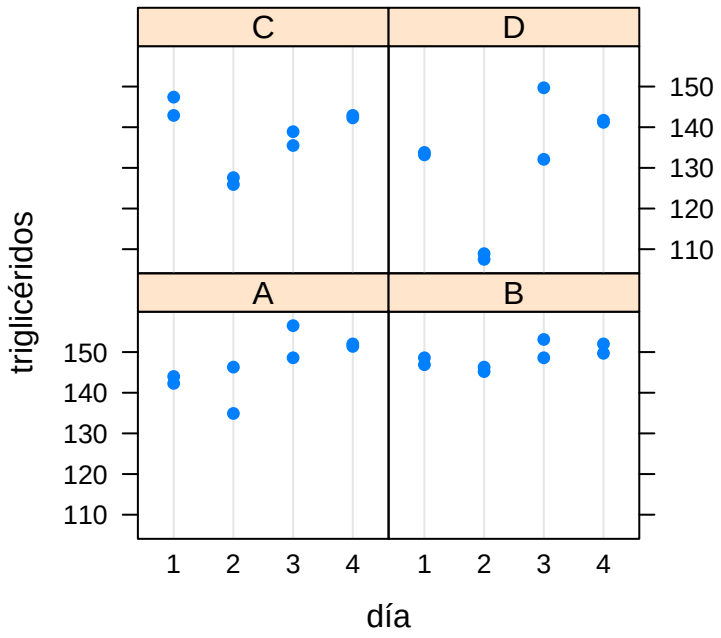
```
trigli=c(142.3, 144.0, 148.6, 146.9, 142.9, 147.4, 133.8,
         133.2, 134.9, 146.3, 145.2, 146.3, 125.9, 127.6,
         108.9, 107.5, 148.6, 156.5, 148.6, 153.1, 135.5,
         138.9, 132.1, 149.7, 152.0, 151.4, 149.7, 152.0,
         142.9, 142.3, 141.7, 141.2)
dia=factor(rep(1:4,each=8))
maq=factor(rep(c("A","B","C","D"),4,each=2))
```

---

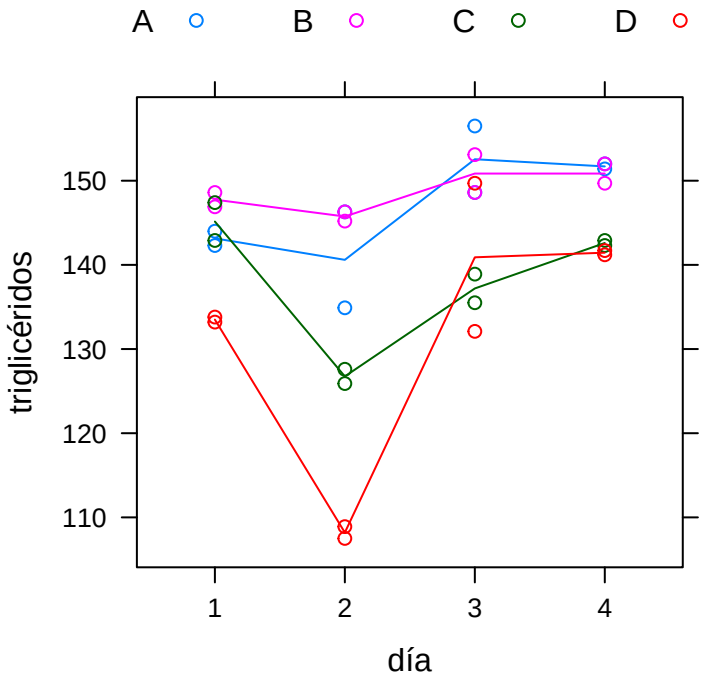
2. Haga gráficos para ver el comportamiento de la respuesta de día a día y de máquina a máquina. Haga un análisis descriptivo.

```
library(lattice)
```

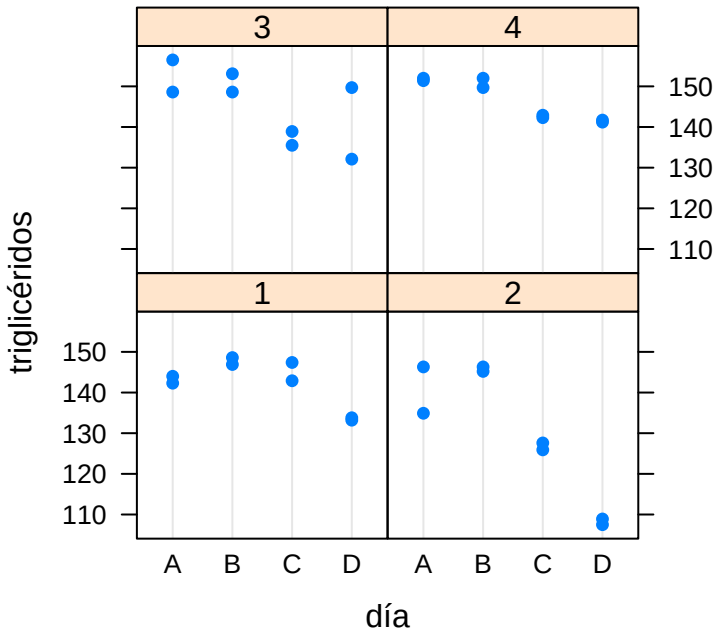
```
dotplot(trigli~dia|maq,xlab="día",ylab="triglicéridos")
```



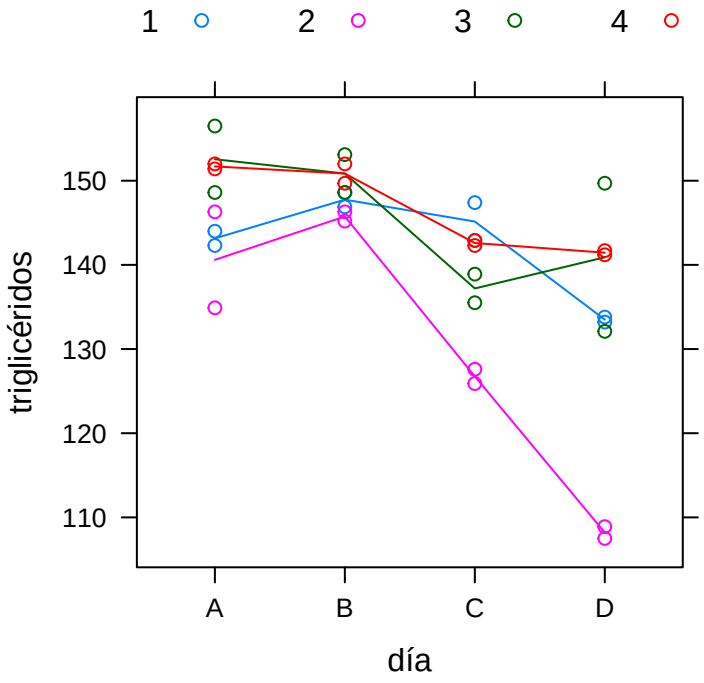
```
xyplot(trigli~dia,groups=maq,type=c("p","a"),xlab="día",
       ylab="triglicéridos",auto.key=list(columns=4))
```




```
dotplot(trigli~maq|dia,xlab="día",ylab="triglicéridos")
```



```
xyplot(trigli~maq,groups=dia,type=c("p","a"),xlab="día",
       ylab="triglicéridos",auto.key=list(columns=4))
```



En los dos primeros gráficos se pueden observar la diferencias de día a día dentro de cada máquina, las que no son muy fuertes, excepto en el día 2 en la máquina D. En el tercero y cuarto gráfico se ven las diferencias de máquina a máquina cada día. Las mayores diferencias se presentan en el día 2.

A partir de los gráficos parece que existe una interacción entre día y máquina, sin embargo, esta interacción podría venir de una situación particular en lugar de tratarse de algo general, pues se vio que en el día 2 la máquina D dio resultados de triglicéridos más bajos de lo normal. Esta observación puede ayudar a comprender el panorama general, si bien se ha dicho que en este tipo de análisis no interesa tanto identificar algo particular.

3. Ajuste el modelo que tiene en la parte de efectos fijos solo el intercepto y en la parte de efectos aleatorios el día `(1|dia)` y la máquina `(1|maq)`. Se podría incluir la interacción `(1|dia:maq)`, sin embargo, en este caso no se hace pues esta interacción no tiene sentido desde el punto de vista teórico.

```
mod1=lmer(trigli~1+(1|dia)+(1|maq))
```

- Obtenga las estimaciones de las variancias correspondientes a este modelo tanto en R como manualmente.

```
summary(mod1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: trigli ~ 1 + (1 | dia) + (1 | maq)
##
## REML criterion at convergence: 222.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2132 -0.4806 -0.1114  0.5861  2.1538
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  dia      (Intercept) 50.24    7.088
##  maq      (Intercept) 63.27    7.955
##  Residual             42.89    6.549
## Number of obs: 32, groups:  dia, 4; maq, 4
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  141.184      5.452    25.9
```

```
vartot=50.24+63.27+42.89
63.27/vartot*100
```

```
## [1] 40.45396
```

```
50.24/vartot*100
```

```
## [1] 32.12276
```

Máquina tiene una variancia de 63.27 que representa un 40.5 % de la variancia total y día una variancia de 50.24 (32.1 %).

- Obtenga e interprete los intervalos de 95 % de confianza para las desviaciones estándar. La estimaciones van en el orden en que aparecen en el summary, es decir, desviación estándar de máquinas (`sig01`), desviación estándar de días (`sig03`) y desviación estándar de error (`sigma`).

```
mod2=lmer(trigli~1+(1|dia)+(1|maq),REML=FALSE)
confint(profile(mod2))
```

```
##                   2.5 %     97.5 %
## .sig01         3.012755  17.244008
## .sig02         3.515762  18.757479
## .sigma         5.084725   8.906165
## (Intercept) 129.709345 152.659461
```

Al observar el intervalo de 95 % de confianza se tiene que la desviación estándar de máquina a máquina está entre 3.0 y 17.2, mientras que la desviación estándar de día a día está entre 3.5 y 18.8. Esto confirma que estas fuentes de variabilidad son importantes. Por lo tanto, hay evidencia que hay fuentes de variabilidad que deben revisarse para que este moldeo de espectofotómetro funcione mejor. Se esperaría mayor consitencia entre máquina, así como de uno a otro día.

Conclusión: el espectofotómetro no está listo para tomar las mediciones de forma precisa ya que estas no son consistentes de una máquina a otra puesto que la variabilidad asociada es cerca del 40 % de la variabilidad total, de igual forma de un día a otro no hay consistencia puesto que esta fuente de variación representa un 32 % de la variabilidad total. Sería importante repetir el experimento puesto que parece que una de las máquinas tuvo un comportamiento especial en el día 2. Dada esta sospecha sería importante verificar si al tener más máquinas y más días en el experimento se repite este tipo de comportamiento.

3.3. Escarabajos

En un estudio se quería identificar cuál tipo de recolecta es mejor para registrar la diversidad de sexo en una especie de escarabajos. Se realizaron recolectas de escarabajos nocturnos con dos métodos: 1) usando alumbrado público y 2) utilizando lámparas. Los muestreos se hicieron sistemáticamente (hora y media a partir de la penumbra) durante los meses de abril a junio en cinco parcelas de cultivo en la zona de San Cristóbal de Las Casas, Chiapas. Se registró el sexo de los escarabajos recolectados para determinar si en un tipo de recolecta la proporción de cada sexo era diferente.

Ejercicios

1. Cargue los datos en el archivo `escarabajos.Rdata`. Defina `parcela` como factor.

- Identifique la variable respuesta y comente sobre la distribución condicional que tiene esta variable.
- Justifique si las parcelas representan un efecto fijo o aleatorio.
- Represente gráficamente los datos y observe si se puede esperar un efecto del método de recolecta.

---

2. Tome las parcelas como bloques fijos y ajuste el modelo logístico con `glm`. Ponga atención al modelo que está usando en R (suma nula o tratamiento referencia).

- Interprete la razón de propensiones (OR) resultante. Debe tener claro cuál probabilidad es la que está obteniendo (de machos o hembras). Esto dependerá de la forma en que especificó su modelo.

---

3. Realice la prueba `LRT` para determinar si existe un efecto del tipo de recolecta en la proporción de machos (o hembras) capturados.

---

4. Ahora tome la parcela como un efecto aleatorio y ajuste el modelo usando la función `glmer` de la librería `lme4`. Obtenga el OR y compárelo con el obtenido anteriormente.

- Realice la prueba adecuada para determinar si existe un efecto del tipo de recolecta.
- Construya intervalos de 95 % para el OR obtenido con cada modelo (con bloques fijos y con efectos aleatorios). Compárelos.

Solución

1. Cargue los datos en el archivo `escarabajos.Rdata`. Defina `parcela` como factor.

```
load("escarabajos.Rdata")
base$parcela=factor(base$parcela)
attach(base)
```

- Identifique la variable respuesta y comente sobre la distribución condicional que tiene esta variable.

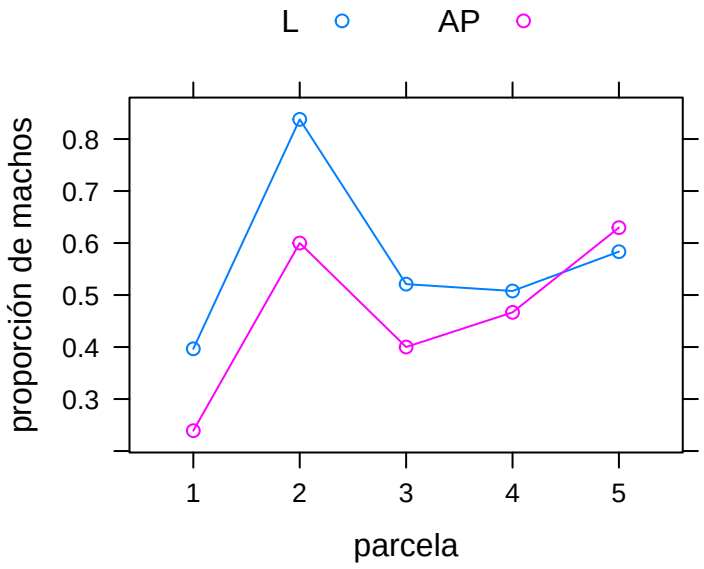
La variable respuesta es el número de escarabajos machos en una parcela. Esta variable tiene una distribución binomial ya que el número de escarabajos machos en una parcela en particular está acotado por el número de escarbajos capturados en esa parcela. La distribución de esta variable tiene dos parámetros: 1) la probabilidad de atrapar machos en la parcela para un método de recolecta en particular, y 2) el número de escarabajos atrapados en la parcela con ese método específico.

- Justifique si las parcelas representan un efecto fijo o aleatorio.

Las parcelas son un factor aleatorio porque los resultados no interesan solo para esas parcelas, sino que estas son una muestra entre muchas parcelas.

- Represente gráficamente los datos y observe si se puede esperar un efecto del método de recolecta.

```
xyplot(M/(M+H)~parcela,groups=tipo,type=c("p","a"),
       auto.key=list(columns=2),ylab="proporción de machos")
```



El método de lampareo consistentemente hace que se recolecte una mayor proporción de machos que el método de alumbrado público.

2. Tome las parcelas como bloques fijos y ajuste el modelo logístico con `glm`. Ponga atención al modelo que está usando en R (suma nula o tratamiento referencia).

```
options(contrasts=c("contr.sum","contr.poly"))
contrasts(tipo)
```

```
##    [,1]
## L     1
## AP   -1
```

```
mod1=glm(cbind(M,H)~tipo+parcela,family="binomial")
summary(mod1)
```

```
##
## Call:
## glm(formula = cbind(M, H) ~ tipo + parcela, family = "binomial")
##
## Deviance Residuals:
##       1        2        3        4        5        6        7        8
##  0.1256   0.7123  -0.0218  -0.2272  -0.8831  -0.5009  -0.8052   0.1767
##       9       10
##  0.6751   0.5652
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.1068     0.1102   0.969 0.332732
## tipo1         0.2783     0.1041   2.673 0.007512 **
## parcela1     -0.8153     0.1245  -6.551 5.72e-11 ***
## parcela2      0.9501     0.2614   3.635 0.000278 ***
## parcela3     -0.3000     0.1168  -2.568 0.010223 *
## parcela4     -0.3137     0.1667  -1.882 0.059809 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 64.2552  on 9  degrees of freedom
## Residual deviance:  3.0608  on 4  degrees of freedom
## AIC: 59.706
##
## Number of Fisher Scoring iterations: 4
```

Se está usando el modelo de suma nula puesto que el tratamiento de referencia que es AP está codificado con -1.

- Interprete la razón de propensiones (OR) resultante. Debe tener claro cuál probabilidad es la que está obteniendo (de machos o hembras). Esto dependerá de la forma en que especificó su modelo.

Puesto que en el modelo se especificó `cbind(M,H)`, la probabilidad obtenida es la de machos. Además, para obtener el OR se debe usar simplemente 2 veces el coeficiente de tipo por tratarse de un modelo de suma nula.

```
b=mod1$coef
OR=exp(2*b[2])
round(OR,2)
```

```
## tipo1
##  1.74
```

En una parcela específica, la propensión de encontrar un macho cuando se usa alumbrado público es 74 % mayor que cuando se usa lampareo.

---

3. Realice la prueba `LRT` para determinar si existe un efecto del tipo de recolecta en la proporción de machos (o hembras) capturados.

```
mod2=glm(cbind(M,H)~parcela,family="binomial")
anova(mod2,mod1,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(M, H) ~ parcela
## Model 2: cbind(M, H) ~ tipo + parcela
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         5    10.4913
## 2         4     3.0608  1   7.4305 0.006413 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En la prueba de razón de verosimilitud se observa que la probabilidad asociada al tipo de recolecta es suficientemente baja, y se puede rechazar la hipótesis nula de que ese factor no tiene ningún efecto sobre la proporción de machos atrapados. Por lo tanto, sí se puede esperar un cambio en la probabilidad de atracción de machos según el tipo de luz que se utilice.

---

4. Ahora tome la parcela como un efecto aleatorio y ajuste el modelo usando la función `glmer` de la librería `lme4`. Obtenga el OR y compárelo con el obtenido anteriormente.

```
mod3=glmer(cbind(M,H)~tipo+(1|parcela),family="binomial")
b=summary(mod3)$coef
b
```

```
##               Estimate Std. Error   z value   Pr(>|z|)
## (Intercept) 0.06282529  0.2742138 0.2291107 0.81878291
## tipo1       0.24406520  0.1029204 2.3713972 0.01772098
```

```
OR=exp(2*b[2,1])
round(OR,2)
```

```
## [1] 1.63
```

Ahora el OR es 1.63 el cual es similar al obtenido cuando los bloques se tomaron como fijos, sin embargo, es un poco menor.

- Realice la prueba adecuada para determinar si existe un efecto del tipo de recolecta.

En la parte de efectos fijos del anova sale la probabilidad asociada al tipo de recolecta para la prueba de Wald (con una z). Esta probabilidad es 0.018 con la cual se tiene evidencia para rechazar la hipótesis nula sobre la ausencia de efecto del tipo de recolecta.

Otra forma de hacer esta prueba es comparando este modelo con uno sin el tipo de recolecta y usando LRT. En este caso la probabilidad asociada es muy parecida (0.016) y se llega a la misma conclusión.

```
mod4=glmer(cbind(M,H)~(1|parcela),family="binomial")
anova(mod4,mod3)
```

```
## Data: NULL
## Models:
## mod4: cbind(M, H) ~ (1 | parcela)
## mod3: cbind(M, H) ~ tipo + (1 | parcela)
##      Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## mod4  2 77.050 77.655 -36.525   73.050
## mod3  3 73.223 74.131 -33.612   67.223 5.8267      1    0.01578 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Construya intervalos de 95 % para el OR obtenido con cada modelo (con bloques fijos y con efectos aleatorios). Compárelos.

```
mod5=glmer(cbind(M,H)~tipo+(1|parcela),family="binomial")
ci4=confint(mod5)
round(exp(2*ci4[3,]),3)
```

```
##  2.5 % 97.5 %
##  1.095  2.459
```

```
ci=confint(mod1)
round(exp(2*ci[2,]),3)
```

```
##  2.5 % 97.5 %
##  1.167  2.646
```

Los intervalos son similares pero no idénticos, y en ambos casos se concluye que el OR es mayor que 1.1 y menor que 2.65. Particularmente el intervalo con efectos aleatorios es un poco más bajo que el que usó bloques.

Conclusión: el objetivo del estudio era determinar si el tipo de recolecta podría llevar a resultados diferentes en cuanto a la composición por sexos de esta especie. El investigador está interesado en determinar la composición real, sin embargo, queda claro que según el método que utilice, la probabilidad de obtener un macho o una hembra se verá alterada, lo cual puede dar una estimación inadecuada de la composición real según el método que se utilice. No queda claro cuál de los dos métodos produce estimaciones más cercanas a la realidad.

