PAC6

Maria Lucas Gascón

2024-01-17

# Problema 1

Cargamos los datos del problema.

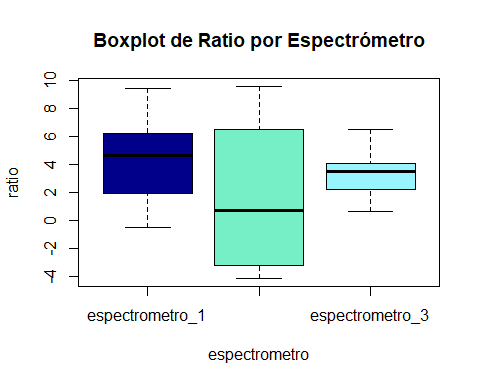
# Seteamos el directorio de trabajo  
setwd("D:/Antiguos estudios/MASTER2/Sem3/Diseño/PAC6/PAC6")  
  
# Importamos los datos  
data <- read.csv("dades\_problema1\_pec6.csv", dec = ",", header = TRUE, sep = ";")  
  
# Marcamos los factores  
data$espectrometro <- as.factor(data$espectrometro)  
data$suelo <- as.factor(data$suelo)  
data$dia <- factor(data$dia, levels = c(1, 2, 3), labels = c("dia1", "dia2", "dia3"))

## 1.1 Análisis descriptivo de los datos

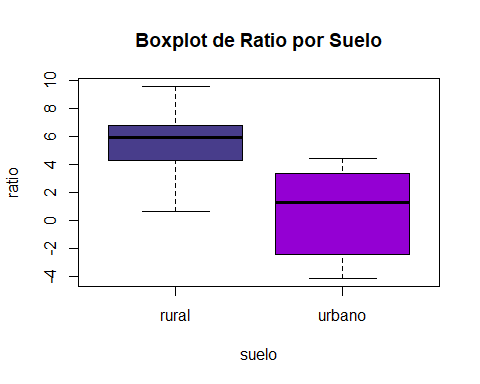
# Primero examinamos la variable ratio  
summary(data)

## ratio espectrometro suelo dia   
## Min. :-4.1356 espectrometro\_1:12 rural :18 dia1:12   
## 1st Qu.: 0.9882 espectrometro\_2:12 urbano:18 dia2:12   
## Median : 3.5102 espectrometro\_3:12 dia3:12   
## Mean : 3.1112   
## 3rd Qu.: 5.7919   
## Max. : 9.5738

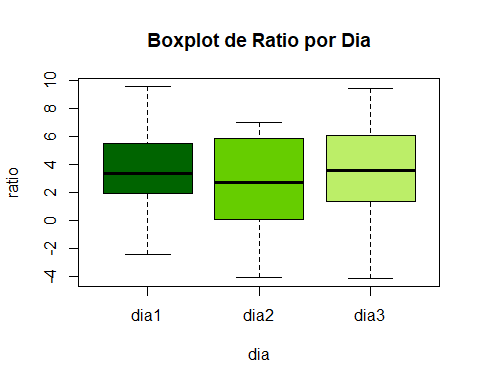
# Box plot de ratio por espectrometro  
boxplot(ratio ~ espectrometro, data = data, col = c("blue4", "aquamarine2", "cadetblue1"), main = "Boxplot de Ratio por Espectrómetro")