---

4. DISEÑOS ANIDADOS

En este laboratorio se presentan dos ejercicios con factores anidados. Los dos ejercicios tienen una estructura similar. En el primer ejercicio llamado **Escuelas**, el factor fijo `instructor` está anidado dentro del factor fijo `escuela`, mientras que en el segundo llamado **Pastas**, el factor aleatorio `barrica` está anidado dentro del factor aleatorio `lote`.

Cuando los factores son fijos se puede usar la función `lm` con una instrucción especial para indicar el anidamiento, y cuando los factores son aleatorios se utiliza la función `lmer` de la librería `lme4` (Bates et al., 2015).

4.1. Escuelas

Una compañía manufacturera tiene tres escuelas de mecánica regionales, una en cada una de sus regiones de operación. Cada escuela tiene dos instructores que dan cursos de tres semanas a 15 mecánicos aproximadamente. La compañía está interesada en conocer el efecto de la escuela (factor A) y el instructor (factor B) en el aprendizaje logrado. Se hace un experimento en el que se forman grupos en cada región y se asigna cada grupo a uno de los instructores. A cada instructor se le asignan dos grupos.

Para determinar el efecto de la escuela y el instructor en el aprendizaje, se hace una prueba a los estudiantes y se obtiene el puntaje promedio del grupo como variable respuesta.

Ejercicios

1. Lea los datos del archivo `escuelas.Rdata`. Defina `escuela` como factor. Ponga nombres a las escuelas: `Región K` (1), `Región D` (2) y `Central` (3). Observe que hay dos variables para referirse a los instructores: `instructor` tiene valores 1 y 2 para los instructores de todas las escuelas, mientras que `instructor1` no repite el número del instructor en diferentes escuelas, por lo que los instructores van del 1 al 6. Redefina esas dos variables como factor.

- Haga una representación gráfica de los datos para ver el comportamiento de la respuesta según escuela e instructor. Use la función `dotplot` en la librería `lattice` de la siguiente forma: `dotplot(puntaje ~ instructor | escuela)`. Hágalo usando tanto `instructor` como `instructor1` y vea cuál de las dos formas es más conveniente.
- ¿Se puede esperar que exista interacción entre el instructor y la escuela?
- Comente si se espera que haya un efecto del instructor sobre el puntaje promedio dentro de cada escuela?
- Busque una forma de analizar gráficamente las diferencias entre escuelas. ¿Qué se puede concluir?

2. Efectos de instructor:

- Obtenga manualmente los efectos de instructor dentro de escuela.
- Obtenga manualmente la suma de cuadrados de instructor dentro de escuela.
- ¿Cuántos grados de libertad tiene el factor anidado instructor?
- Calcule el cuadrado medio de instructor dentro de escuela y explique su significado.

---

3. Comparación de instructores.

- Establezca con palabras la hipótesis nula al comparar los instructores.
- Haga el análisis de variancia. Introduzca `escuela` e `instructor` pero no agregue interacción. Note que instructor no resulta significativo y que tiene únicamente un grado de libertad cuando en realidad hay 6 instructores distintos.
- Estime de nuevo el mismo modelo pero usando `instructor1` y llámelo `mod2`. ¿Qué es lo que está pasando? ¿Cuál es el análisis correcto, con `instructor` o con `instructor1`?
- Escriba el modelo con `aov` y haga de nuevo el análisis de variancia. Introduzca como factores `escuela` e `instructor` pero agregue la interacción. Observe que ahora instructor sí da significativo, lo mismo que la interacción, pero persiste el problema de los grados de libertad.
- Use `model.tables(mod)` para obtener las estimaciones de los efectos. Observe que hay un efecto para el instructor 1 y un efecto para el instructor 2, cuando en realidad son 6 instructores diferentes.
- Ajuste nuevamente el modelo pero usando `instructor1`, llámelo `mod4`. Observe los efectos de instructor y compárelos con los que se obtienen de `mod2`.

---

4. Uso del factor anidado:

- Se puede usar el factor anidado con el formato en que aparece `instructor`, pero indicando la estructura anidada dentro de la función `lm` de la siguiente forma: `lm(puntaje~escuela+escuela/instructor)`. De esta forma se indica que el factor `instructor` está anidado dentro del factor `escuela`. Use el modelo de suma nula y llámelo `mod5`.
- Obtenga el anova del modelo anidado y note que tiene una línea que parece una interacción que dice `escuela:instructor`, pero que en realidad debería leerse `instructor(escuela)`, es decir, instructor dentro de escuela.
- Observe el cuadrado medio de instructor y compárelo con el obtenido anteriormente. ¿Se rechaza la hipótesis nula referente al instructor?
- Obtenga los efectos de instructor y compárelos con los anteriores.

---

5. Comparaciones entre instructores:

- Obtenga los coeficientes del modelo. Hágalo con `mod2`, `mod4` y `mod5`. Observe que `mod2` y `mod4` tienen la desventaja de que hay muchos NA, mientras que `mod5` da solo los coeficientes que se han estimado.
- ¿A qué corresponden los últimos 3 coeficientes en `mod5`?
- Haga las comparaciones que tengan sentido entre los instructores. No tiene sentido comparar todos los pares de instructores, puesto que no todos pertenecen a las mismas escuelas. Por ejemplo, no tiene sentido comparar un instructor de la Región D con un instructor de la Región K. Use la corrección de Bonferroni con un nivel de significancia de 0.10 y confianza de 90 %.

---

6. Haga las comparaciones entre las diferentes escuelas usando Tukey con 10 % de significancia en pruebas de una cola.

- Obtenga cotas inferiores de 90 % de confianza para la diferencia de medias entre los pares de escuelas en que encontró diferencias.

---

7. ¿Cómo se alteran las conclusiones si se considera que hay muchos instructores en cada región y en el experimento los que participaron fueron una muestra?

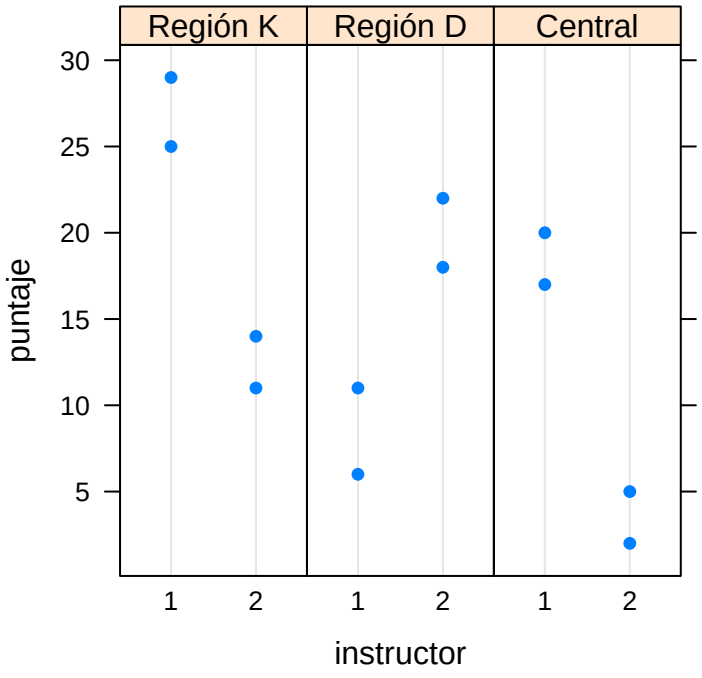
Solución

1. Lea los datos del archivo `escuelas.Rdata`. Defina `escuela` como factor. Ponga nombres a las escuelas: `Región K` (1), `Región D` (2) y `Central` (3). Observe que hay dos variables para referirse a los instructores: `instructor` tiene valores 1 y 2 para los instructores de todas las escuelas, mientras que `instructor1` no repite el número del instructor en diferentes escuelas, por lo que los instructores van del 1 al 6. Redefina esas dos variables como factor.

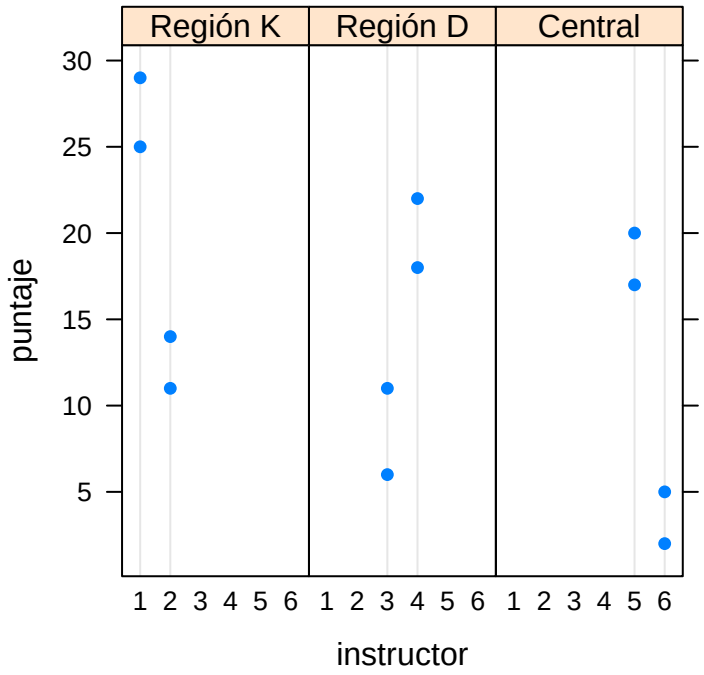
```
load("escuelas.Rdata")
base$escuela=factor(base$escuela)
base$instructor=factor(base$instructor)
base$instructor1=factor(base$instructor1)
levels(base$escuela)=c("Región K","Región D","Central")
attach(base)
```

- Haga una representación gráfica de los datos para ver el comportamiento de la respuesta según escuela e instructor. Use la función `dotplot` en la librería `lattice` de la siguiente forma: `dotplot(puntaje ~ instructor | escuela)`. Hágalo usando tanto `instructor` como `instructor1` y vea cuál de las dos formas es más conveniente.

```
library(lattice)
dotplot(puntaje ~ instructor | escuela,xlab="instructor",
        layout = c(3,1))
```



```
dotplot(puntaje ~ instructor1| escuela,xlab="instructor",
        layout = c(3,1))
```



Es más conveniente usar instructor porque la otra forma guarda el campo para todos los instructores en todas las escuelas, lo cual no tiene sentido.

- ¿Se puede esperar que exista interacción entre el instructor y la escuela?

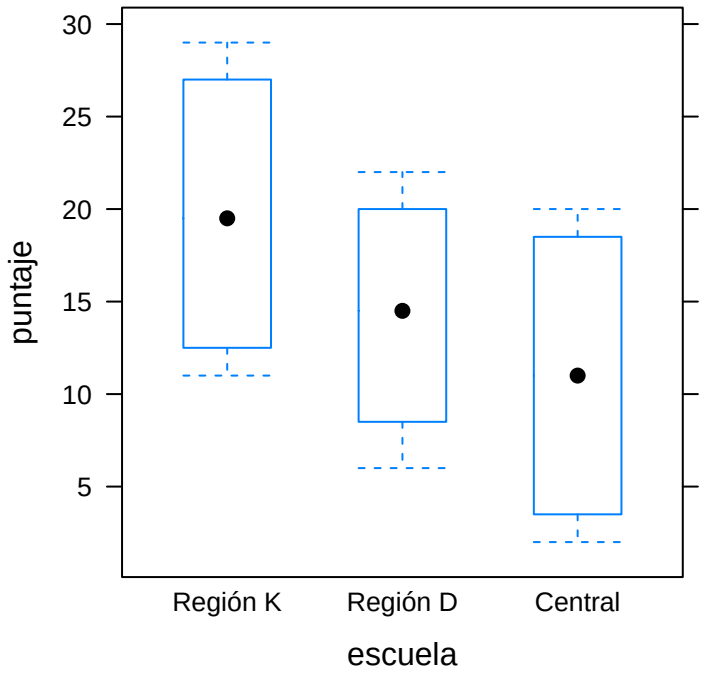
No se analiza la interacción cuando el diseño es anidado ya que los instructores que están en una escuela no son los mismos que están en otra escuela.

- Comente si se espera que haya un efecto del instructor sobre el puntaje promedio dentro de cada escuela?

A partir de este gráfico, sí parece haber diferencias entre los puntajes promedio de los diferentes instructores dentro de cada escuela.

- Busque una forma de analizar gráficamente las diferencias entre escuelas. ¿Qué se puede concluir?

```
bwplot(puntaje~escuela,xlab="escuela",ylab="puntaje")
```



Se nota que algunas escuelas tienen puntajes promedio más altos que otras, por lo que sí se esperaría un efecto de la escuela en el aprendizaje. En particular, se notan mayores diferencias entre la escuela Central y Región K. Sin embargo, aquí no se está aislando la variabilidad que producen los instructores, lo que hace que esas diferencias no sean tan claras.

2. Efectos de instructor:

- Obtenga manualmente los efectos de instructor dentro de escuela.

```
medinst=tapply(puntaje,list(instructor,escuela),mean)
medinst
```

```
##   Región K Región D Central
## 1     27.0      8.5    18.5
## 2     12.5     20.0     3.5
```

```
medesc=tapply(puntaje,escuela,mean)
medesc
```

```
## Región K Región D  Central
##    19.75    14.25    11.00
```

```
bj.1=medinst[,1]-medesc[1]
bj.2=medinst[,2]-medesc[2]
bj.3=medinst[,3]-medesc[3]

efinst=rbind(bj.1,bj.2,bj.3)
efinst
```

```
##          1     2
## bj.1  7.25 -7.25
## bj.2 -5.75  5.75
## bj.3  7.50 -7.50
```

- Obtenga manualmente la suma de cuadrados de instructor dentro de escuela.

```
table(instructor,escuela)
```

```
##            escuela
## instructor Región K Región D Central
##          1        2        2       2
##          2        2        2       2
```

```
r=2
scinst=sum(r*efinst^2)
scinst
```

```
## [1] 567.5
```

- ¿Cuántos grados de libertad tiene el factor anidado instructor?

```
gl=(2-1)*3
gl
```

```
## [1] 3
```

Dentro de cada escuela hay dos instructores diferentes por lo que hay un grado de libertad para los instructores de cada escuela. Como son 3 escuelas se tiene en total 3 grados de libertad.

- Calcule el cuadrado medio de instructor dentro de escuela y explique su significado.

```
cminst=scinst/gl
cminst
```

```
## [1] 189.1667
```

El cuadrado medio de instructor dentro de escuela es 189.2. Esta es una medida de la variabilidad entre los promedios obtenidos por los instructores de cada escuela.

---

3. Comparación de instructores.

- Establezca con palabras la hipótesis nula al comparar los instructores.

Los puntajes promedio para cada par de instructores dentro de cada escuela son iguales.

- Haga el análisis de variancia. Introduzca `escuela` e `instructor` pero no agregue interacción. Note que instructor no resulta significativo y que tiene únicamente un grado de libertad cuando en realidad hay 6 instructores distintos.

```
mod1=lm(puntaje~escuela+instructor)
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: puntaje
##            Df Sum Sq Mean Sq F value Pr(>F)
## escuela     2  156.5  78.250  1.2483 0.3374
## instructor  1  108.0 108.000  1.7228 0.2257
## Residuals   8  501.5  62.688
```

El factor instructor tiene solo un grado de libertad cuando debería tener 3. La probabilidad asociada es 0.22, por lo que no se rechaza la hipótesis de igualdad de medias de instructores dentro de cada escuela.

- Estime de nuevo el mismo modelo pero usando `instructor1` y llámelo `mod2`. ¿Qué es lo que está pasando? ¿Cuál es el análisis correcto, con `instructor` o con `instructor1`?

```
mod2=lm(puntaje~escuela+instructor1)
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: puntaje
##             Df Sum Sq Mean Sq F value   Pr(>F)
## escuela      2  156.5   78.25  11.179 0.009473 **
## instructor1  3  567.5  189.17  27.024 0.000697 ***
## Residuals    6   42.0    7.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cuando se usa instructor no se considera la estructura anidada y por eso se tiene solo un grado de libertad, aunque en realidad sean 6. En cambio cuando se usa instructor1 se tiene un grado de libertad para instructor dentro de cada escuela, por lo que resultan 3 grados de libertad. El análisis con instructor1 es correcto.

- Escriba el modelo con `aov` y haga de nuevo el análisis de variancia. Introduzca como factores `escuela` e `instructor` pero agregue la interacción. Observe que ahora instructor sí da significativo, lo mismo que la interacción, pero persiste el problema de los grados de libertad.

```
mod3=aov(puntaje~escuela*instructor)
anova(mod3)
```

```
## Analysis of Variance Table
##
## Response: puntaje
##                    Df Sum Sq Mean Sq F value    Pr(>F)
## escuela             2  156.5   78.25  11.179 0.0094725 **
## instructor          1  108.0  108.00  15.429 0.0077312 **
## escuela:instructor  2  459.5  229.75  32.821 0.0005874 ***
## Residuals           6   42.0    7.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Aunque se introduzca la interacción no se corrige el problema de los grados de libertad, por lo que el análisis no es correcto.

- Use `model.tables(mod)` para obtener las estimaciones de los efectos. Observe que hay un efecto para el instructor 1 y un efecto para el instructor 2, cuando en realidad son 6 instructores diferentes.

```
model.tables(mod3)
```

```
## Tables of effects
##
##  escuela
## escuela
## Región K Región D  Central
##     4.75    -0.75    -4.00
##
##  instructor
## instructor
##  1  2
##  3 -3
##
##  escuela:instructor
##           instructor
## escuela    1     2
##   Región K  4.25 -4.25
##   Región D -8.75  8.75
##   Central   4.50 -4.50
```

Es incorrecto tener solo dos efectos, deberían ser 6 efectos para instructor pues hay 6 instructores.

- Ajuste nuevamente el modelo pero usando `instructor1`, llámelo `mod4`. Observe los efectos de instructor y compárelos con los que se obtienen de `mod2`.

```
mod4=aov(puntaje~escuela*instructor1)
anova(mod4)
```

```
## Analysis of Variance Table
##
## Response: puntaje
##             Df Sum Sq Mean Sq F value   Pr(>F)
## escuela      2  156.5   78.25  11.179 0.009473 **
## instructor1  3  567.5  189.17  27.024 0.000697 ***
## Residuals    6   42.0    7.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model.tables(mod4)
```

```
## Tables of effects
##
##  escuela
## escuela
## Región K Región D  Central
##     4.75    -0.75    -4.00
##
##  instructor1
## instructor1
##     1     2     3     4     5     6
##  7.25 -7.25 -5.75  5.75  7.50 -7.50
```

```
model.tables(aov(mod2))
```

```
## Tables of effects
##
##  escuela
## escuela
## Región K Región D  Central
##     4.75    -0.75    -4.00
##
##  instructor1
## instructor1
##     1     2     3     4     5     6
##  7.25 -7.25 -5.75  5.75  7.50 -7.50
```

El hecho de usar interacción cuando se usa instructor1 no produce ningún cambio con respecto a cuando no se usó interacción. Los efectos están bien estimados y se observa que los efectos de cada par de instructores suman cero dentro de cada escuela. Por ejemplo, el efecto del instructor 1 es 7.25 y el del instructor 2 es -7.25. Así con el 3 y 4 (-5.75 y 5.75) y con el 5 y 6 (7.5 y -7.5).

4. Uso del factor anidado:

   - Se puede usar el factor anidado con el formato en que aparece `instructor`, pero indicando la estructura anidada dentro de la función `lm` de la siguiente forma: `lm(puntaje~escuela+escuela/instructor)`. De esta forma se indica que el factor `instructor` está anidado dentro del factor `escuela`. Use el modelo de suma nula y llámelo `mod5`.

```
options(contrasts=c("contr.sum","contr.poly"))
mod5=lm(puntaje~escuela+escuela/instructor)
```

   - Obtenga el anova del modelo anidado y note que tiene una línea que parece una interacción que dice `escuela:instructor`, pero que en realidad debería leerse `instructor(escuela)`, es decir, instructor dentro de escuela.

```
anova(mod5)
```

```
## Analysis of Variance Table
##
## Response: puntaje
##                    Df Sum Sq Mean Sq F value   Pr(>F)
## escuela             2  156.5   78.25  11.179 0.009473 **
## escuela:instructor  3  567.5  189.17  27.024 0.000697 ***
## Residuals           6   42.0    7.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El resultado es idéntico al de mod4.

   - Observe el cuadrado medio de instructor y compárelo con el obtenido anteriormente. ¿Se rechaza la hipótesis nula referente al instructor?

El cuadrado medio de instructor es 189.2, el cual coincide con el obtenido anteriormente. El cuadrado medio residual es 7, con esto se obtiene un valor de F igual a 27.0, el cual tiene una probabilidad asociada muy pequeña (0.0007). De esta forma se decide rechazar la hipótesis de igualdad de medias para los instructores dentro de cada escuela. En al menos una escuela los instructores no tienen los mismos puntajes promedio.

- Obtenga los efectos de instructor y compárelos con los anteriores.

```
model.tables(aov(mod5))
```

```
## Tables of effects
##
##  escuela
## escuela
## Región K Región D  Central
##     4.75    -0.75    -4.00
##
##  escuela:instructor
##           instructor
## escuela    1     2
##   Región K  7.25 -7.25
##   Región D -5.75  5.75
##   Central   7.50 -7.50
```

Los efectos son los mismos que los obtenidos anteriormente, pero la tabla es más ordenada.

---

5. Comparaciones entre instructores:

- Obtenga los coeficientes del modelo. Hágalo con `mod2`, `mod4` y `mod5`. Observe que `mod2` y `mod4` tienen la desventaja de que hay muchos NA, mientras que `mod5` da solo los coeficientes que se han estimado.

```
mod2$coef
```

```
##  (Intercept)     escuela1     escuela2 instructor11 instructor12
##         15.0         -8.5          5.0         20.5          6.0
## instructor13 instructor14 instructor15
##        -11.5           NA           NA
```

```
mod4$coef
```

```
##           (Intercept)              escuela1              escuela2
##                  15.0                  -8.5                   5.0
##          instructor11          instructor12          instructor13
##                  20.5                   6.0                 -11.5
##          instructor14          instructor15 escuela1:instructor11
##                    NA                    NA                    NA
## escuela2:instructor11 escuela1:instructor12 escuela2:instructor12
##                    NA                    NA                    NA
## escuela1:instructor13 escuela2:instructor13 escuela1:instructor14
##                    NA                    NA                    NA
## escuela2:instructor14 escuela1:instructor15 escuela2:instructor15
##                    NA                    NA                    NA
```

```
mod5$coef
```

```
##                  (Intercept)                      escuela1
##                        15.00                          4.75
##                      escuela2 escuelaRegión K:instructor1
##                        -0.75                          7.25
## escuelaRegión D:instructor1  escuelaCentral:instructor1
##                        -5.75                          7.50
```

- ¿A qué corresponden los últimos 3 coeficientes en `mod5`?

```
mod5$coef[4:6]
```

```
## escuelaRegión K:instructor1 escuelaRegión D:instructor1
##                        7.25                       -5.75
##  escuelaCentral:instructor1
##                        7.50
```

Esos 3 coeficientes son los efectos del primer instructor dentro de cada escuela.

- Haga las comparaciones que tengan sentido entre los instructores. No tiene sentido comparar todos los pares de instructores, puesto que no todos pertenecen a las mismas escuelas. Por ejemplo, no tiene sentido comparar un instructor de la Región D con un instructor de la Región K. Use la corrección de Bonferroni con un nivel de significancia de 0.10 y confianza de 90 %.

Para hacer comparaciones entre instructores hay que escribir el modelo adecuadamente de la siguiente forma:

$$\mu_{ij} = \mu + \alpha_i + \beta_{j(i)}$$

Por ejemplo los dos instructores de la escuela 1 serían:

$\mu_{11} = \mu + \alpha_1 + \beta_{1(1)}$

$\mu_{12} = \mu + \alpha_1 + \beta_{2(1)}$

Entonces la diferencia sería:

$\mu_{11} - \mu_{12} = \beta_{1(1)} - \beta_{2(1)} = \beta_{1(1)} + \beta_{1(1)} = 2\beta_{1(1)}$

Similarmente:

$\mu_{22} - \mu_{21} = \beta_{2(2)} - \beta_{1(2)} = -\beta_{1(2)} - \beta_{1(2)} = -2\beta_{1(2)}$

$\mu_{31} - \mu_{32} = \beta_{1(3)} - \beta_{2(3)} = \beta_{1(3)} + \beta_{1(3)} = 2\beta_{1(3)}$

Todas las diferencias se expresan simplemente como el doble de los coeficientes de los instructores dentro de cada escuela. Por lo tanto, los contrastes están en función de los últimos tres coeficientes solamente. No debe hacerse ningún ajuste porque las comparaciones son independientes entre escuelas.

```r
c1=c(2,0,0)
c2=c(0,-2,0)
c3=c(0,0,2)
contr=cbind(c1,c2,c3)
L=t(contr)%*%mod5$coef[4:6]
ee=sqrt(diag(t(contr)%*%vcov(mod5)[4:6,4:6]%*%contr))
t=L/ee
p=pt(t,6,lower.tail = F)
row.names(p)=c("K1-K2","D2-D1","C1-C2")
round(p,3)
```

```
##         [,1]
## K1-K2 0.001
## D2-D1 0.002
## C1-C2 0.001
```

Como no se hace ningún ajuste, las probabilidades resultantes se comparan con α directamente. Se observan diferencias entre los pares de instructores dentro de todas las escuelas.

Ahora se construyen las tres cotas inferiores con 90 % confianza de forma independiente.

```r
qt=qt(0.9,6); LIM=L-qt*ee
row.names(LIM)=row.names(p)
round(LIM,1)
```

```
##        [,1]
## K1-K2 10.7
## D2-D1  7.7
## C1-C2 11.2
```

En la Región K, el puntaje promedio del instructor 1 es al menos 10.7 puntos mayor que el del instructor 2, en la Región D, el puntaje promedio del instructor 2 es al menos 7.7 puntos mayor que el del instructor 1, y en la Región Central, el puntaje promedio del instructor 1 es al menos 11.2 puntos mayor que el del instructor 2. Todo esto se puede asegurar con 90 % de confianza.

---

6. Haga las comparaciones entre las diferentes escuelas usando Tukey con 10 % de significancia en pruebas de una cola.

```
contrasts(escuela)
```

```
##          [,1] [,2]
## Región K    1    0
## Región D    0    1
## Central    -1   -1
```

```
tapply(puntaje,escuela,mean)
```

```
## Región K Región D  Central
##    19.75    14.25    11.00
```

```
K=c(1,0)
D=c(0,1)
C=c(-1,-1)
KD=K-D
KC=K-C
DC=D-C
contr=cbind(KD,KC,DC)
L=t(contr)%*%mod5$coef[2:3]
L
```

```
##    [,1]
## KD 5.50
## KC 8.75
## DC 3.25
```

```
ee=sqrt(diag(t(contr)%*%vcov(mod5)[2:3,2:3]%*%contr))
q=L/ee
ptukey(q*sqrt(2),3,6,lower.tail = F)
```

```
##           [,1]
## KD 0.058612875
## KC 0.008122427
## DC 0.267659579
```

Puesto que las pruebas son de una cola con 10 % de significancia, las probabilidades deben compararse contra $\alpha = 0{,}10$. Se observan diferencias entre la Región K y las otras dos escuelas.

- Obtenga cotas inferiores de 90 % de confianza para la diferencia de medias entre los pares de escuelas en que encontró diferencias.

```
t=qt(1-0.10/2,6)

LIM=L[-3]-t*ee[-3]
names(LIM)=c("K-D","K-Central")
round(LIM,2)
```

```
##       K-D K-Central
##      1.86      5.11
```

Con 90 % de confianza global se puede esperar que el puntaje promedio de la Región K esté al menos 1.86 puntos sobre el de la Región D, y al menos 5.11 puntos sobre el de la Región Central.

---

7. ¿Cómo se alteran las conclusiones si se considera que hay muchos instructores en cada región y en el experimento los que participaron fueron una muestra?

Para analizar el efecto de instructor no se altera el cálculo de la F, pero sí cambia la hipótesis. En tal caso $H_0 : \sigma^2_\beta = 0$. Además no tendría sentido hacer comparaciones entre instructores a pesar de que se haya demostrado un efecto del instructor. En las comparaciones entre escuelas, para comparar las escuelas se construye la F dividiendo el cuadrado medio de escuela entre el cuadrado medio de instructor dentro de escuela.

```
anova(mod5)
```

```
## Analysis of Variance Table
##
## Response: puntaje
##                    Df Sum Sq Mean Sq F value   Pr(>F)
## escuela             2  156.5   78.25  11.179 0.009473 **
## escuela:instructor  3  567.5  189.17  27.024 0.000697 ***
## Residuals           6   42.0    7.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
CME=anova(mod5)[1,3]; CMI=anova(mod5)[2,3]
F=CME/CMI; 1-pf(F,2,3)
```

```
## [1] 0.6939704
```

La probabilidad asociada es 0.69, lo cual hace que no se rechace la hipótesis nula referente a la igualdad de medias entre escuelas. Sin embargo, se siguen viendo diferencias entre los instructores dentro de escuelas pero no interesa comparar los instructores de forma específica, y no tiene sentido hacer intervalos de confianza.

Conclusión: para hacer el análisis de este ejercicio se compararon diferentes formas de estimar el modelo, y se vio que la forma más simple es especificar el anidamiento entre dos factores dentro de la función `lm`. Es importante verificar siempre que los grados de libertad estén bien calculados para el factor anidado.

Se encontró que la escuela de la Región K obtiene calificaciones promedio más altas que las otras dos escuelas, pero además dentro de cada escuela, hay diferencias entre los dos instructores. Del gráfico que se presentó al inicio, se observa que uno de los instructores de la Región K tiene calificaciones mayores y, aunque su compañero tiene calificaciones más bajas que las de él, juntos hacen que esta escuela sobresalga. Cabe preguntarse si el desempeño de estos dos instructores se debe a mejores condiciones en esta escuela, a que los estudiantes que llegan tienen mayor motivación, o a que realmente estos instructores dan más atención a sus estudiantes. El desempeño de la escuela no puede desligarse del desempeño de los instructores, y la causa de los puntajes altos en la Región K no es clara puesto que los estudiantes no fueron asignados aleatoriamente a las escuelas.

4.2. Cemento

El cemento es el material más activo de la mezcla de concreto, por tanto sus características y sobre todo su contenido (proporción) dentro de la mezcla tienen una gran influencia en la resistencia del concreto a cualquier edad. La resistencia a la compresión simple es la característica mecánica principal del concreto. Se define como la capacidad para soportar una carga por unidad de área, y se expresa en términos de esfuerzo, generalmente en kg/cm^2.

Un experimento trata de analizar la calidad en la producción de cemento, el cual se produce por lotes y se empaca en sacos. Se sospecha que hay variabilidad entre los lotes en que se ha producido el cemento, así como puede haberla entre los sacos provenientes de cada lote. Rutinariamente se seleccionan aleatoriamente 3 sacos de cada lote para tenerlas como referencia. Se escogen aleatoriamente 10 lotes y se hacen 2 pruebas analíticas de resistencia de cada una de los 30 sacos resultantes.

Ejercicios

1. Justifique si los factores están cruzados o anidados.

---

2. Cuando los factores están anidados se debe tener cuidado con la definición del factor anidado, ya que no pueden repetirse los niveles de este factor en los diferentes niveles del factor externo. Abra el archivo `cemento.Rdata`, observe que los sacos tienen los mismos nombres en todos los lotes. Para comprender los datos haga una tabla cruzada de las variables lote y saco.

   - Puede cruzar las dos variables y hacer una nueva variable llamada `saco1` usando la combinación de ambas variables: `base$lote:base$saco`.
   - Obtenga la tabla cruzada de `lote` y `saco1`. Note la diferencia.

---

3. Haga un gráfico para ver el comportamiento de la respuesta en los sacos de cada lote.

---

4. Analice las fuentes de variabilidad que pueden estar incidiendo en la resistencia del cemento. Haga el análisis de diversas formas. ¿Son consistentes los resultados obtenidos con los diferentes enfoques?

Solución

1. Justifique si los factores están cruzados o anidados.

El factor `saco` está anidado dentro del factor `lote`, ya que los 3 sacos que se eligen dentro de cada lote obviamente son diferentes en los diferentes lotes.

---

2. Cuando los factores están anidados se debe tener cuidado con la definición del factor anidado, ya que no pueden repetirse los niveles de este factor en los diferentes niveles del factor externo. Abra el archivo `cemento.Rdata`, observe que los sacos tienen los mismos nombres en todos los lotes. Para comprender los datos haga una tabla cruzada de las variables lote y saco.

```
load("cemento.Rdata")
with(base,table(saco,lote))
```

```
    lote
saco A B C D E F G H I J
   a 2 2 2 2 2 2 2 2 2 2
   b 2 2 2 2 2 2 2 2 2 2
   c 2 2 2 2 2 2 2 2 2 2
```

Aquí se observa que los sacos se llaman siempre a, b y c en todos los lotes llamados A, B, C, etc. Además en cada combinación hay 2 datos, es decir, hay dos réplicas en cada saco dentro de cada lote.

- Puede cruzar las dos variables y hacer una nueva variable llamada `saco1` usando la combinación de ambas variables: `base$lote:base$saco`.

```
base$saco1=base$lote:base$saco
```

- Obtenga la tabla cruzada de `lote` y `saco1`. Note la diferencia.

```
attach(base)
head(table(saco1,lote))
```

```
##      lote
## saco1 A B C D E F G H I J
##   A:a 2 0 0 0 0 0 0 0 0 0
##   A:b 2 0 0 0 0 0 0 0 0 0
##   A:c 2 0 0 0 0 0 0 0 0 0
##   B:a 0 2 0 0 0 0 0 0 0 0
##   B:b 0 2 0 0 0 0 0 0 0 0
##   B:c 0 2 0 0 0 0 0 0 0 0
```

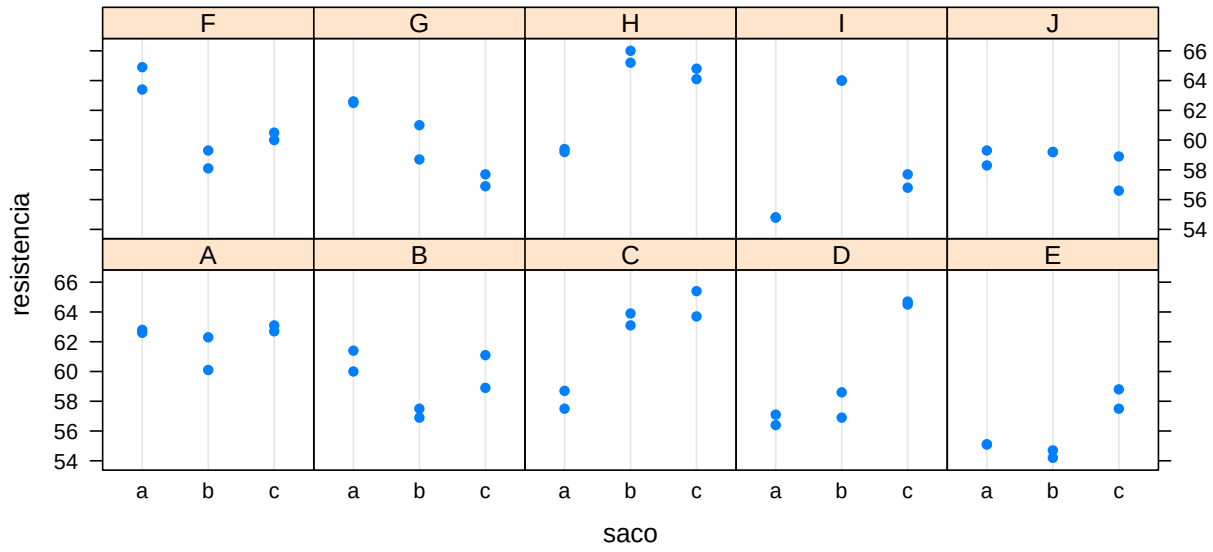
Ahora solo los sacos llamados A:a, A:b y A:c tienen datos en el lote A que es el lote al que pertenecen esos sacos. Aquí se nota la estructura anidada que exige la función `lmer`.

---

3. Haga un gráfico para ver el comportamiento de la respuesta en los sacos de cada lote.

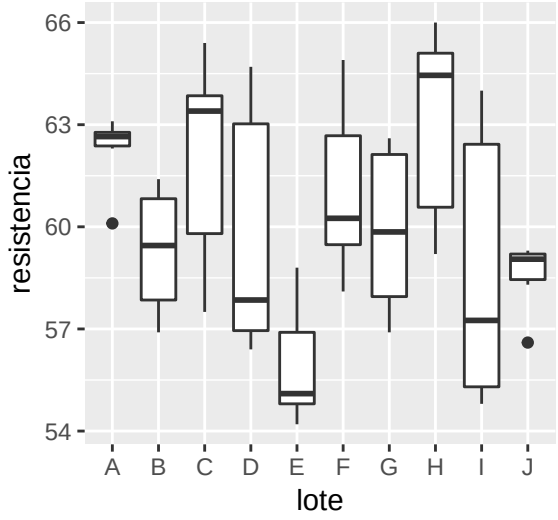
Hay que notar que aquí hay que usar la variable saco que tiene los mismos nombres para todos los lotes para que haga bien el gráfico.

```
dotplot(resist~saco|lote,xlab="saco",ylab="resistencia",layout=c(5,2))
```



Dentro de algunos lotes los promedios de resistencia son muy diferentes de un saco a otro, por ejemplo en los lotes C, D, I.

```
qplot(lote,resist,geom="boxplot",ylab="resistencia")
```



Hay grandes diferencias en resistencia entre algunos de los lotes, por ejemplo, el lote A tiene un promedio bastante alto y mucho mayor que el del lote E cuyo promedio es muy bajo.

4. Analice las fuentes de variabilidad que pueden estar incidiendo en la resistencia del cemento. Haga el análisis de diversas formas. ¿Son consistentes los resultados obtenidos con los diferentes enfoques?

Un primer enfoque es obtener los componentes de variancia y ver el porcentaje que cada componente aporta a la variabilidad total. Hay que notar que no se incluye la interacción en ningún momento debido a que los factores están anidados.

```
library(lme4)
mod0=lmer(resist~1+(1|saco1)+(1|lote))
summary(mod0)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: resist ~ 1 + (1 | saco1) + (1 | lote)
##
## REML criterion at convergence: 247
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.4798 -0.5156  0.0095  0.4720  1.3897
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  saco1    (Intercept) 8.434    2.9041
##  lote     (Intercept) 1.657    1.2874
##  Residual             0.678    0.8234
## Number of obs: 60, groups:  saco1, 30; lote, 10
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  60.0533     0.6769   88.72
```

La variancia total es 10.768 (8.434+1.657+0.678). La proporción que corresponde al saco es 0.78 (8.434/10.768), mientras que la parte que corresponde a lote es 0.15 (1.657/10.768). Entonces, se ve claramente que la mayor fuente de variabilidad se debe a las diferencias entre los sacos de un mismo lote, mientras que entre un lote y otro no hay tanta variabilidad, ya que la de saco es el 78 % mientras que la de lote es solo el 15 %.

Ahora se construyen intervalos de 95 % de confianza para las desviaciones estándar.

```
  confint(profile(mod0))
```

```
##                    2.5 %    97.5 %
## .sig01         2.1579337  4.053589
## .sig02         0.0000000  2.946591
## .sigma         0.6520234  1.085448
## (Intercept) 58.6636504 61.443016
```

En estos resultados, `sig01` es la desviación estándar de `saco`, mientras que `sig02` es la desviación estándar de `lote`. El intervalo correspondiente a lote tiene como límite inferior cero, lo cual es una indicación de que esa fuente de variabilidad no es importante.

Podríamos probar el efecto del saco (dentro de lote) comparando dos modelos.

```
  mod2=lmer(resist~1+(1|saco1)+(1|lote),REML=F)
  mod3=lmer(resist~1+(1|lote),REML=F)
  anova(mod2,mod3,test="LRT")
```

```
## Data: NULL
## Models:
## mod3: resist ~ 1 + (1 | lote)
## mod2: resist ~ 1 + (1 | saco1) + (1 | lote)
##      Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
## mod3  3 308.60 314.88 -151.3   302.60
## mod2  4 255.99 264.37 -124.0   247.99 54.605      1  1.474e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Es claro el efecto del saco dentro de lote, puesto que se obtiene un valor de Chi-cuadrado de 54.6, con un grado de libertad, y la probabilidad asociada es menor a 0.0001. Con este método no se prueba el efecto del lote porque no tendría sentido hacer un modelo sin el lote, ya que los sacos están anidados en los lotes.

Aún existe otra forma de analizar estos efectos con las esperanzas de los cuadrados medios. En este caso habría que usar `saco` anidado dentro de `lote` y ajustar la F para probar el efecto del lote.

```
mod3=lm(resist~lote+(lote/saco))
anova(mod3)
```

```
## Analysis of Variance Table
##
## Response: resist
##           Df Sum Sq Mean Sq F value    Pr(>F)
## lote       9 247.40  27.489  40.544 2.280e-14 ***
## lote:saco 20 350.91  17.545  25.878 9.791e-14 ***
## Residuals 30  20.34   0.678
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
1-pf(27.489/17.545,9,20)
```

```
## [1] 0.1925487
```

La probabilidad asociada a la hipótesis sobre el efecto del lote es 0.19 y la de la hipótesis asociada al saco es <0.0001, con lo cual los resultados son consistentes con los obtenidos anteriormente.

Conclusión: en este ejercicio se utilizaron 4 formas de análisis: 1) descomposición de la variabilidad total y análisis de la contribución de cada fuente de variación (con la función `lmer` y la variable `saco1`), 2) criterio con intervalos de confianza para las desviaciones estándar (con la función `confint` a partir del mismo modelo del punto anterior), 3) comparación de dos modelos (modelo completo contra modelo eliminando `saco1` también con `lmer`), y 4) prueba F a partir de las esperanzas de los cuadrados medios, ajustando la F obtenida de la función `lm`.