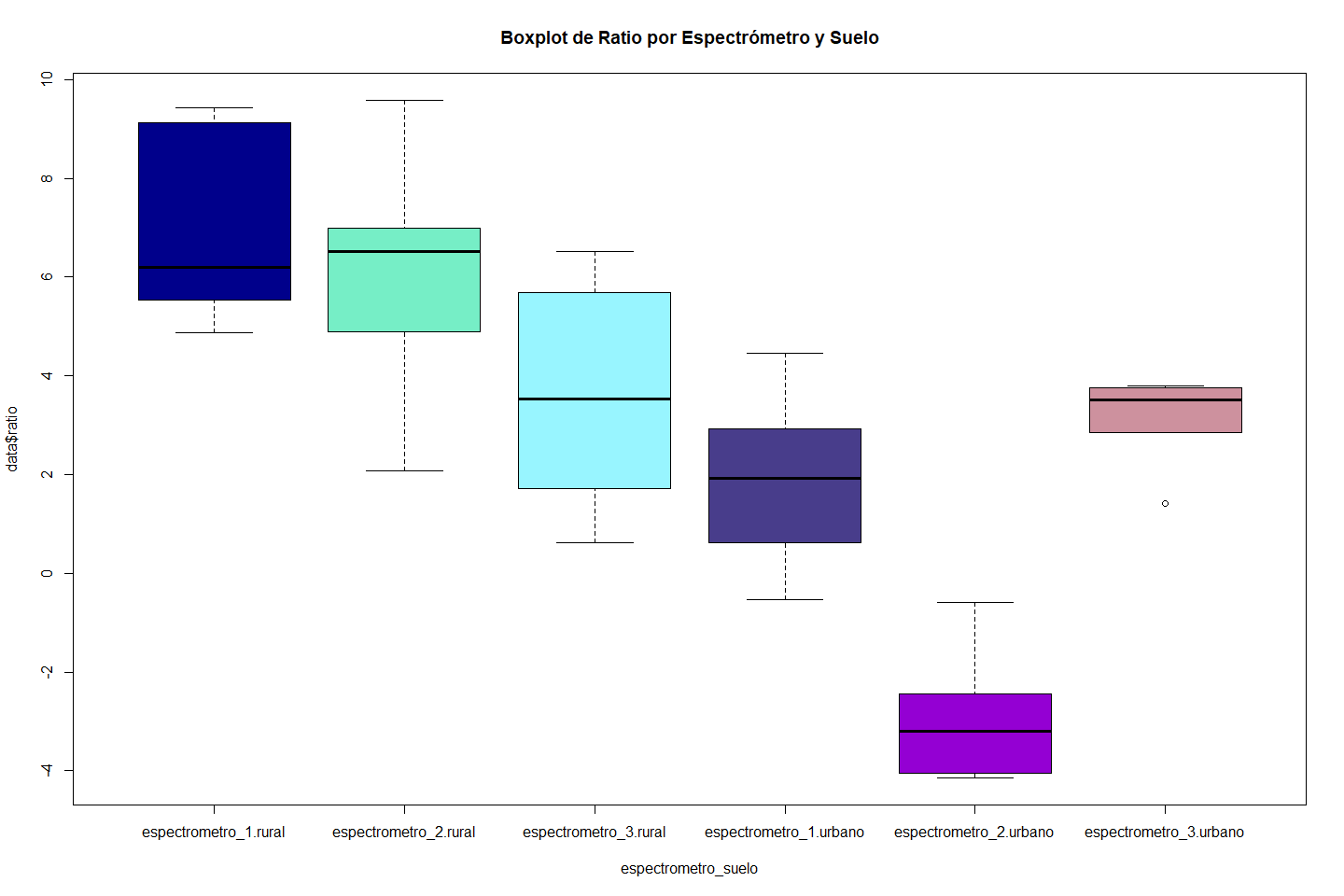
# Box plot de ratio por suelo  
boxplot(ratio ~ suelo, data = data, col = c("darkslateblue", "darkviolet"), main = "Boxplot de Ratio por Suelo")



# Box plot de ratio por dia  
boxplot(ratio ~ dia, data = data, col = c("darkgreen", "chartreuse3", "darkolivegreen2"), main = "Boxplot de Ratio por Dia")



# Create an interaction factor of espectrometro and suelo  
espectrometro\_suelo <- interaction(data$espectrometro, data$suelo)  
  
# Box plot of ratio by espectrometro and suelo  
boxplot(data$ratio ~ espectrometro\_suelo, col = c("blue4", "aquamarine2", "cadetblue1", "darkslateblue", "darkviolet", "pink3"),   
 main = "Boxplot de Ratio por Espectrómetro y Suelo")



En el análisis del boxplot correspondiente al “Ratio por Espectrómetro”, se observa una ligera disparidad en las mediciones entre los espectrómetros. Esta discrepancia se evidencia mediante la diferencia en la altura de las cajas y la variación en la media de las mediciones, destacada por una línea negra en cada caja.

En cuanto al boxplot relativo al “Ratio por Suelo”, se destaca que el suelo rural presenta, en general, un ratio de 14N a 15N superior al observado en el suelo urbano. Esta diferencia se refleja tanto en la altura como en el ancho de las cajas, indicando que el suelo rural exhibe un ratio más elevado en comparación con el suelo urbano.

Adicionalmente, se observa una notable consistencia en las mediciones diarias, ya que muestran una variación mínima.

Al combinar los datos de espectrómetro y suelo en un único gráfico, se puede evaluar la interacción entre ambos factores. Se aprecia que, para el suelo urbano, las mediciones de los espectrómetros presentan una mayor disparidad en comparación con el suelo rural, donde las mediciones son más uniformes. Específicamente, se destaca que el espectrómetro 2 en el suelo urbano proporciona mediciones notablemente más bajas.

## 1.2 Modelo lineal

En este problema, se busca modelar la proporción de 14N a 15N en dos tipos de suelos (urbano, rural) utilizando tres espectrómetros distintos. Se realizaron mediciones en 18 muestras de cada tipo de suelo. Cada muestra se asignó a uno de los tres espectrómetros al azar, y para cada combinación de suelo y espectrómetro, se eligieron 3 días al azar para realizar las mediciones. El objetivo es analizar la influencia de los factores suelo, espectrómetro y día, así como sus interacciones.