El primer enfoque da resultados que ayudan a ver que la mayor fuente de variabilidad es de saco a saco, mientras que la variabilidad de lote a lote no es muy alta. Con este enfoque no se hace ninguna prueba, pero se complementa con el segundo enfoque, donde el intervalo de confianza de la desviación estándar de lote apoya la conclusión de que esta no es una fuente de variabilidad importante. El tercer enfoque tiene la limitante de que no permite probar si hay un efecto del lote, pues no se puede hacer un modelo eliminando el factor `lote`. El último enfoque confirma lo obtenido en los casos anteriores, aunque actualmente se recomienda usar modelos mixtos en lugar de esperanzas de cuadrados medios.

---

5. PARCELAS DIVIDIDAS

En este laboratorio se presentan dos ejercicios con una estructura conocida como parcelas divididas. En este tipo de diseños que tienen dos factores no hay una asignación totalmente aleatoria de las unidades a los tratamientos. El diseño se hace en dos etapas: primero se asignan unidades primarias a los niveles de un factor que se llama **parcela**, luego cada una de las parcelas se subdivide en unidades secundarias llamadas **subparcelas** y estas se asignan a los niveles del segundo factor. En el primer ejercicio llamado **Papel 1** se tienen parcelas que no presentan una estructura de bloques, mientras que en el segundo ejercicio llamado **Papel 2** se presenta una variante del primero, en el que las parcelas están agrupadas en bloques.

Se utiliza función `lme` de la librería `nlme` (Pinheiro et al., 2017) para la estimación de modelos mixtos. Además se usa la librería `lattice` (Sarkar, 2008) para visualización de los datos.

5.1. Papel 1

Un fabricante de papel está interesado en tres métodos para preparar la pulpa y cuatro temperaturas de cocción de la pulpa. Aunque la temperatura es una variable continua, solo se van a estudiar 4 temperaturas, por lo que se toma como un factor fijo.

El fabricante desea estudiar el efecto del método y de la temperatura sobre la resistencia a la tensión del papel, que es el esfuerzo máximo a tensión obtenido durante una prueba hasta la ruptura bajo unas condiciones prescritas. El esfuerzo es expresado como la fuerza por unidad del ancho de la muestra puesta a prueba, medido en kg/cm.

Cada réplica de un factorial requiere 12 observaciones y el investigador ha decidido correr tres réplicas. El experimento se lleva a cabo de la siguiente forma:

- Produce 3 lotes de pulpa con cada uno de los 3 métodos que está estudiando en un orden aleatorio. En total produce 9 lotes de pulpa.
- Cada vez que produce un lote lo divide en cuatro partes o muestras y realiza la cocción de cada muestra con una temperatura diferente asignada aleatoriamente.

Ejercicios

1. Utilice los datos que se encuentran en el archivo `papel1.Rdata`. Asegúrese que están bien definidos los factores método y temperatura.

- Las variables `lote` y `lote1` indican los lotes de pulpa, en `lote` se enumeran de 1 a 3 dentro de cada método y en `lote1` se enumeran de 1 a 9 para diferenciar todos los lotes. Más adelante se verá en qué casos de se debe usar una u otra. Ponga juntas las variables `metodo`, `lote` y `lote1` para observar la correspondencia de los lotes con los métodos.
- Observe el orden en que fue realizada la cocción de la pulpa. De esta forma puede justificar el tipo de aleatorización que se realizó.

- Haga una representación gráfica de los datos para ver el comportamiento de la respuesta según método y temperatura. Analice primero la interacción entre método y temperatura. Use `type=.ª"` en la función `xyplot`. En los diseños de parcela divididas se pone más énfasis al factor que está en la subparcela, por lo que ese factor debe colocarse en `groups`, mientras que el factor de parcela se coloca en el eje X.
- ¿Qué implicaciones tendría una interacción entre método y temperatura?

---

2. Calcule manualmente la suma de cuadrados de error de parcela (`SCEa`). Para esto, calcule la media de cada lote dentro de cada método, luego calcule la `SCEa` usando las desviaciones de estos promedios con respecto a la media de cada método. Recuerde ponderar por el número de datos en cada combinacion de lote y método.

- Los grados de libertad de parcela se obtienen recordando que los lotes están anidados dentro de cada método. Entonces se tienen `a(r-1)` grados de libertad, donde `a` es el número de métodos y `r` el número de lotes por cada método. Usando los grados de libertad adecuados obtenga la estimación de la variancia del error de parcela, la cual está dada por el cuadrado medio de error de parcela (`CMEa`).
- Obtenga el cuadrado medio de método (`CMmet`) de la forma clásica, es decir, basándose en las distancias de los promedios de cada método al promedio general.
- Pruebe la hipótesis sobre el efecto del método. Use el valor de F obtenido al dividir `CMmet` sobre `CMEa` y los grados de libertad adecuados.

---

3. Para obtener la estimación de la variancia residual en la subparcela, es necesario tener la suma de cuadrados total (`SCTot`) y sustraer las sumas de cuadrados consideradas en el modelo. Estime un modelo de la forma usual para obtener las sumas de cuadrados. Incluya en el modelo la interacción entre `metodo` y `temp`.

- Obtenga la `SCTot`.
- Obtenga la suma de cuadrados de error de subparcela (`SCEb`) por sustracción: $SCEb = SCTot - SCmet - SCtemp - SCInt - SCEa$, donde `SCInt` es la suma de cuadrados de la interacción.
- Obtenga los grados de libertad del error de subparcela también por sustracción.
- Obtenga el cuadrado medio de error de subparcela (`CMEb`) y haga la prueba correspondiente a la interacción.
- Asumiendo que no hay interacción entre `metodo` y `temp`, ajuste el `CMEb`. Para esto debe sumar la `SCInt` a la `SCEb` anterior y lo mismo con los grados de libertad. Luego se hace la división nuevamente para obtener el `CMEb`.

- Use la función `lme` de la librería `nlme` para hacer el análisis automático. Debe colocar dos fórmula, en la primera se coloca del modo usual una fórmula con los factores fijos que en este caso son `metodo` y `temp`. Aquí se puede incluir la interacción entre ambos. Luego se coloca en la parte aleatoria las subparcela dentro de cada repetición. La instrucción completa debe quedar de la siguiente forma: `lme(Y~P*SP,ramdom=~1|rep)`. En esta función debe usarse una variable de repetición que identifique a cada lote con números únicos, por esto se usa `lote1`. Haga primero la prueba correspondiente a la interacción.
- Use un modelo sin interacción y haga la prueba sobre el efecto de la temperatura.
- Compare los resultados en los dos modelos anteriores en relación al efecto del método.

---

4. Puesto que no hay interacción entre método y temperatura, para hacer comparaciones múltiples se pueden hacer comparaciones de Tukey. Al hacer las comparaciones es importante tomar en cuenta el error adecuado: cuando se comparan métodos se debe tomar el `CMEa` y cuando se comparan temperaturas el `CMEb` (el último que se obtuvo cuando ya se había eliminado la interacción). En este caso solo se rechazó la hipótesis de igualdad de medias para las temperaturas, por lo que solo para ese caso interesa hacer comparaciones.

- Haga un gráfico para comparar la resistencia entre las diferentes temperaturas.
- Compare los promedios y obtenga límites inferiores para los casos en que tenga sentido.

Solución

1. Utilice los datos que se encuentran en el archivo `papel1.Rdata`. Asegúrese que están bien definidos los factores método y temperatura.

```
load("papel1.Rdata")
base$metodo = factor(base$metodo)
base$lote=factor(base$lote)
base$lote1=factor(base$lote1)
levels(base$metodo)=c("M1","M2","M3")
base$temp = factor(base$temp)
attach(base)
```

- Las variables `lote` y `lote1` indican los lotes de pulpa, en `lote` se enumeran de 1 a 3 dentro de cada método y en `lote1` se enumeran de 1 a 9 para diferenciar todos los lotes. Más adelante se verá en que casos de se debe usar una u otra. Ponga juntas las variables `metodo`, `lote` y `lote1` para observar la correspondencia de los lotes con los métodos.

```
cbind(metodo,lote,lote1)[1:24,]
```

```
##       metodo lote lote1
##  [1,]      1    1     1
##  [2,]      1    1     1
##  [3,]      1    1     1
##  [4,]      1    1     1
##  [5,]      2    1     2
##  [6,]      2    1     2
##  [7,]      2    1     2
##  [8,]      2    1     2
##  [9,]      2    2     3
## [10,]      2    2     3
## [11,]      2    2     3
## [12,]      2    2     3
## [13,]      1    2     4
## [14,]      1    2     4
## [15,]      1    2     4
## [16,]      1    2     4
## [17,]      3    1     5
## [18,]      3    1     5
## [19,]      3    1     5
## [20,]      3    1     5
## [21,]      1    3     6
## [22,]      1    3     6
## [23,]      1    3     6
## [24,]      1    3     6
```

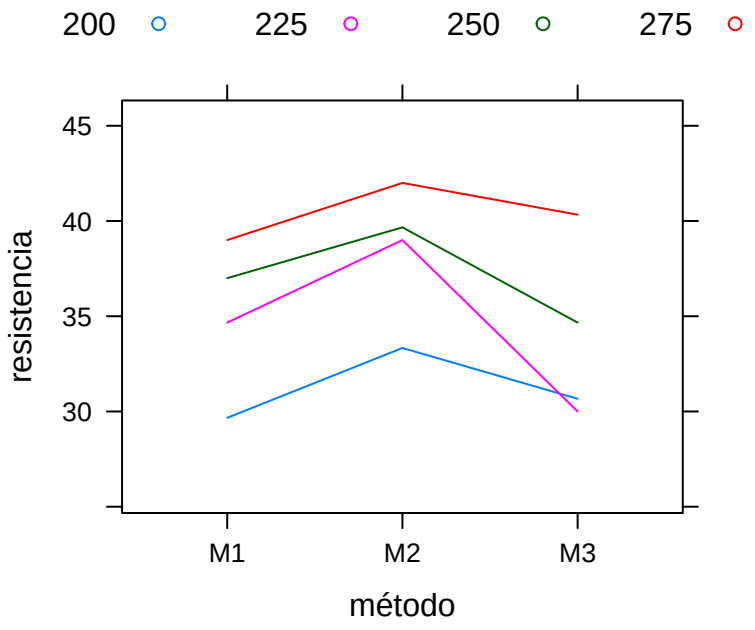
Los lotes 1, 4 y 6 fueron producidos con el método 1. En la variable `lote` se denominan 1, 2 y 3. Los lotes 2, 3 y 8 fueron producidos con el método 2 y también se denominan 1, 2 y 3 en la variable `lote`. Finalmente, los lotes 5, 7, 9 fueron producidos con el método 3 y también se denominan 1, 2 y 3 en la variable `lote`.

- Observe el orden en que fue realizada la cocción de la pulpa. De esta forma puede justificar el tipo de aleatorización que se realizó.

Se observa que cada cuatro observaciones fueron producidas con un mismo método que corresponde a la producción de la pulpa. El grupo de cuatro observaciones es un lote y se llama parcela. A su vez, ese lote fue subdividido en cuatro cocciones con diferentes temperaturas, las cuales no siempre siguieron el mismo orden. En el primer lote, por ejemplo, las cocciones se hicieron con el siguiente orden de temperatura: 200, 275, 225 y 250. En el segundo lote el orden fue: 275, 250, 200 y 225.

- Haga una representación gráfica de los datos para ver el comportamiento de la respuesta según método y temperatura. Analice primero la interacción entre método y temperatura. Use `type=."a"` en la función `xyplot`. En los diseños de parcela divididas se pone más énfasis al factor que está en la subparcela, por lo que ese factor debe colocarse en `groups`, mientras que el factor de parcela se coloca en el eje X.

```
library(lattice)
xyplot(res ~ metodo, groups = temp,type="a",auto.key=list(columns=4),
       xlab="método",ylab="resistencia")
```



En el gráfico no parece haber interacción entre temperatura y método.

- ¿Qué implicaciones tendría una interacción entre método y temperatura?

Si hubiera interacción entre método y temperatura se esperaría que el efecto que tiene la temperatura fuera diferente para cada método. En este caso parece ser consistente que a mayor temperatura, la resistencia promedio aumenta, independientemente del método.

---

2. Calcule manualmente la suma de cuadrados de error de parcela (`SCEa`). Para esto, calcule la media de cada lote dentro de cada método, luego calcule la `SCEa` usando las desviaciones de estos promedios con respecto a la media de cada método. Recuerde ponderar por el número de datos en cada combinacion de lote y método.

```
m.lot=tapply(res,list(lote,metodo),mean)
m.met=tapply(res,metodo,mean)
m.met1=array(rep(m.met,each=3),list(3,3))
table(lote,metodo)
```

```
##     metodo
## lote M1 M2 M3
##    1  4  4  4
##    2  4  4  4
##    3  4  4  4
```

```
SCEa=sum((m.lot-m.met1)^2)*4
SCEa
```

```
## [1] 102.8333
```

- Los grados de libertad de parcela se obtienen recordando que los lotes están anidados dentro de cada método. Entonces se tienen `a(r-1)` grados de libertad, donde `a` es el número de métodos y `r` el número de lotes por cada método. Usando los grados de libertad adecuados obtenga la estimación de la variancia del error de parcela, la cual está dada por el cuadrado medio de error de parcela (`CMEa`).

```
gl.a=3*(3-1)
gl.a
```

```
## [1] 6
```

```
CMEa=SCEa/gl.a
CMEa
```

```
## [1] 17.13889
```

- Obtenga el cuadrado medio de método (`CMmet`) de la forma clásica, es decir, basándose en las distancias de los promedios de cada método al promedio general.

```
table(metodo)
```

```
## metodo
## M1 M2 M3
## 12 12 12
```

```
gl.met=2
SCmet=12*sum((m.met-mean(res))^2)
CMmet=SCmet/gl.met
CMmet
```

```
## [1] 68.08333
```

- Pruebe la hipótesis sobre el efecto del método. Use el valor de F obtenido al dividir `CMmet` sobre `CMEa` y los grados de libertad adecuados.

```
Fa=CMmet/CMEa
Fa
```

```
## [1] 3.972447
```

```
pa=pf(Fa,gl.met,gl.a,lower.tail = F)
pa
```

```
## [1] 0.07965408
```

La probabilidad asociada a esta prueba es mayor que el nivel de significancia (p=0.08), por lo que no se rechaza la hipótesis de igualdad de medias entre métodos. No hay evidencia para decir que existe un efecto del método sobre la resistencia promedio.

---

3. Para obtener la estimación de la variancia residual en la subparcela, es necesario tener la suma de cuadrados total (`SCTot`) y sustraer las sumas de cuadrados consideradas en el modelo. Estime un modelo de la forma usual para obtener las sumas de cuadrados. Incluya en el modelo la interacción entre `metodo` y `temp`.

```
mod1=lm(res~metodo*temp)
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: res
##             Df Sum Sq Mean Sq F value    Pr(>F)
## metodo       2 136.17  68.083  9.1455  0.001116 **
## temp         3 412.11 137.370 18.4527 2.002e-06 ***
## metodo:temp  6  58.06   9.676  1.2998  0.295043
## Residuals   24 178.67   7.444
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Obtenga la `SCTot`.

```
SCTot=sum(anova(mod1)[,2])
SCTot
```

```
## [1] 785
```

```
n=nrow(base)
n
```

```
## [1] 36
```

```
var(res)*(n-1)
```

```
## [1] 785
```

La SCTot se puede calcular a partir de la variancia de la respuesta y también como la suma de las sumas de cuadrados en el análisis de variancia. En ambos casos se obtiene 785.

- Obtenga la suma de cuadrados de error de subparcela (`SCEb`) por sustracción: $SCEb = SCTot - SCmet - SCtemp - SCInt - SCEa$, donde `SCInt` es la suma de cuadrados de la interacción.

```
SCtemp=anova(mod1)[2,2]
SCInt=anova(mod1)[3,2]
SCEb=SCTot-SCmet-SCtemp-SCInt-SCEa
SCEb
```

```
## [1] 75.83333
```

- Obtenga los grados de libertad del error de subparcela también por sustracción.

```
gl.b=35-2-3-6-gl.a
gl.b
```

```
## [1] 18
```

- Obtenga el cuadrado medio de error de subparcela (`CMEb`) y haga la prueba correspondiente a la interacción.

```
CMEb=SCEb/gl.b
CMInt=anova(mod1)[3,3]
CMInt
```

```
## [1] 9.675926
```

```
Fint=CMInt/CMEb
pint=pf(Fint,6,gl.b,lower.tail = F)
pint
```

```
## [1] 0.08015924
```

La probabilidad asociada a esta prueba es mayor que el nivel de significancia (p=0.08), por lo que no se rechaza la hipótesis de no interacción. De esta forma se asume que el efecto de la temperatura es el mismo para cada método.

- Asumiendo que no hay interacción entre `metodo` y `temp`, ajuste el `CMEb`. Para esto debe sumar la `SCInt` a la `SCEb` anterior y lo mismo con los grados de libertad. Luego se hace la división nuevamente para obtener el `CMEb`.

```
SCEb=SCEb+SCInt
gl.b=gl.b+6
CMEb=SCEb/gl.b
CMtemp=anova(mod1)[2,3]
Fb=CMtemp/CMEb
pb=pf(Fb,3,gl.b,lower.tail = F)
pb
```

```
## [1] 1.675399e-07
```

La probabilidad asociada a esta prueba es muy pequeña (p<0.0001), por lo que se rechaza la hipótesis nula de no efecto de la temperatura. Se concluye, con un nivel de significancia de 5 %, que la temperatura sí tiene un efecto sobre la resitencia promedio, independientemente del método. Hay que investigar para cuáles pares de temperaturas se detectan diferencias entre las resistencias promedio.

- Use la función `lme` de la librería `nlme` para hacer el análisis automático. Debe colocar dos fórmula, en la primera se coloca del modo usual una fórmula con los factores fijos que en este caso son `metodo` y `temp`. Aquí se puede incluir la interacción entre ambos. Luego se coloca en la parte aleatoria las subparcela dentro de cada repetición. La instrucción completa debe quedar de la siguiente forma: `lme(Y~P*SP,ramdom=~1|rep)`. En esta función debe usarse una variable de repetición que identifique a cada lote con números únicos, por esto se usa `lote1`. Haga primero la prueba correspondiente a la interacción.

```
library(nlme)
mod2=lme(res~metodo*temp,random= ~1|lote1)
anova(mod2)
```

```
##             numDF denDF   F-value p-value
## (Intercept)     1    18 2697.0827  <.0001
## metodo          2     6    3.9724  0.0797
## temp            3    18   32.6066  <.0001
## metodo:temp     6    18    2.2967  0.0802
```

- Use un modelo sin interacción y haga la prueba sobre el efecto de la temperatura.

```
mod3=lme(res~metodo+temp,random= ~1|lote1)
anova(mod3)
```

```
##             numDF denDF   F-value p-value
## (Intercept)     1    24 2697.0827  <.0001
## metodo          2     6    3.9724  0.0797
## temp            3    24   24.6241  <.0001
```

- Compare los resultados en los dos modelos anteriores en relación al efecto del método.

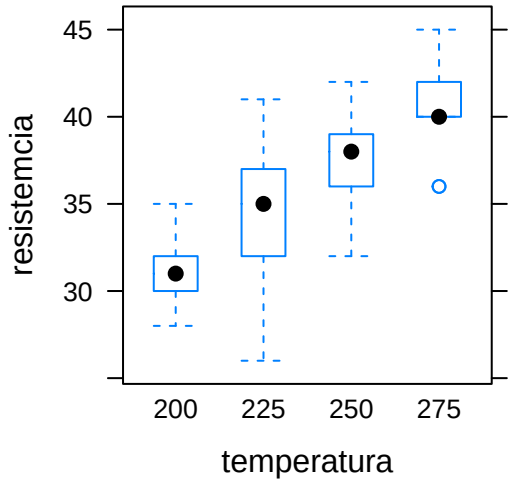
La prueba sobre el efecto del método no se ve alterada si se incluye o no la interacción, sin embargo, la interpretación del efecto del método solo es válida en el caso en que no se incluyó interacción.

---

4. Puesto que no hay interacción entre método y temperatura, para hacer comparaciones múltiples se pueden hacer comparaciones de Tukey. Al hacer las comparaciones es importante tomar en cuenta el error adecuado: cuando se comparan métodos se debe tomar el CMEa y cuando se comparan temperaturas el CMEb (el último que se obtuvo cuando ya se había eliminado la interacción). En este caso solo se rechazó la hipótesis de igualdad de medias para las temperaturas, por lo que solo para ese caso interesa hacer comparaciones.

- Haga un gráfico para comparar la resistencia entre las diferentes temperaturas.

```
bwplot(res~temp,ylab="resistemcia",xlab="temperatura")
```



A mayor temperatura mayor resistencia promedio. Se observan diferencias claras entre los promedios de todas las temperaturas, excepto entre 225 y 250.

- Compare los promedios y obtenga límites inferiores para los casos en que tenga sentido.

```
n.b=table(temp)[1]
med = tapply(res,temp, mean)
m275.200=med[4]-med[1]
m275.225=med[4]-med[2]
m275.250=med[4]-med[3]
m250.200=med[3]-med[1]
m250.225=med[3]-med[2]
m225.200=med[2]-med[1]
dif=c(m225.200,m250.200,m275.200,m250.225,m275.225,m275.250)
names(dif)=c("225-200","250-200","275-200","250-225","275-225","275-250")
dif
```

```
##  225-200  250-200  275-200  250-225  275-225  275-250
## 3.333333 5.888889 9.222222 2.555556 5.888889 3.333333
```

```
ee  = sqrt(2*CMEb/n.b)
q=dif/ee
p=ptukey(q*sqrt(2),4,gl.b,lower.tail = F)
round(p,3)
```

```
## 225-200 250-200 275-200 250-225 275-225 275-250
##   0.030   0.000   0.000   0.127   0.000   0.030
```

```
q   = qt(1-0.05/5,gl.b)
ic  = dif[-4]-q*ee
names(ic)=names(dif)[-4]
round(ic,2)
```

```
## 225-200 250-200 275-200 275-225 275-250
##    0.56    3.11    6.45    3.11    0.56
```

No se observan diferencias entre las medias de las temperaturas 250 y 225, pero sí las hay entre todos los demás pares. Las que más se diferencian son 200 con 275 pues la resistencia promedio para 275 es al menos 6.45kg/cm mayor que la de 200.

Conclusión: no se encontró evidencia de interacción entre método y temperatura, lo cual facilita la intepretación de los resultados ya que el efecto que pueda tener la temperatura es el mismo para cualquier método. Tampoco se pudo probar que existan diferencias en la resitencia promedio del papel al ser elaborado con uno u otro método. Cabe preguntarse si el número de lotes utilizados es suficiente para detectar diferencias, por lo que valdría la pena hacer un estudio de la potencia que tiene este experimento. Finalmente, sí se encontró que algunas temperaturas producen una mayor resitencia promedio. A mayor temperatura se obtiene una resitencia promedio también mayor, por lo que el promedio de resistencia para una temperatura de 275 supera el que se obtiene con temperaturas más bajas como 200 y 225, sin embargo, la diferencia con el obtenido para 250 es apenas de 0.56kg/cm.

5.2. Papel 2

Como variante del ejercicio anterior, suponga que la capacidad de la planta solo permite realizar 15 corridas por día y se sospecha que de un día a otro se pueden experimentar diferencias en la resistencia. Entonces el fabricante decide considerar los días como bloques y corre una réplica en cada uno de tres días.

Cada día lleva a cabo el experimento así:

- Produce tres lotes de pulpa, cada lote con uno de los tres métodos que está estudiando en un orden aleatorio.
- Cada lote se divide en cuatro partes o muestras y realiza la cocción de cada muestra con una temperatura diferente, asignada aleatoriamente.

Ejercicios

1. Utilice los datos que se encuentran en el archivo `papel2.Rdata`. Asegúrese que están bien definidos los factores `metod` y`temp`, así como `dia` y `lote`.

---

2. Empiece con el modelo con interacción entre `metodo` y `temp` e incluya `dia` como bloque. Debe reconocerse que los días son aleatorios. Además, dentro de cada día se seleccionaron 3 lotes, de forma también aleatoria, a los cuales se les asignó un método específico. Esta estructura debe indicarse en la segunda parte de la función `lme`, de la siguiente forma: `random=~1|dia/lote1`, esto quiere decir que el lote está anidado dentro de cada día y que ambos son aleatorios. Si por alguna razón el día no fuera aleatorio, entonces se debe indicar solamente `random=~1|lote1`, pero recordando que siempre se debería colocar `dia` en la parte fija.

- Pruebe la hipótesis de no interacción.
- Use el modelo sin interacción y pruebe si hay un efecto del método y de la temperatura.

---

3. Como se vio que el método tiene un efecto sobre la resistencia promedio, es importante comparar los promedios de resistencia entre los diferentes métodos.

- Haga un gráfico para comparar la resistencia según los diferentes métodos.
- Para hacer las pruebas de diferencias de medias, el mejor camino es usando los contrastes, ya que no se tiene el `CMEa` de forma explícita. Escriba los contrastes adecuados y verifique las hipótesis de igualdad de pares de medias usando Tukey.

---

4. También se vio que la temperatura tiene un efecto sobre la resistencia promedio. Haga un gráfico para comparar la resistencia según las diferentes temperaturas.

- Haga las comparaciones de pares de promedios usando contrastes.
- Para el caso de la subparcela, la función `nlme` sí arroja la estimación de la desviación estándar residual. Se obtiene con `mod5$sigma`. Entonces se puede obtener el `CMEb` elevando esta estimación al cuadrado. De esta forma, es fácil obtener el error estándar en las comparaciones del punto anterior. Obtenga ese error estándar y compárelo con el obtenido en el punto anterior.

Solución

1. Utilice los datos que se encuentran en el archivo `papel2.Rdata`. Asegúrese que están bien definidos los factores `metod` y`temp`, así como `dia` y `lote`.

```
load("papel2.Rdata")
str(base)
```

```
## 'data.frame':    36 obs. of  6 variables:
##  $ res   : int  30 35 37 36 34 41 38 42 29 37 ...
##  $ dia   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ metodo: int  1 1 1 1 2 2 2 2 3 3 ...
##  $ temp  : int  200 225 250 275 200 225 250 275 200 225 ...
##  $ lote  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ lote1 : int  1 1 1 1 2 2 2 2 3 3 ...
```

```
base$dia = factor(base$dia)
base$metodo = factor(base$metodo)
base$lote=factor(base$lote)
base$lote1=factor(base$lote1)
levels(base$metodo)=c("M1","M2","M3")
base$temp = factor(base$temp)
attach(base)
```

---

2. Empiece con el modelo con interacción entre `metodo` y `temp` e incluya `dia` como bloque. Debe reconocerse que los días son aleatorios. Además, dentro de cada día se seleccionaron 3 lotes, de forma también aleatoria, a los cuales se les asignó un método específico. Esta estructura debe indicarse en la segunda parte de la función `lme`, de la siguiente forma: `random=~1|dia/lote1`, esto quiere decir que el lote está anidado dentro de cada día y que ambos son aleatorios. Si por alguna razón el día no fuera aleatorio, entonces se debe indicar solamente `random=~1|lote1`, pero recordando que siempre se debería colocar `dia` en la parte fija.

- Pruebe la hipótesis de no interacción.

```
mod4=lme(res~metodo*temp,random=~1|dia/lote1)
anova(mod4)
```

```
##             numDF denDF   F-value p-value
## (Intercept)     1    18 1692.2493  <.0001
## metodo          2     4    8.0055  0.0400
## temp            3    18   24.7855  <.0001
## metodo:temp     6    18    1.2506  0.3278
```

No se detecta que haya interacción entre método y temperatura ya que la probabilidad asociada es mayor que el nivel de significancia (p=0.33).

- Use el modelo sin interacción y pruebe si hay un efecto del método y de la temperatura.

```
mod5=lme(res~metodo+temp,random=~1|dia/lote1)
anova(mod5)
```

```
##             numDF denDF   F-value p-value
## (Intercept)     1    24 1692.2493  <.0001
## metodo          2     4    7.5974  0.0434
## temp            3    24   23.5222  <.0001
```

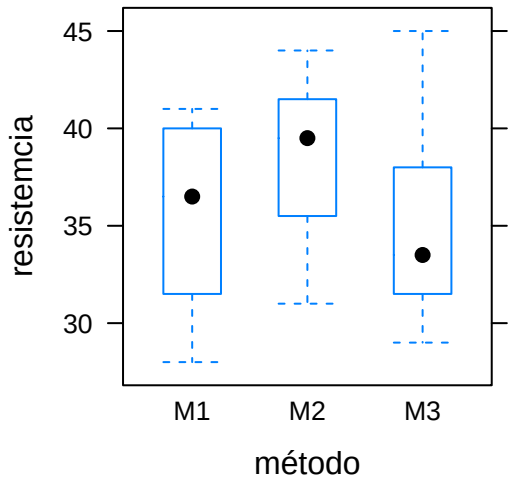
Tanto método como temperatura tienen probabilidades asociadas menores al nivel de significancia (p=0.04 para método y p<0.0001 para temperatura), por lo que se concluye que ambos factores tienen un efecto sobre la resistencia promedio.

---

3. Como se vio que el método tiene un efecto sobre la resistencia promedio, es importante comparar los promedios de resistencia entre los diferentes métodos.

- Haga un gráfico para comparar la resistencia según los diferentes métodos.

```
bwplot(res~metodo,ylab="resistemcia",xlab="método")
```



Se podría esperar que haya diferencia entre los promedios de los métodos 2 y 3, mientras que el método 1 está ubicado en el medio de los otros 2.

- Para hacer las pruebas de diferencias de medias, el mejor camino es usando los contrastes, ya que no se tiene el `CMEa` de forma explícita. Escriba los contrastes adecuados y verifique las hipótesis de igualdad de pares de medias usando Tukey.

```
options(contrasts=c("contr.treatment","contr.poly"))
contrasts(metodo)
```

```
##    M2 M3
## M1  0  0
## M2  1  0
## M3  0  1
```

```
tapply(res,metodo,mean)
```

```
##       M1       M2       M3
## 35.66667 38.50000 34.83333
```

```
mod5=lme(res~metodo+temp,random=~1|dia/lote1)
b=mod5$coef$fixed[2:3]
c2.1=c(1,0)
c2.3=c(1,-1)
c1.3=c(0,-1)
cont=cbind(c2.1,c2.3,c1.3)
L=t(cont)%*%b
ee=sqrt(diag(t(cont)%*%vcov(mod5)[2:3,2:3]%*%cont))
q=L/ee

p=ptukey(q*sqrt(2),3,4,lower.tail = F)
round(p,3)
```

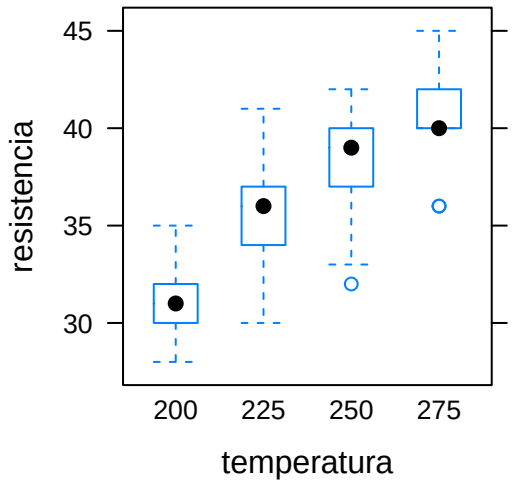
```
##       [,1]
## c2.1 0.094
## c2.3 0.044
## c1.3 0.699
```

Solamente se encuentran diferencias entre los promedios del método 2 y el método 3.

---

4. También se vio que la temperatura tiene un efecto sobre la resistencia promedio. Haga un gráfico para comparar la resistencia según las diferentes temperaturas.

```
bwplot(res~temp,ylab="resistencia",xlab="temperatura")
```



En este caso parece haber diferencias entre todos los pares de temperaturas, siendo que a mayor temperatura se tiene una resistencia promedio mayor.

- Haga las comparaciones de pares de promedios usando contrastes.

Primero establecemos el modelo a utilizar y escogemos el de tratamiento referencia (también se puede usar el de suma nula). Luego se estima nuevamente el modelo y se construyen los contrastes.

```
b=mod5$coef$fixed[4:6]
c225.200=c(1,0,0)
c250.200=c(0,1,0)
c275.200=c(0,0,1)
c250.225=c(-1,1,0)
c275.225=c(-1,0,1)
c275.250=c(0,-1,1)
cont=cbind(c225.200,c250.200,c275.200,c250.225,c275.225,c275.250)
L=t(cont)%*%b
ee=sqrt(diag(t(cont)%*%vcov(mod5)[4:6,4:6]%*%cont))
q=L/ee
```

```
p=ptukey(q*sqrt(2),4,24,lower.tail = F)
round(p,3)
```

```
##           [,1]
## c225.200 0.003
## c250.200 0.000
## c275.200 0.000
## c250.225 0.274
## c275.225 0.002
## c275.250 0.140
```

No se detectaron diferencias en todos los pares; específicamente no se puede decir que el promedio de resistencia para una temperatura de 250 grados sea mayor que el de la temperatura de 225, así como tampoco que el de 275 grados sea mayor que el de 250 grados. Se ha probado que la resistencia promedio a una temperatura de 275 sí es mayor que la de temperaturas bajas como 200 y 225 grados. De igual forma, para las temperaturas de 225 y 250 sí se tiene una resitencia promedio mayor que para 200 grados.

- Para el caso de la subparcela, la función `nlme` sí arroja la estimación de la desviación estándar residual. Se obtiene con `mod5$sigma`. Entonces se puede obtener el `CMEb` elevando esta estimación al cuadrado. De esta forma, es fácil obtener el error estándar en las comparaciones del punto anterior. Obtenga ese error estándar y compárelo con el obtenido en el punto anterior.

```
CMEb=mod5$sigma^2
table(temp)
```

```
## temp
## 200 225 250 275
##   9   9   9   9
```

```
ee1=sqrt(2*CMEb/9)
ee1
```

```
## [1] 1.138744
```

```
ee
```

```
## c225.200 c250.200 c275.200 c250.225 c275.225 c275.250
## 1.138744 1.138744 1.138744 1.138744 1.138744 1.138744
```

Se obtiene el mismo resultado en ambos casos (1.139).

Conclusión: para lograr altas resistencias se puede recomendar el método 2, el cual produce resistencias promedio que son mayores a las del método 3, aunque no se alejan significativamente del método 1. En cuanto a la tempertura, se puede ver que el efecto que esta tiene sobre la resistencia promedio es el mismo en todos los métodos, lo cual simplifica la decisión de escoger una temperatura para la producción del papel. Aunque los resultados para 275 no son necesariamente mejores que los de 250, sí lo son con respecto a las temperaturas más bajas.

---

6. MEDIDAS REPETIDAS

En este laboratorio se presentan 4 ejercicios con mediciones repetidas sobre sujetos que han sido tomados aleatoriamente de una población. Los ejercicios se llaman **Sueño**, **Riqueza**, **Ortodoncia** y **Arbustos**. En tres de estos ejercicios se toman medidas a lo largo del tiempo, por lo que se utiliza como predictor una variable relacionada con el tiempo (`dias`, `edad` y `año`), mientras que en el último ejercicio, el predictor es una variable diferente (`NAP`). En algunos casos los sujetos son personas (conductores o niños), mientras que en otros son entidades físicas como playas o parcelas de cultivo. La primera parte del análisis consiste en determinar si existe correlación lineal entre las pendientes y los interceptos al aplicar una regresión a cada sujeto. A partir de ahí, se procede a verificar si las pendientes de todos los sujetos bajo un mismo tratamiento se pueden considerar iguales. Finalmente se comparan las pendientes generales de los diferentes tratamientos y se hacen las interpretaciones.

Se usa la función `lmer` de la librería `lme4` (Bates et al., 2015) para todos los análisis de los modelos mixtos. Además se usa la librería `lattice` (Sarkar, 2008) para visualización de los datos.

6.1. Sueño

Se investiga el efecto que tiene la privación del sueño sobre el tiempo de reacción en las personas. Se selecciona una muestra de conductores de camiones de larga distancia, los cuales son asignados aleatoriamente a grupos. En cada grupo se les restringe el número de horas de sueño a un máximo de horas por noche durante el período del experimento. Se mide el tiempo de reacción varias veces cada día del experimento para cada conductor. El experimento se sigue por 10 días. La variable respuesta es el promedio de tiempos de reacción en milisegundos para un sujeto en un día determinado.

Se toman solamente lo sujetos en el grupo en que se restringió el número de horas de sueño a 3 horas cada noche. Se cuenta con 18 sujetos en ese grupo. Puede ser interesante comparar grupos en los que se ha restringido el sueño a diferentes números de horas, sin embargo, en este ejercicio, no se hacen comparaciones, sino que se analiza un único grupo.

Ejercicios

1. Abra el archivo `sueño.Rdata`.

---

2. En primer lugar se va a analizar si existe relación entre las pendientes y los interceptos al hacer una regresión simple para cada sujeto usando `dias` como predictor y `reac` como respuesta.

- Ajuste una regresión para cada sujeto y almacene en un vector llamado `beta0` los interceptos de estas regresiones y en otro llamado `beta1` las pendientes.
- Haga las líneas de regresión en un solo gráfico, con la función `xyplot` de la librería `lattice`, indicando `type='r'`. Visualice si existe una relación entre los valores de las pendientes y los interceptos.
- Haga un gráfico que relacione los interceptos de `beta0` con las pendientes de `beta1`. Además obtenga el coeficiente de correlación de estos dos vectores.
- Ajuste un modelo lineal ordinario con tiempo de reacción en función del tiempo (días) sin tomar en cuenta la tendencia en cada sujeto. Escriba la ecuación resultante, la cual deberá comparar más adelante con los resultados obtenidos en el modelo mixto.
- Se van a ajustar dos modelos: 1) un modelo mixto asumiendo que hay correlación entre las pendientes y los interceptos, 2) un modelo que asume que no hay correlación entre las pendientes y los interceptos. Ambos modelos se deben ajustar con máxima verosimilitud, pues la idea es compararlos con la prueba de razón de verosimilitud (LRT). Use la función `lmer` de la librería `lme4`. Para el primer modelo se pone la línea de tendencia general que se considera un efecto fijo, luego se toma el sujeto como efecto aleatorio y dentro de cada sujeto se estima una regresión. El 1 representa el intercepto y se puede omitir, pero aunque se omita siempre van a estimarse los interceptos: `lmer(reac~1+dias+(1+dias|sujeto),REML=F)` o `lmer(reac~dias+(dias|sujeto),REML=F)`. Se agrega `REML=F` para indicar que se haga el ajuste por máxima verosimilitud en vez de usar REML.
- Obtenga el `summary` de este modelo y observe el valor de la correlación entre pendientes e interceptos.
- Ahora ajuste el modelo pero asumiendo que no hay correlación entre pendientes e interceptos. Para esto debe ajustar los interceptos dentro de cada sujeto de forma independiente de las pendientes: `lmer(reac~1+dias+(1|sujeto)+(0+dias|sujeto),REML=F)`. El cero en la parte de las pendientes obliga a que no estime el intercepto junto con la pendiente dentro de cada sujeto.
- Compare los dos modelos anteriores con la prueba LRT. Se usa un `anova` de los dos modelos anidados y se indica `test="LRT"`.

---

3. Ahora se quiere verificar si se puede asumir que todos los sujetos tienen la misma pendiente. Para esto se ajusta un modelo donde se elimina la pendiente de la parte aleatoria y luego se compara con el `mod3`, que fue el ganador en el punto anterior.

- Interprete lo que significa el resultado anterior en términos del problema.

---

4. Estime el modelo escogido con REML para hacer las interpretaciones finales. Obtenga el `summary` del modelo y escriba la ecuación general.

- Compare los errores estándar de los coeficientes obtenidos con la regresión ordinaria y con el modelo mixto.
- Obtenga intervalos de 95 % de confianza para los parámetros del modelo e interprételos. Use `confint(profile(mod))`. Aquí va a obtener primero los intevalos para las desviaciones estándar de los componentes aleatorios llamados `.sig01` para los interceptos aleatorios y `.sig02` para las pendientes aleatorias, seguida de la desviación estándar residual. Luego seguirán los términos de la parte fija que son el intercepto general y la pendiente general.

Solución

1. Abra el archivo `sueño.Rdata`.

```
rm(list=ls(all=TRUE))
load("sueño.Rdata")
attach(base)
```

---

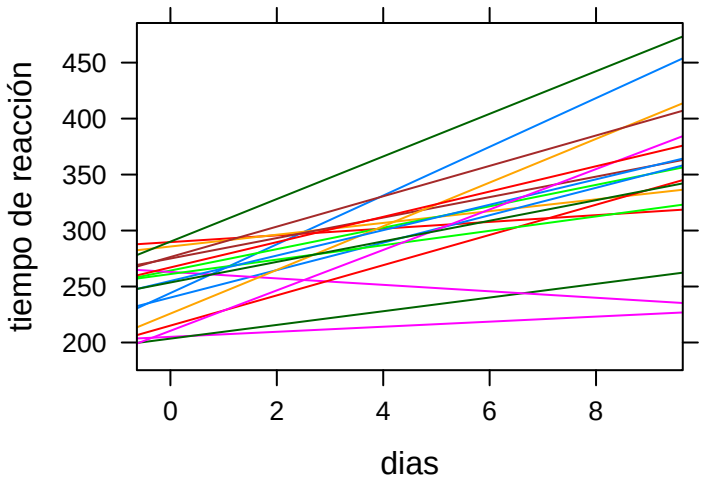
2. En primer lugar se va a analizar si existe relación entre las pendientes y los interceptos al hacer una regresión simple para cada sujeto usando `dias` como predictor y `reac` como respuesta.

- Ajuste una regresión para cada sujeto y almacene en un vector llamado `beta0` los interceptos de estas regresiones y en otro llamado `beta1` las pendientes.

```
beta0=beta1=c()
ind=as.numeric(names(table(sujeto)))
for(i in 1:18)  {
  mod=lm(reac~dias,base[sujeto==ind[i],])
  beta0[i]=mod$coef[1]
  beta1[i]=mod$coef[2]
}
```

- Haga las líneas de regresión en un solo gráfico, con la función `xyplot` de la librería `lattice`, indicando `type='r'`. Visualice si existe una relación entre los valores de las pendientes y los interceptos.

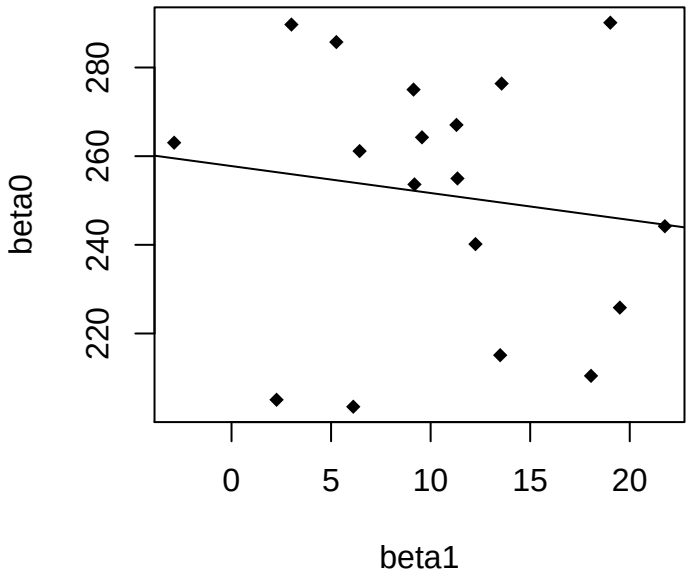
```
library(lattice)
xyplot(reac~dias,group=sujeto,
       ylab="tiempo de reacción",type="r")
```



No parece haber una relación ya que hay líneas que cortan el eje Y en valores similares pero tienen pendientes diferentes.

- Haga un gráfico que relacione los interceptos de `beta0` con las pendientes de `beta1`. Además obtenga el coeficiente de correlación de estos dos vectores.

```
plot(beta1,beta0,pch=18)
abline(lm(beta0~beta1))
```




```
cor(beta0,beta1)
```

```
## [1] -0.1375534
```

El coeficiente de correlación es bastante bajo y el gráfico no muestra una relación estrecha entre pendientes e interceptos.

- Ajuste un modelo lineal ordinario con tiempo de reacción en función del tiempo (días) sin tomar en cuenta la tendencia en cada sujeto. Escriba la ecuación resultante, la cual deberá comparar más adelante con los resultados obtenidos en el modelo mixto.

```
mod1=lm(reac~dias)
mod1$coef
```

```
## (Intercept)        dias
##   251.40510    10.46729
```

La ecuación general es:

$$\hat{y} = 251{,}4 + 10{,}5X$$

- Se van a ajustar dos modelos: 1) un modelo mixto asumiendo que hay correlación entre las pendientes y los interceptos, 2) un modelo que asume que no hay correlación entre las pendientes y los interceptos. Ambos modelos se deben ajustar con máxima verosimilitud, pues la idea es compararlos con la prueba de razón de verosimilitud (LRT). Use la función `lmer` de la librería `lme4`. Para el primer modelo se pone la línea de tendencia general que se considera un efecto fijo, luego se toma el sujeto como efecto aleatorio y dentro de cada sujeto se estima una regresión. El 1 representa el intercepto y se puede omitir, pero aunque se omita siempre van a estimarse los interceptos: `lmer(reac~1+dias+(1+dias|sujeto),REML=F)` o `lmer(reac~dias+(dias|sujeto),REML=F)`. Se agrega `REML=F` para indicar que se haga el ajuste por máxima verosimilitud en vez de usar REML.

```
library(lme4)
mod2=lmer(reac~1+dias+(1+dias|sujeto),REML=F)
```

- Obtenga el `summary` de este modelo y observe el valor de la correlación entre pendientes e interceptos. Puede extraer la parte aleatoria con `summary(mod)$varcor`.

```
summary(mod2)$varcor
```

```
##  Groups   Name        Std.Dev. Corr
##  sujeto   (Intercept) 23.7806
##           dias         5.7168  0.081
##  Residual             25.5918
```

Se obtiene una correlación de 0.08 la cual es muy baja como se esperaba de los gráficos.

- Ahora ajuste el modelo pero asumiendo que no hay correlación entre pendientes e interceptos. Para esto debe ajustar los interceptos dentro de cada sujeto de forma independiente de las pendientes: `lmer(reac~1+dias+(1|sujeto)+(0+dias|sujeto),REML=F)`. El cero en la parte de las pendientes obliga a que no estime el intercepto junto con la pendiente dentro de cada sujeto.

```
mod3=lmer(reac~1+dias+(1|sujeto)+(0+dias|sujeto),REML=F)
```

- Compare los dos modelos anteriores con la prueba LRT. Se usa un `anova` de los dos modelos anidados y se indica `test="LRT"`.

```
anova(mod2,mod3,test="LRT")
```

```
## Data: NULL
## Models:
## mod3: reac ~ 1 + dias + (1 | sujeto) + (0 + dias | sujeto)
## mod2: reac ~ 1 + dias + (1 + dias | sujeto)
##      Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## mod3  5 1762.0 1778.0 -876.00   1752.0
## mod2  6 1763.9 1783.1 -875.97   1751.9 0.0639      1     0.8004
```

Al comparar las deviancias de estos dos modelos se obtiene una diferencia de 0.0639, con un grado de libertad, lo cual arroja una probabilidad asociada de 0.80. No se rechaza la hipótesis nula que establece que no hay correlación entre pendientes e interceptos. Por lo tanto, se asume que las pendientes y los interceptos son independientes y se usa el segundo modelo (`mod3`).

---

3. Ahora se quiere verificar si se puede asumir que todos los sujetos tienen la misma pendiente. Para esto se ajusta un modelo donde se elimina la pendiente de la parte aleatoria y luego se compara con el `mod3`, que fue el ganador en el punto anterior.

```
mod4=lmer(reac~1+dias+(1|sujeto),REML=F)
anova(mod3,mod4)
```

```
## Data: NULL
## Models:
## mod4: reac ~ 1 + dias + (1 | sujeto)
## mod3: reac ~ 1 + dias + (1 | sujeto) + (0 + dias | sujeto)
##      Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## mod4  4 1802.1 1814.8 -897.04   1794.1
## mod3  5 1762.0 1778.0 -876.00   1752.0 42.075      1  8.782e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Al comparar los dos modelos se obtiene una diferencia de las dos deviancias de 42.1, con 1 grado de libertad y una probabilidad asociada muy pequeña (p<0.001). Por lo tanto, se concluye que los sujetos no tienen la misma pendiente y se prefiere el `mod3`.

- Interprete lo que significa el resultado anterior en términos del problema.

Puesto que cada sujeto tiene una pendiente diferente, se puede esperar que conforme pasan los días, el tiempo de reacción aumenta a una tasa diferente entre los conductores. Esto indica que la privación de sueño afecta de forma diferente a los conductores.

---

4. Estime el modelo escogido con REML para hacer las interpretaciones finales. Escriba la ecuación general de la parte fija. Puede extraer los coeficientes fijos con `summary(mod)$coef`.

```
mod5=lmer(reac~1+dias+(1|sujeto)+(0+dias|sujeto))
summary(mod5)$coef
```

```
##              Estimate Std. Error   t value
## (Intercept) 251.40510   6.885382 36.512879
## dias         10.46729   1.559566  6.711666
```

La ecuación general es igual a la obtenida con la regresión ordinaria:

$$\hat{y} = 251{,}4 + 10{,}5X$$

- Compare los errores estándar de los coeficientes obtenidos con la regresión ordinaria y con el modelo mixto.

```
ee1=summary(mod1)$coef[,2]
ee5=summary(mod5)$coef[,2]
cbind(ee1,ee5)
```

```
##                  ee1      ee5
## (Intercept) 6.610154 6.885382
## dias        1.238195 1.559566
```

Los errores estándar son más grandes en el modelo mixto, lo cual tiene sentido porque se tienen individuos tomados de forma aleatoria de una población mayor en lugar de considerarlos fijos. De esta forma hay una incertidumbre mayor en las estimaciones de los parámetros la cual se ve reflejada en los errores estándar. Aunque estos resultados son menos precisos con el modelo mixto, son más adecuados que la regresión ordinaria, puesto que reflejan el comportamiento aleatorio de los individuos.

- Obtenga intervalos de 95 % de confianza para los parámetros del modelo e interprételos. Use `confint(profile(mod))`. Aquí va a obtener primero los intevalos para las desviaciones estándar de los componentes aleatorios llamados `.sig01` para los interceptos aleatorios y `.sig02` para las pendientes aleatorias, seguida de la desviación estándar residual. Luego seguirán los términos de la parte fija que son el intercepto general y la pendiente general.

```
confint(profile(mod5))
```

```
##                   2.5 %     97.5 %
## .sig01        15.258647  37.786473
## .sig02         3.964074   8.769159
## .sigma        22.880555  28.787598
## (Intercept) 237.572148 265.238062
## dias          7.334067  13.600505
```

Los primeros dos intervalos indican la importancia de los interceptos y pendientes aleatorias, los cuales tienen una variabilidad no despreciable. En la parte fija el coeficiente más importante es la pendiente que, si bien es cierto que puntualmente da igual que con un modelo ordinario, el intervalo cambia puesto que tiene un error estándar más grande. En este caso se espera con 95 % de confianza que, en promedio, para todos los conductores de la población, por cada día adicional el tiempo de reacción promedio aumente entre 7.3 y 13.6 milisegundos.

Conclusión: en el grupo de conductores a los que se les restringió el número de horas de sueño a 3 horas, no hay una relación estrecha entre el tiempo de reacción al inicio del experimento y la rapidez en que ese tiempo aumenta. Además la privación de sueño afecta a los conductores de diferentes maneras, hay conductores que son más sensibles y conforme pasan los días van perdiendo su capacidad de reaccionar más rápidamente que otros (aumenta el tiempo de reacción a una mayor tasa). Se puede obtener una estimación de la velocidad en que este tiempo de reacción va creciendo en promedio para el grupo en general.

6.2. Ortodoncia

Se realizó un estudio para dar seguimiento al crecimiento óseo de la maxila o mandíbula de 27 niños (16 hombres y 11 mujeres) desde los 8 a los 14 años. Cada dos años se midió la distancia entre la pituitaria y la escotadura pterygomaxilar, dos puntos que son fácilmente identificados con rayos X. Esta distancia se utiliza para conocer el grado de maduración esquelética del individuo y está medida en milímetros. Se compara el ritmo de crecimiento de esta distancia entre hombres y mujeres.

Ejercicios

1. Abra el archivo `ortodoncia.Rdata`.

---

2. Haga un gráfico donde se pueda ver el crecimiento de la distancia en función del tiempo. En el gráfico se deben apreciar las diferencias entre los diferentes niños sin importar el sexo. Primero se tiene que hacer una nueva variable llamada edad1=edad-8 de tal forma que el intercepto represente la distancia a la edad inicial que es 8 años.

- A partir de lo que se observa en el gráfico, ¿qué se puede adelantar sobre el crecimiento de la distancia? ¿Se puede decir que entre más alta la distancia a los 8 años va a haber un mayor crecimiento de esa distancia en el tiempo?

---

3. Haga gráficos donde se pueda ver el crecimiento de la distancia en función del tiempo para hombres y para mujeres. En un caso ponga las líneas de cada sujeto y en otro ponga solo la línea de tendencia de hombres y de mujeres.

- A partir de los gráficos, ¿se puede adelantar que en alguno de los sexos haya un mayor crecimiento de la distancia?
- Usando un modelo verifique si la distancia crece al mismo ritmo para hombres y para mujeres. En su análisis debe decidir si descarta la posibilidad de correlación entre pendientes e interceptos.
- Escriba el modelo utilizado para las observaciones individuales.
- Escriba el modelo para la media condicional de cada sexo.
- Estime cuánto crece la distancia cada año en promedio entre los hombres y cuánto crece entre mujeres. Para hacer las estimaciones debe usar un modelo que se haya estimado con REML.
- Estime la distancia promedio a los 8 años para cada sexo. ¿En cuánto difiere la distancia entre hombres y mujeres a esa edad? ¿En cuánto difiere la distancia a los 14 años?

Solución

1. Abra el archivo `ortodoncia.Rdata`.

```
load("ortodoncia.Rdata")
```

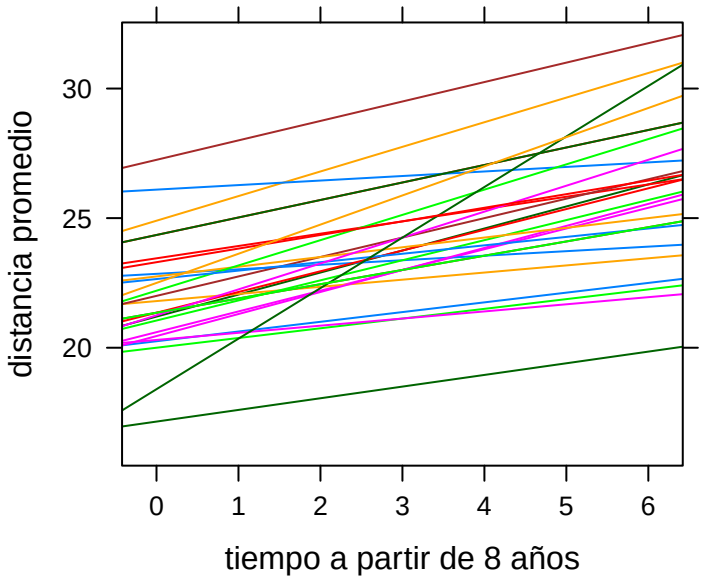
---

2. Haga un gráfico donde se pueda ver el crecimiento de la distancia en función del tiempo. En el gráfico se deben apreciar las diferencias entre los diferentes niños sin importar el sexo. Primero se tiene que hacer una nueva variable llamada edad1=edad-8 de tal forma que el intercepto represente la distancia a la edad inicial que es 8 años.

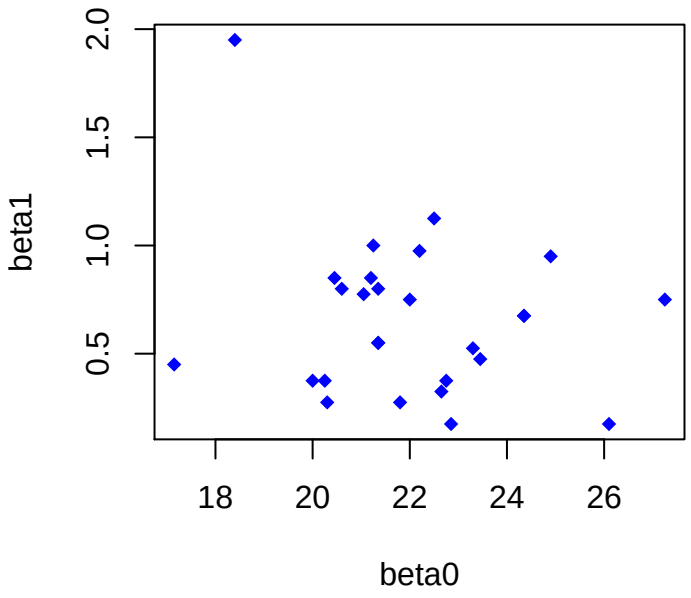
```
library(lattice)
base$edad1=base$edad-8
attach(base)
beta0=beta1=c()
length(table(sujeto))
```

```
## [1] 27
```

```
suj=as.numeric(sujeto)
for(i in 1:27)  {
  mod=lm(distancia~edad1,base[suj==i,])
  beta0[i]=mod$coef[1]
  beta1[i]=mod$coef[2]
}
xyplot(distancia~edad1,group=sujeto,pch=18,
       xlab="tiempo a partir de 8 años",
       ylab="distancia promedio",type="r")
```



```
plot(beta0,beta1,pch=18, col=4)
```




```
cor(beta0,beta1)
```

```
## [1] -0.2083781
```

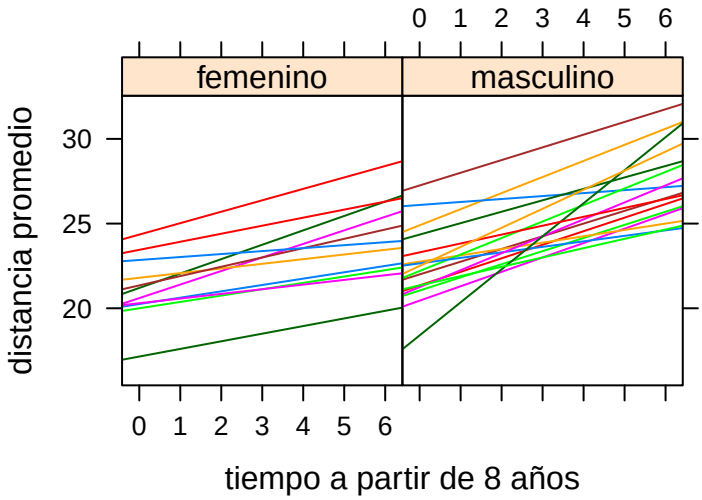
- A partir de lo que se observa en el gráfico, ¿qué se puede adelantar sobre el crecimiento de la distancia? ¿Se puede decir que entre más alta la distancia a los 8 años va a haber un mayor crecimiento de esa distancia en el tiempo?

En general se ve que las pendientes no van decreciendo conforme aumenta el intercepto. La correlación es -0.2 la cual no es muy alta. No parece haber una relación fuerte entre la distancia a los 8 años y la pendiente de cada niño o niña.

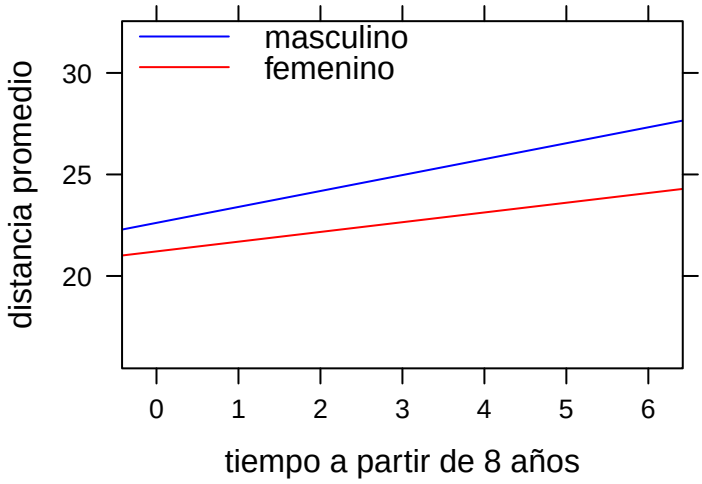
---

3. Haga gráficos donde se pueda ver el crecimiento de la distancia en función del tiempo para hombres y para mujeres. En un caso ponga las líneas de cada sujeto y en otro ponga solo la línea de tendencia de hombres y de mujeres.

```
xyplot(distancia~edad1|sexo,group=sujeto,pch=18,
       xlab="tiempo a partir de 8 años",
       ylab="distancia promedio",type=c("r"))
```




```
xyplot(distancia~edad1,group=sexo,col=c(2,4),
       xlab="tiempo a partir de 8 años",
       ylab="distancia promedio",type="r",
       key=list(corner=c(0,1),lines=list(col=c(4,2),lty=1),
                text=list(c("masculino","femenino"))))
```



- A partir de los gráficos, ¿se puede adelantar que en alguno de los sexos haya un mayor crecimiento de la distancia?

En los gráficos se puede ver que las líneas de crecimiento solo para hombres o solo para mujeres muestran pendientes muy parecidas dentro de cada grupo. Además, haciendo solo la línea de tendencia para hombres y mujeres (segundo gráfico) se nota un crecimiento más alto para los hombres. Podría ser que la interacción fija entre edad y sexo sea significativa. De todas formas se puede apreciar cómo independientemente de la edad, la distancia promedio para los hombres es siempre mayor que la de las mujeres.

- Usando un modelo verifique si la distancia crece al mismo ritmo para hombres y para mujeres. En su análisis debe decidir si descarta la posibilidad de correlación entre pendientes e interceptos.

Primero se descarta que haya correlación entre las pendientes y los interceptos.

```
library(lme4)
 options(contrasts=c("contr.sum","contr.poly"))
 mod1=lmer(distancia~edad1*sexo+(1+edad1|sujeto),REML=F)
 mod2=lmer(distancia~edad1*sexo+(1|sujeto)+(0+edad1|sujeto),REML=F)
 anova(mod1,mod2)
```

```
## Data: NULL
## Models:
## mod2: distancia ~ edad1 * sexo + (1 | sujeto) + (0 + edad1 | sujeto)
## mod1: distancia ~ edad1 * sexo + (1 + edad1 | sujeto)
##      Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## mod2  7 441.81 460.58 -213.91   427.81
## mod1  8 443.81 465.26 -213.90   427.81 0.0031      1     0.9555
```

Como la probabilidad asociada al comparar los dos modelos es alta (0.96), no se rechaza la hipótesis que dice que no hay correlación entre pendientes e interceptos. Se escoge el `mod2` para continuar con el análisis. Luego se descarta que las pendientes sean diferentes entre niños de un mismo sexo. Se hace un modelo que no tenga las pendientes aleatorias (`mod3`) y se compara con el modelo que sí tenía esas pendientes (`mod2`).

```
  mod3=lmer(distancia~edad1*sexo+(1|sujeto),REML=F)
  anova(mod3,mod2)
```

```
## Data: NULL
## Models:
## mod3: distancia ~ edad1 * sexo + (1 | sujeto)
## mod2: distancia ~ edad1 * sexo + (1 | sujeto) + (0 + edad1 | sujeto)
##      Df    AIC    BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
## mod3  6 440.64 456.73 -214.32   428.64
## mod2  7 441.81 460.58 -213.91   427.81  0.83      1     0.3623
```

Nuevamente la probabilidad es alta (0.36), y se escoge el modelo más simple que descarta pendientes específicas para cada sujeto. Ahora se verifica si la interacción entre edad y sexo es significativa. Para esto se hace un modelo similar al `mod3` pero donde se ha quitado la interacción entre edad y sexo (`mod4`).

```
  mod4=lmer(distancia~edad1+sexo+(1|sujeto),REML=F)
  anova(mod4,mod3)
```

```
## Data: NULL
## Models:
## mod4: distancia ~ edad1 + sexo + (1 | sujeto)
## mod3: distancia ~ edad1 * sexo + (1 | sujeto)
##      Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## mod4  5 444.86 458.27 -217.43   434.86
## mod3  6 440.64 456.73 -214.32   428.64 6.2174      1    0.01265 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Al obtener una probabilidad tan baja (0.013), se rechaza la hipótesis de no interacción. Como la interacción sí es significativa, se concluye que la tasa de crecimiento de la distancia no es la misma en niños que en niñas.

- Escriba el modelo utilizado para las observaciones individuales.

En el modelo que se utiliza para la Y, i representa el individuo, j representa el sexo y E corresponde a la edad:

$$Y_{ij,E} = \beta_0 + \beta_1 E + \tau_j + \tau_j^* E + \beta_{0,i} + e_{ij} = (\beta_0 + \beta_{0,i}) + \tau_j + (\beta_1 + \tau_j^*)E + e_{ij}$$

- Escriba el modelo para la media condicional de cada sexo.

Se verifica que el código 1 es para mujeres y -1 para hombres.

```
contrasts(sexo)
```

```
##           [,1]
## femenino     1
## masculino   -1
```

Hombres:

$\mu_{H,E} = \beta_0 + \beta_1 E - \tau_1 - \tau_1^* E = (\beta_0 - \tau_1) + (\beta_1 - \tau_1^*)E$

Mujeres:

$\mu_{M,E} = \beta_0 + \beta_1 E + \tau_1 + \tau_1^* E = (\beta_0 + \tau_1) + (\beta_1 + \tau_1^*)E$

- Estime cuánto crece la distancia cada año en promedio entre los hombres y cuánto crece entre mujeres. Para hacer las estimaciones debe usar un modelo que se haya estimado con REML.

De los coeficientes de la parte fija se pueden derivar las ecuaciones para los promedios estimados de la distancia para hombres y para mujeres.

```
mod5=lmer(distancia~edad1*sexo+(1|sujeto))
summary(mod5)$coef
```

```
##               Estimate Std. Error   t value
## (Intercept) 21.9123580 0.42203167 51.921122
## edad1        0.6319602 0.06071045 10.409415
## sexo1       -0.7032670 0.42203167 -1.666385
## edad1:sexo1 -0.1524148 0.06071045 -2.510520
```

Hombres:

$\hat{y}_H = (21{,}912 + 0{,}703) + (0{,}632 + 0{,}152)E = 22{,}615 + 0{,}784E$

Mujeres:

$\hat{y}_M = (21{,}912 - 0{,}703) + (0{,}632 - 0{,}152)E = 21{,}209 + 0{,}480E$

De forma puntual, se puede decir que en los hombres, por cada año, la distancia aumenta en promedio 0.784mm, mientras que en las mujeres, esta distancia aumenta en promedio 0.48mm por año. Se puede ver que la tasa de aumento para los niños es mayor.

- Estime la distancia promedio a los 8 años para cada sexo. ¿En cuánto difiere la distancia entre hombres y mujeres a esa edad? ¿En cuánto difiere la distancia a los 14 años?

```
22.615-21.209
```

```
## [1] 1.406
```

```
22.615-21.209+(0.784-0.480)*10
```

```
## [1] 4.446
```

La distancia a los 8 años está dada por los interceptos de las ecuaciones escritas anteriormente. A los 8 años la distancia promedio para los hombres es 22.615mm y para las mujeres 21.209mm, entonces el promedio de esta distancia es 1.406mm mayor para los hombres que para las mujeres. A los 14 años esta diferencia aumenta considerablemente, ya que ahora la distancia promedio para los hombres es 4.446mm mayor que para las mujeres.

Conclusión: desde que los niños entran al estudio a los 8 años se tiene una diferencia importante en la distancia media estudiada entre hombres y mujeres; sin embargo, con el tiempo, al ir creciendo esta distancia también se acentúan las diferencias.

6.3. Arbustos

Se hace un estudio para analizar el efecto que tiene la presencia de herbívoros y depredadores en la cobertura de arbustos. Se investigan 3 tratamientos y se desea detectar si la disminución en la cobertura de arbustos pueden reducirse con la aplicación de alguno de esos tratamientos. El primer tratamiento es un control que contiene depredadores y herbívoros (C), el segundo consiste en excluir los depredadores de la parcela (nD) y el tercero consiste en excluir los herbívoros de la parcela (nH). Se cuenta con 4 parcelas en cada uno de los 3 tratamientos. Además, se cuenta con la cantidad de lluvia (en milímetros) en el día de la medición en cada una de las parcelas. Se midió e porcentaje de cobertura en las parcelas cada año durante el período 2001-2013.

Ejercicios

1. Abra el archivo `arbustos.Rdata`.

- Primero se tiene que hacer una nueva variable llamada `tiempo`, de tal forma que el intercepto represente la cobertura en el año inicial que es 2001, entonces `tiempo=año-2001`. También hay que notar que se tienen 12 parcelas en la variable `parcela`, las cuales están enumeradas de 1 a 15, pero faltan la 4, la 7 y la 12. Para evitar confusiones con esto, hay que hacer un truco: convertir primero `parcela` a factor y luego a numérica para que asigne de forma corrida los números del 1 al 12.

---

2. Haga un gráfico donde se pueda ver el cambio de cobertura en función del tiempo. En el gráfico se deben apreciar las diferencias entre las diferentes parcelas sin importar el tratamiento. A partir de lo que se observa en el gráfico, ¿qué se puede adelantar sobre el cambio en la cobertura?

- Haga la prueba formal para decidir si se puede decir que entre más alta la cobertura al inicio va a haber un mayor aumento de esa cobertura en el tiempo. Use `periodo`, además incluya en todos los análisis la variable `lluvia` como una covariable que puede estar metiendo ruido.

---

3. Haga un gráfico donde se pueda ver el cambio en la cobertura en función del tiempo para cada tratamiento. A partir del gráfico, ¿se puede adelantar que en alguno de los tratamientos hay un mayor decrecimiento de la cobertura?

---

4. Estimaciones.

- Escriba el modelo utilizado que incluya interacción entre tiempo y tratamiento.
- Usando el modelo, estime cuánto crece o decrece la cobertura cada año en promedio en cada tratamiento.

---

5. ¿Se puede concluir que el crecimiento o decrecimiento de la cobertura es más rápido en alguno de los tratamientos?

- Dé una estimación de la tasa de crecimiento general y construya un intervalo de confianza.

Solución

1. Abra el archivo `arbustos.Rdata`.

```
load("arbustos.Rdata")
```

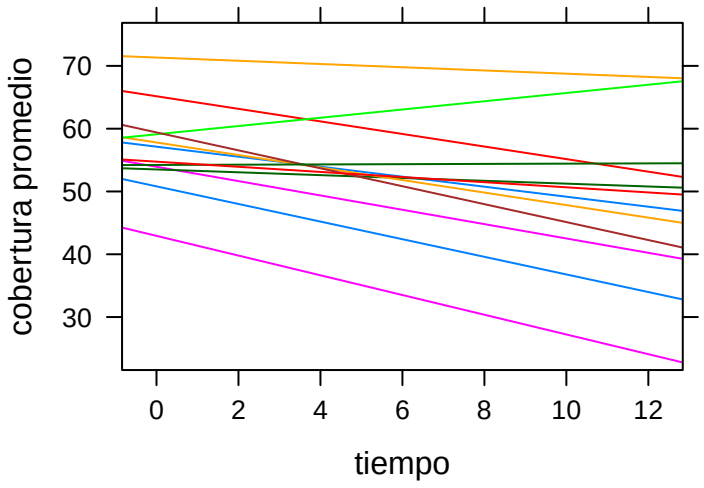
- Primero se tiene que hacer una nueva variable llamada `tiempo`, de tal forma que el intercepto represente la cobertura en el año inicial que es 2001, entonces `tiempo=año-2001`. También hay que notar que se tienen 12 parcelas en la variable `parcela`, las cuales están enumeradas de 1 a 15, pero faltan la 4, la 7 y la 12. Para evitar confusiones con esto, hay que hacer un truco: convertir primero `parcela` a factor y luego a numérica para que asigne de forma corrida los números del 1 al 12.

```
base$tiempo=base$año-2001
base$parcela=as.numeric(factor(base$parcela))
attach(base)
```

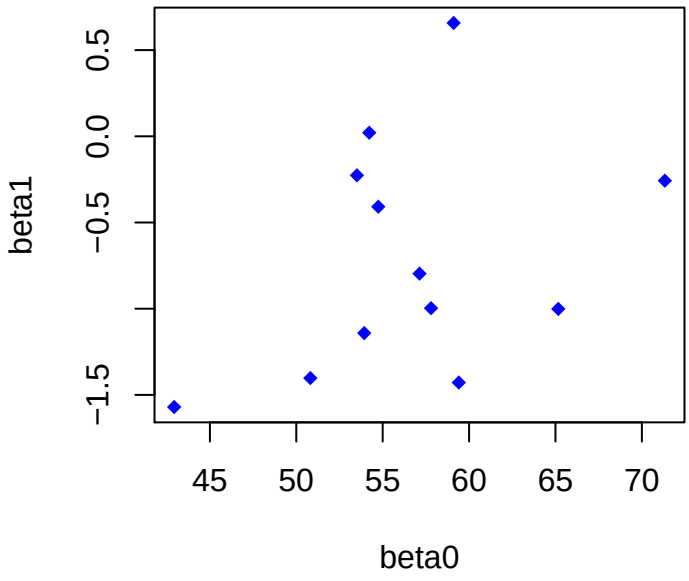
---

2. Haga un gráfico donde se pueda ver el cambio de cobertura en función del tiempo. En el gráfico se deben apreciar las diferencias entre las diferentes parcelas sin importar el tratamiento. A partir de lo que se observa en el gráfico, ¿qué se puede adelantar sobre el cambio en la cobertura?

```
beta0=beta1=c()
for(i in 1:12)  {
  mod=lm(tcov~tiempo,base[parcela==i,])
  beta0[i]=mod$coef[1]
  beta1[i]=mod$coef[2]
}
xyplot(tcov~tiempo,group=parcela,pch=18,
       ylab="cobertura promedio",type=c("r"))
```



```
plot(beta0,beta1,pch=18,col=4)
```




```
cor(beta0,beta1)
```

```
## [1] 0.3476235
```

En general, hay una leve tendencia a que las pendientes vayan creciendo conforme aumenta el intercepto, sin embargo, la correlación entre pendientes e interceptos no es demasiado alta (0.35).

- Haga la prueba formal para decidir si se puede decir que entre más alta la cobertura al inicio va a haber un mayor aumento de esa cobertura en el tiempo. Use `periodo`, además incluya en todos los análisis la variable `lluvia` como una covariable que puede estar metiendo ruido.

```
options(contrasts=c("contr.sum","contr.poly"))
mod1=lmer(tcov~tiempo*trt+lluvia+(1+tiempo|factor(parcela)),REML=F)
mod2=lmer(tcov~tiempo*trt+lluvia+(1|factor(parcela))+
            (0+tiempo|factor(parcela)),REML=F)
```

```r
anova(mod1,mod2)
```

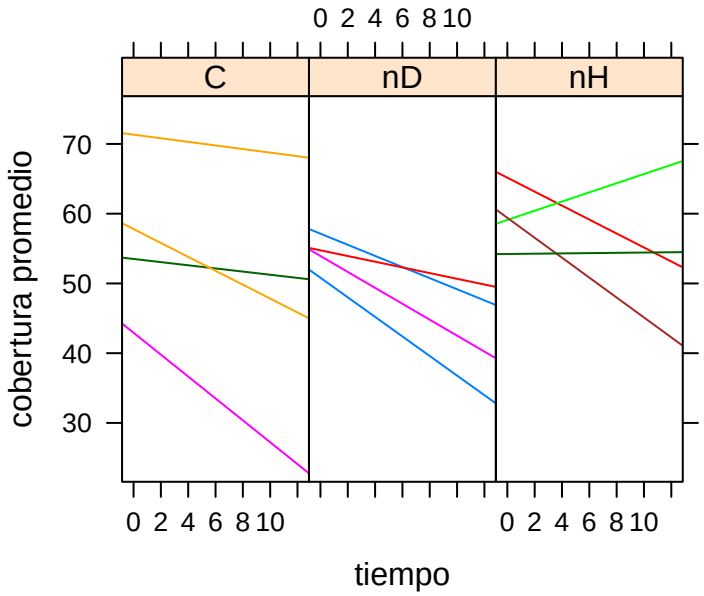
```
## Data: NULL
## Models:
## mod2: tcov ~ tiempo * trt + lluvia + (1 | factor(parcela)) + (0 + tiempo |
## mod2:     factor(parcela))
## mod1: tcov ~ tiempo * trt + lluvia + (1 + tiempo | factor(parcela))
##      Df    AIC    BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
## mod2 10 957.79 988.28 -468.89   937.79
## mod1 11 957.85 991.40 -467.93   935.85 1.934      1     0.1643
```

Al hacer la prueba formal para ver si la correlación entre pendientes e interceptos es nula, se tiene que no hay evidencia para rechazar esa hipótesis, por lo que se puede asumir, en lo sucesivo, que ambos parámetros son independientes.

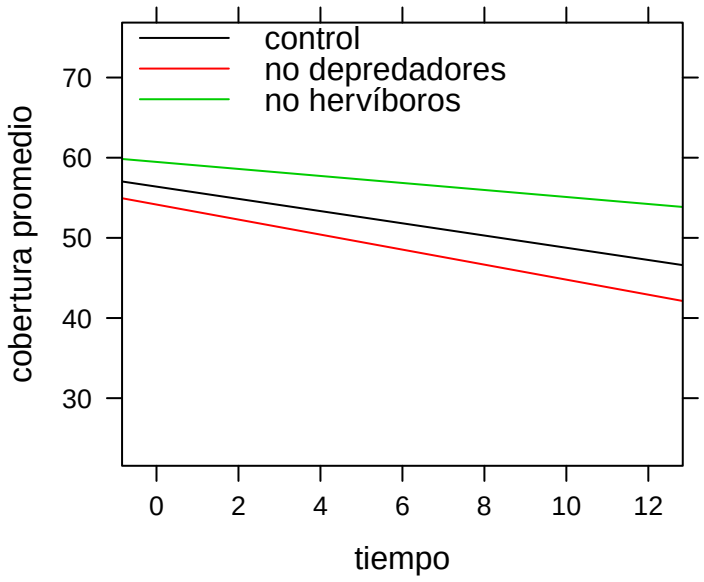
---

3. Haga un gráfico donde se pueda ver el cambio en la cobertura en función del tiempo para cada tratamiento. A partir del gráfico, ¿se puede adelantar que en alguno de los tratamientos hay un mayor decrecimiento de la cobertura?

```r
xyplot(tcov~tiempo|trt,group=parcela,pch=18,
       ylab="cobertura promedio",type=c("r"),layout=c(3,1))
```



```
xyplot(tcov~tiempo,group=trt,col=1:3,ylab="cobertura promedio",type=c("r"),
       key=list(corner=c(0,1),lines=list(col=1:3,lty=1),
                text=list(c("control","no depredadores","no hervíboros"))))
```



En los primeros gráficos se puede ver que en algunas parcelas la cobertura baja con el tiempo, en otras se mantiene e incluso en una aumenta. Eso hace pensar que las líneas no son paralelas y que se debería mantener el coeficiente aleatorio para pendientes por parcela. Además, la tendencia general promedio es a bajar de forma muy similar en los tres tratamientos. La que baja más suavemente es la del tratamiento C y las otras dos son muy parecidas

---

4. Estimaciones.

- Escriba el modelo utilizado que incluya interacción entre tiempo y tratamiento.

En el modelo se utiliza i para la parcela, j para el tratamiento, T para el tiempo y L para la lluvia que es una covariable.

$$Y_{ij,T,L} = (\beta_0 + \beta_{0,i}) + (\beta_1 + \tau_j^* + \beta_{1,i})T + \beta_2 L + e_{ij}$$

$$\mu_{j,T,L} = \beta_0 + (\beta_1 + \tau_j^*)T + \beta_2 L$$

- Usando el modelo, estime cuánto crece o decrece la cobertura cada año en promedio en cada tratamiento.

En un punto anterior se descartó que haya correlación entre las pendientes y los interceptos, por lo que siempre se parte de un modelo que asume que no existe esta correlación. Luego se verifica si las pendientes son diferentes entre parcelas de un mismo tratamiento.

```
mod3=lmer(tcov~tiempo*trt+lluvia+(1|factor(parcela)),REML=F)
anova(mod3,mod2)
```

```
## Data: NULL
## Models:
## mod3: tcov ~ tiempo * trt + lluvia + (1 | factor(parcela))
## mod2: tcov ~ tiempo * trt + lluvia + (1 | factor(parcela)) + (0 + tiempo |
## mod2:     factor(parcela))
##      Df    AIC     BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## mod3  9 977.67 1005.12 -479.83   959.67
## mod2 10 957.79  988.28 -468.89   937.79 21.884      1  2.896e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se puede rechazar la hipótesis de que estas pendientes son iguales, y se concluye que no todas las rectas son paralelas, es decir, que al pasar el tiempo, la tasa de disminución en la cobertura promedio no se da por igual en todas las parcelas con un mismo tratamiento.

A pesar de la conclusión anterior, se quiere ver cuál es la tendencia de la cobertura a lo largo del tiempo en los distintos tratamientos. Para ver la tendencia por tratamiento se ven los coeficientes fijos tomando en cuenta la interacción entre `tiempo` y `trt`.

```
summary(mod2)$coef
```

```
##                 Estimate Std. Error    t value
## (Intercept) 55.451585756 2.19338130 25.2813251
## tiempo      -0.635986510 0.19622424 -3.2411210
## trt1        -0.281913919 2.74881731 -0.1025583
## trt2        -2.518040293 2.74881731 -0.9160450
## lluvia       0.005867431 0.00487769  1.2029119
## tiempo:trt1 -0.050151099 0.26250794 -0.1910460
## tiempo:trt2 -0.224642857 0.26250794 -0.8557564
```

```
  b=summary(mod2)$coef[,1]
  b
```

```
##  (Intercept)       tiempo         trt1         trt2       lluvia
## 55.451585756 -0.635986510 -0.281913919 -2.518040293  0.005867431
##  tiempo:trt1  tiempo:trt2
## -0.050151099 -0.224642857
```

```
  b[2]+b[6]
```

```
##     tiempo
## -0.6861376
```

```
  b[2]+b[7]
```

```
##     tiempo
## -0.8606294
```

```
  b[2]-b[6]-b[7]
```

```
##     tiempo
## -0.3611926
```

Las pendientes para `periodo` según tratamiento son: -0.69 para trt1 (C), -0.86 para trt2 (nD), y -0.36 para trt3 (nH).

Se ve puntualmente que los tratamientos que presentan decrecimientos más pronunciados son C y nD.

---

5. ¿Se puede concluir que el crecimiento o decrecimiento de la cobertura es más rápido en alguno de los tratamientos?

Para contestar a esta pregunta basta ver si la interacción entre `tiempo` y `trt` está presente.

```
  mod4=lmer(tcov~tiempo+trt+lluvia+(1|factor(parcela))+
              (0+tiempo|factor(parcela)),REML=F)
  anova(mod4,mod2)
```

```
## Data: NULL
## Models:
## mod4: tcov ~ tiempo + trt + lluvia + (1 | factor(parcela)) + (0 + tiempo |
## mod4:     factor(parcela))
## mod2: tcov ~ tiempo * trt + lluvia + (1 | factor(parcela)) + (0 + tiempo |
## mod2:     factor(parcela))
##      Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## mod4  8 954.97 979.37 -469.48   938.97
## mod2 10 957.79 988.28 -468.89   937.79 1.1838      2     0.5533
```

Al comparar un modelo con interacción contra el modelo que no la tiene, se obtiene una probabilidad asociada de 0.55, por lo que no existe evidencia para decir que hay interacción y se puede asumir que todos los tratamientos presentan una tendencia decreciente similar.

- Dé una estimación de la tasa de decrecimiento general y construya un intervalo de confianza.

```
summary(mod4)$coef[2,]
```

```
##   Estimate Std. Error    t value
## -0.6359865  0.2049787 -3.1026955
```

```
confint(profile(mod4))[5,]
```

```
##      2.5 %     97.5 %
## -1.0661251 -0.2058479
```

La tasa de decrecimiento anual se estima en 0.64 %. Con 95 % de confianza se espera que la tasa de decrecimiento anual esté entre 0.21 % y 1.07 % para cualquier tratamiento.

Conclusión: hay mucha variabilidad en la forma en que la cobertura decrece en las distintas parcelas ya que se prueba que no todas tienen el mismo ritmo de cambio, sin embargo, como comportamiento promedio por tratamiento, no se logra diferenciar entre ellos si este ritmo de cambio es mayor para alguno de los tratamientos.

6.4. Riqueza

Se llama bentos a la comunidad formada por los organismos que habitan el fondo de los ecosistemas acuáticos. Se utilizan datos de bentos marino procedente de nueve playas (zonas intermareales) de la costa holandesa recogidos por el instituto holandés RIKZ en el verano de 2002. En cada playa se tomaron cinco muestras de la macro-fauna y variables abióticas.

Se quiere ver si existe alguna relación entre la riqueza de especies y la altura de cada estación de muestreo con respecto al nivel medio de la marea (`NAP`). Como la riqueza de especies es un conteo, sería más apropiado utilizar un modelo lineal generalizado (GLM) con una distribución Poisson. Sin embargo, para simplificar utilizaremos un modelo con errores normales.

Ejercicios

1. Abra el archivo `riqueza.Rdata`.

- La riqueza de especies se puede medir como el número de especies registradas en un sitio y en un momento dado. Utilice las columnas 2 a 76 que contienen el número de individuos registrados (abundancia) para cada una de las 75 especies presentes. Busque una forma creativa de obtener la variable riqueza a partir de estas 75 variables sin necesidad de contar manualmente. Para cada línea debe tomar en cuenta las especies que registran al menos un individuo.
- Haga un gráfico con una línea de regresión por playa, donde se muestren los puntos de las observaciones, para determinar si se justifica una relación lineal entre riqueza promedio y NAP. Use `riq~NAP|playa` en la función `xyplot` de la librería `lattice`.

---

2. Haga un gráfico con todas las líneas de regresión en un solo gráfico para tratar de visualizar si existe una relación entre el valor de la pendiente con el intercepto. Use `riq~NAP, group=playa` en la función `xyplot`.

- Obtenga el ajuste de las regresiones por separado para cada playa y guarde en dos vectores diferentes las pendientes y los interceptos. Haga un gráfico para visualizar la relación entre pendientes e interceptos. ¿Tiene sentido pensar en una correlación entre interceptos y pendientes?

- Obtenga dos modelos para probar si es conveniente considerar la correlación entre pendientes e interceptos. En el primer modelo se permite correlación, por lo que se indica en la parte aleatoria simplemente intercepto (1) y pendiente (NAP) dentro de la misma playa con `(1+NAP|playa)`. El modelo se escribe `mod1=lmer(riq~1+NAP+(1+NAP|playa),REML=F)`. En el segundo modelo se deben especificar el intercepto y la pendiente por separado para que no haya correlación, pero en la parte de la pendiente se debe poner un 0 para indicar que ahí no se quiere intercepto, entonces se escribe `mod2=lmer(riq~1+NAP+(1|playa)+(0+NAP|playa),REML=F)`. En ambos casos se pone `REML=F`, pues el objetivo es comparar ambos modelos mediante la prueba de razón de verosimilitud y para esto se requiere que el ajuste no sea por REML, sino por máxima verosimilitud.
- Compare ambos modelos mediante la prueba de razón de verosimilitud (LRT).

---

3. Pruebe si el NAP tiene un aporte importante en el modelo. Primero se prueba si se puede considerar que todas las rectas son paralelas, usando un modelo en el que se elimina la pendiente en la parte aleatoria (`mod3`). Lo adecuado es comparar este nuevo modelo con el modelo que permite que haya correlación entre interceptos y pendientes (`mod1`), puesto eso fue lo que se concluyó en el punto anterior.

- Obtenga los intervalos de confianza para los parámetros del modelo e interprételos.

---

4. Ahora vamos a realizar la prueba de interacción entre playa y NAP de la forma en que muchas veces se realiza, sin considerar la estructura de dependencia de las observaciones dentro de una misma playa. Ponga playa como un factor en un modelo lineal ordinario donde se incluya la interacción entre playa y NAP. Utilice el modelo suma para que tenga sentido la pendiente general.

- Haga la prueba para verificar si existe interacción entre playa y NAP. Compare el resultado con el obtenido en el modelo mixto.
- Haga el intervalo de 95 % de confianza para la pendiente general. Compárelo con el obtenido con el modelo mixto.
- ¿Qué problemas existen?

Solución

1. Abra el archivo `riqueza.Rdata`.

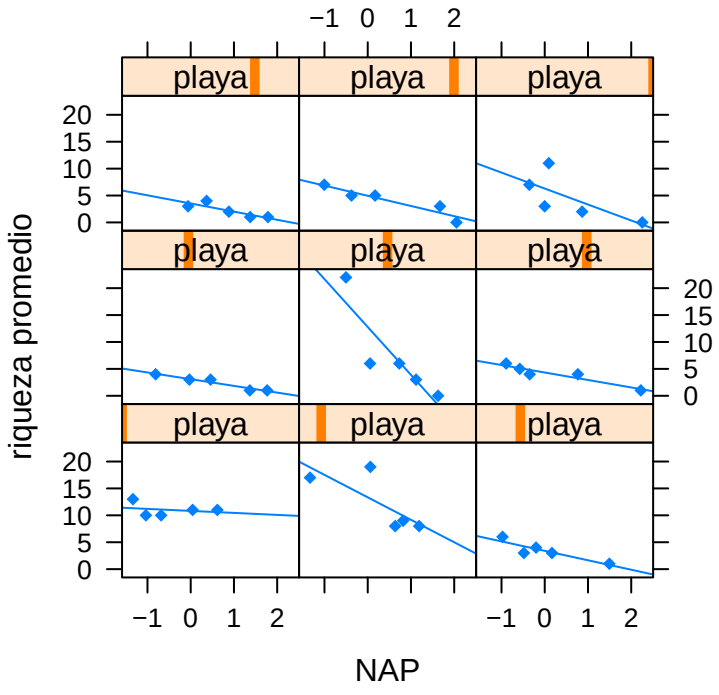
```
load("riqueza.Rdata")
```

- La riqueza de especies se puede medir como el número de especies registradas en un sitio y en un momento dado. Utilice las columnas 2:76 que contiene el número de individuos registrados (abundancia) para cada una de las 75 especies presentes. Busque una forma creativa de obtener la variable riqueza a partir de estas 75 variables sin necesidad de contar manualmente. Para cada línea debe tomar en cuenta las especies que registran al menos un individuo.

```
base$riq = apply(base[, 2:76] > 0,1,sum)
attach(base)
```

- Haga un gráfico con una línea de regresión por playa, donde se muestren los puntos de las observaciones, para determinar si se justifica una relación lineal entre riqueza promedio y NAP. Use `riq~NAP|playa` en la función `xyplot` de la librería `lattice`.

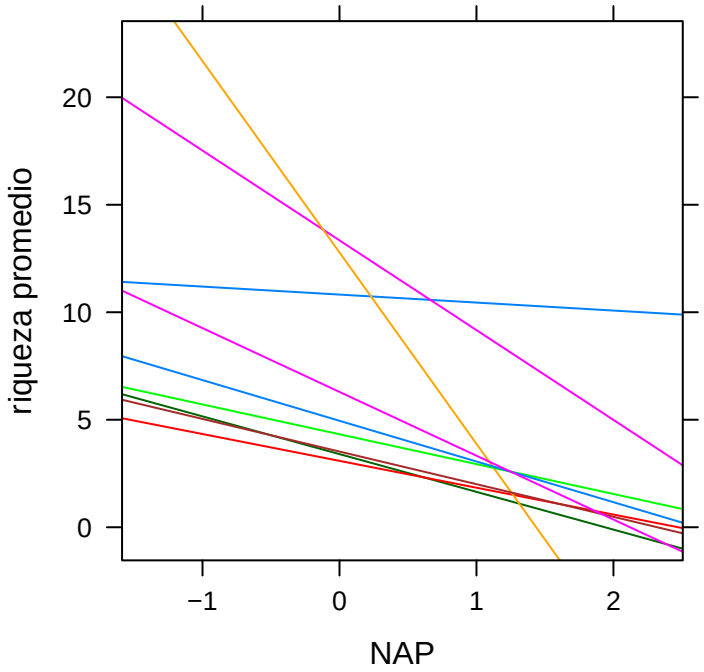
```
library(lattice)
xyplot(riq~NAP|playa,pch=18,ylab="riqueza promedio",type=c("p","r"))
```



En la mayoría de las playas, los puntos siguen una tendencia bastante lineal, por lo que parece acertado seguir con una regresión lineal.

2. Haga un gráfico con todas las líneas de regresión en un solo gráfico para tratar de visualizar si existe una relación entre el valor de la pendiente con el intercepto. Use `riq~NAP, group=playa` en la función `xyplot`.

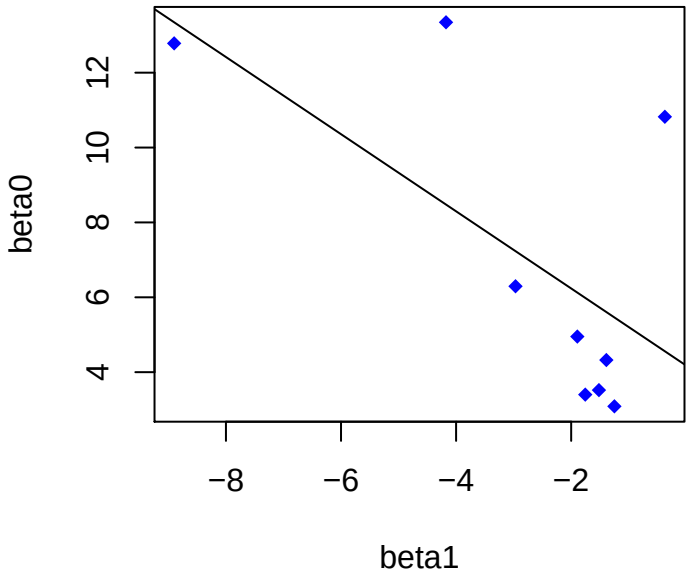
```
library(lattice)
xyplot(riq~NAP,group=playa,pch=18,ylab="riqueza promedio",type=c("r"))
```



Hay que observar que el NAP puede tomar valores negativos, por lo que el gráfico puede ser engañoso para visualizar los interceptos, ya que estos deben ubicarse a la altura del NAP igual a cero. Se puede ver que la mayoría de las playas tienen pendientes muy bajas, excepto dos que presentan pendientes muy fuertes y a la vez son las que tienen el intercepto más alto. Esto indica que en esas playas el aumento del NAP provoca un fuerte decrecimiento en la riqueza promedio, mientras que en las otras playas el aumento del NAP hace que la riqueza promedio baje muy poco. Parece que sí existe una correlación no despreciable entre interceptos y pendientes.

- Obtenga el ajuste de las regresiones por separado para cada playa y guarde en dos vectores diferentes las pendientes y los interceptos. Haga un gráfico para visualizar la relación entre pendientes e interceptos. ¿Tiene sentido pensar en una correlación entre interceptos y pendientes?

```
beta0=beta1=c()
for(i in 1:9)  {
  mod=lm(riq~NAP,base[base$playa==i,])
  beta0[i]=mod$coef[1]
  beta1[i]=mod$coef[2]
}
plot(beta1,beta0,pch=18,col=4)
abline(lm(beta0~beta1))
```




```
cor(beta0,beta1)
```

```
## [1] -0.632037
```

Parece que sí existe una correlación entre interceptos y pendientes, porque las playas con pendientes más pronunciadas (aunque negativas) también tienen interceptos más altos. Se puede notar que la correlación entre ambos es -0.63, con lo cual se confirma la observación del gráfico.

- Obtenga dos modelos para probar si es conveniente considerar la correlación entre pendientes e interceptos. En el primer modelo se permite correlación, por lo que se indica en la parte aleatoria simplemente intercepto (1) y pendiente (NAP) dentro de la misma playa con `(1+NAP|playa)`. El modelo se escribe `mod1=lmer(riq~1+NAP+(1+NAP|playa),REML=F)`. En el segundo modelo se deben especificar el intercepto y la pendiente por separado para que no haya correlación, pero en la parte de la pendiente se debe poner un 0 para indicar que ahí no se quiere intercepto, entonces se escribe `mod2=lmer(riq~1+NAP+(1|playa)+(0+NAP|playa),REML=F)`. En ambos casos se pone `REML=F`, pues el objetivo es comparar ambos modelos mediante la prueba de razón de verosimilitud y para esto se requiere que el ajuste no sea por REML, sino por máxima verosimilitud.

```
library(lme4)
mod1=lmer(riq~1+NAP+(1+NAP|playa),REML=F)
mod2=lmer(riq~1+NAP+(1|playa)+(0+NAP|playa),REML=F)
```

- Compare ambos modelos mediante la prueba de razón de verosimilitud (LRT).

```
anova(mod1,mod2,test="LRT")
```

```
## Data: NULL
## Models:
## mod2: riq ~ 1 + NAP + (1 | playa) + (0 + NAP | playa)
## mod1: riq ~ 1 + NAP + (1 + NAP | playa)
##      Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## mod2  5 251.21 260.25 -120.61   241.21
## mod1  6 246.66 257.50 -117.33   234.66 6.5556      1    0.01046 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Al comparar el modelo que considera correlación entre pendientes e interceptos contra el modelo que las considera independientes, se obtiene un valor de la probabilidad asociada a la prueba LRT de 0.01, con lo cual se puede rechazar la hipótesis de no correlación. Se concluye que sí existe correlación entre pendientes e interceptos.

---

3. Pruebe si el NAP tiene un aporte importante en el modelo. Primero se prueba si se puede considerar que todas las rectas son paralelas, usando un modelo en el que se elimina la pendiente en la parte aleatoria (`mod3`). Lo adecuado es comparar este nuevo modelo con el modelo que permite que haya correlación entre interceptos y pendientes (`mod1`), puesto eso fue lo que se concluyó en el punto anterior.

```
mod3=lmer(riq~1+NAP+(1|playa),REML=F)
anova(mod1,mod3,test="LRT")
```

```
## Data: NULL
## Models:
## mod3: riq ~ 1 + NAP + (1 | playa)
## mod1: riq ~ 1 + NAP + (1 + NAP | playa)
##      Df    AIC    BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
## mod3  4 249.83 257.06 -120.92   241.83
## mod1  6 246.66 257.50 -117.33   234.66 7.173      2    0.02769 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Al hacer esta comparación se rechaza la hipótesis nula que dice que todas la pendientes son iguales, lo cual tiene sentido, pues se había encontrado que las playas que tienen una riqueza más alta en el nivel medio de la marea, van a tender a disminuir más rápidamente la riqueza conforme aumenta el NAP. Evidentemente no todas las playas van a tener una misma pendiente, sino que esta depende en cierta medida de la riqueza que la playa tenga cuando el NAP es cero. Por lo tanto, se debe mantener el primer modelo (`mod1`).

- Obtenga los intervalos de confianza para los parámetros del modelo e interprételos.

```
confint(profile(mod1))
```

```
##                  2.5 %     97.5 %
## .sig01        1.959389  6.0172181
## .sig02       -1.000000 -0.3593062
## .sig03        0.440480  3.4051925
## .sigma        2.068354  3.4418049
## (Intercept)  3.971115  9.1793732
## NAP         -4.409507 -1.3556937
```

En la parte aleatoria, `.sig01` es la desviación estándar de los interceptos aleatorios, `.sig02` es la correlación entre interceptos y pendientes, `.sig03` es la desviación estándar de las pendientes aleatorias y `.sigma` es la desviación estándar del error. Es evidente que el intervalo para la correlación no incluye al cero pues va de -1 a -0.36, lo que indica que sí hay una correlación importante entre los interceptos y las pendientes.

El intercepto representa la riqueza promedio cuando el NAP es cero, es decir cuando se ubica en el nivel medio de la marea, entonces, entre más alta es la riqueza promedio al nivel medio de la marea, se va a tener un mayor decrecimiento en la riqueza al aumentar el NAP, lo cual tiene sentido.

Se puede dar una interpretación al coeficiente fijo de NAP que corresponde a la pendiente general fija. Aquí se puede decir que como medida general de la tendencia (sin tomar en cuenta una playa particular), al aumentar el NAP en una unidad, la cantidad de especies disminuye en promedio entre 1.3 y 4.4 especies, con 95 % de confianza. Además, el intercepto general está entre 3.97 y 9.18 lo cual representa el rango en que se espera que se encuentre el número de especies promedio cuando se ubica en el nivel medio de la marea.

---

4. Ahora vamos a realizar la prueba de interacción entre playa y NAP de la forma en que muchas veces se realiza, sin considerar la estructura de dependencia de las observaciones dentro de una misma playa. Ponga playa como un factor en un modelo lineal ordinario donde se incluya la interacción entre playa y NAP. Utilice el modelo suma para que tenga sentido la pendiente general.

```
options(contrasts=c("contr.sum","contr.poly"))
mod4=lm(riq~NAP*factor(playa))
```

- Haga la prueba para verificar si existe interacción entre playa y NAP. Compare el resultado con el obtenido en el modelo mixto.

```
drop1(mod4,test="F")
```

```
## Single term deletions
##
## Model:
## riq ~ NAP * factor(playa)
##                   Df Sum of Sq    RSS     AIC F value   Pr(>F)
## <none>                          165.92  94.719
## NAP:factor(playa)  8    161.82 327.74 109.350  3.2915 0.009434 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Si se pone playa como factor y se incluyen interacciones, se puede hacer la prueba para ver si hay interacción entre playa y NAP, y se observa que sí es significativa. La conclusión es la misma que se obtuvo anteriormente, pero la probabilidad asociada se triplica (0.0094 vs 0.0028).

- Haga el intervalo de 95 % de confianza para la pendiente general. Compárelo con el obtenido con el modelo mixto.

```
confint(mod4)[2,]
```

```
##     2.5 %    97.5 %
## -3.583646 -1.798116
```

La tendencia general del NAP es negativa con un intervalo entre -1.8 y -3.6. Este resultado es similar al obtenido con el modelo mixto (-1.4, -4.4) pero no es igual. De hecho, el intervalo del modelo mixto es más amplio porque permite más variabidad entre las pendientes debido a que las 9 playas son solo una muestra de un conjunto mayor de playas y eso debe considerarse en las posibles pendientes

- ¿Qué problemas existen?

El problema principal con un modelo ordinario es que los resultados son válidos solo para esas 9 playas. Además, no se está tomando en cuenta la correlación que existe en las observaciones dentro de cada playa sino que se asume que son independientes.

Conclusión: se verificó que los datos son consistentes con una tendencia lineal para justificar el uso del modelo de regresión lineal. El modelo utilizado considera que hay correlación entre los interceptos y las pendientes de las regresiones asociadas a cada playa. Esto se justifica porque las playas que tienen una mayor riqueza promedio cuando se encuentra al nivel promedio de la marea, son las que experimentan una mayor disminución de la riqueza promedio al aumentar el NAP.

BIBLIOGRAFIA

Adler, D., Murdoch, D. and others (2017). rgl: 3D Visualization Using OpenGL. R package version 0.98.1. https://CRAN.R-project.org/package=rgl

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Kleiber, C. & Zeileis, A. (2008). Applied Econometrics with R. New York: Springer-Verlag. ISBN 978-0-387-77316-2. URL https://CRAN.R-project.org/package=AER

Lenth, R. V. (2009). Response-Surface Methods in R, Using rsm. Journal of Statistical Software, 32(7), 1-17. URL http://www.jstatsoft.org/v32/i07/.

Pinheiro J., Bates D., DebRoy S., Sarkar D. and R Core Team (2017). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-131, <URL: https://CRAN.R-project.org/package=nlme>.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sarkar, D. (2008). Lattice: Multivariate Data Visualization with R. Springer, New York. ISBN 978-0-387-75968-5

Venables, W. N. & Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0