El modelo lineal puede expresarse de la siguiente manera:

* es la medida de la proporción de 14N a 15N para la k-ésima observación en el suelo i, espectrómetro j y día k.
* es la media global.
* es el efecto del i-ésimo nivel del factor suelo. La suma de todos los efectos del factor suelo debe ser cero:
* es el efecto del j-ésimo nivel del factor espectrómetro. La suma de todos los efectos del factor espectrómetro debe ser cero:
* es la interacción entre el suelo i y el espectrómetro j. La suma de las interacciones debe ser cero:
* es el efecto del k-ésimo nivel del factor día anidado en la combinación suelo i y espectrómetro j. Como es un efecto aleatorio, no hay restricciones específicas, pero se asume que sigue una distribución normal con media cero:
* es el error aleatorio asociado con la k-ésima observación en la combinación suelo i, espectrómetro j y día k. Se asume que sigue una distribución normal con media cero:

## 1.3 Hipótesis de interés

Para el factor suelo:

* . No hay diferencia significativa en la proporción de 14N a 15N entre los suelos urbanos y rurales.
* Almenos un . Hay al menos una diferencia significativa en la proporción de 14N a 15N entre los suelos urbanos y rurales.

Para el factor espectrómetro:

* . No hay diferencia significativa en la proporción de 14N a 15N entre los espectrómetros.
* Almenos un . Hay al menos una diferencia significativa en la proporción de 14N a 15N entre los tres espectrómetros distintos.

Para la interacción suelo-espectrómetro:

* para todos los . No hay interacción significativa entre el tipo de suelo y el espectrómetro.
* Existe al menos una . Hay al menos una interacción significativa entre el tipo de suelo y el espectrómetro.

Para el efecto aleatorio del día:

* . No hay variabilidad significativa asociada con los días anidados en la combinación de suelo y espectrómetro.
* . Existe variabilidad significativa asociada con los días anidados en la combinación de suelo y espectrómetro.

Para el error aleatorio:

* . No hay variabilidad significativa no explicada por los factores considerados en el modelo.
* . Existe variabilidad significativa no explicada por los factores considerados en el modelo.

## 1.4 Estudiar las hipótesis

# Definimos el modelo  
library(lme4)

## Loading required package: Matrix

library(car)

## Loading required package: carData

modelo <- lmer(ratio ~ suelo + espectrometro + suelo:espectrometro + (1 | dia:suelo:espectrometro), data = data)  
  
summary(modelo)

## Linear mixed model fit by REML ['lmerMod']  
## Formula:   
## ratio ~ suelo + espectrometro + suelo:espectrometro + (1 | dia:suelo:espectrometro)  
## Data: data  
##   
## REML criterion at convergence: 128.3  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -1.8173 -0.5379 0.0471 0.4748 1.8520   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## dia:suelo:espectrometro (Intercept) 2.071 1.439   
## Residual 1.841 1.357   
## Number of obs: 36, groups: dia:suelo:espectrometro, 18  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 6.8943 0.9985 6.904  
## suelourbano -5.0119 1.4122 -3.549  
## espectrometroespectrometro\_2 -0.8031 1.4122 -0.569  
## espectrometroespectrometro\_3 -3.2959 1.4122 -2.334  
## suelourbano:espectrometroespectrometro\_2 -4.0167 1.9971 -2.011  
## suelourbano:espectrometroespectrometro\_3 4.5518 1.9971 2.279  
##   
## Correlation of Fixed Effects:  
## (Intr) sulrbn espc\_2 espc\_3 slr:\_2  
## suelourbano -0.707   
## espctrmtr\_2 -0.707 0.500   
## espctrmtr\_3 -0.707 0.500 0.500   
## slrbn:spc\_2 0.500 -0.707 -0.707 -0.354   
## slrbn:spc\_3 0.500 -0.707 -0.354 -0.707 0.500

# Intervalos de confianza  
library(broom)

## Warning: package 'broom' was built under R version 4.3.2

confint(modelo)

## Computing profile confidence intervals ...

## 2.5 % 97.5 %  
## .sig01 0.000000 1.8099950  
## .sigma 1.010479 1.9580075  
## (Intercept) 5.207159 8.5814518  
## suelourbano -7.397879 -2.6259082  
## espectrometroespectrometro\_2 -3.189038 1.5829331  
## espectrometroespectrometro\_3 -5.681922 -0.9099516  
## suelourbano:espectrometroespectrometro\_2 -7.391029 -0.6424435  
## suelourbano:espectrometroespectrometro\_3 1.177523 7.9261089

En la síntesis del modelo, no se evidencian los factores que resultan significativos. No obstante, es posible calcular los valores de confianza mediante el empleo de un valor crítico de la distribución t de Student. En esta distribución, un nivel de confianza del 0.05 implicaría un valor crítico de 1.96.

El intervalo de confianza se formula como el estimado del coeficiente más o menos 1.96 veces el error estándar del coeficiente. Para mayor conveniencia, este cálculo se realiza de forma automática mediante el paquete “broom”.

Al examinar los intervalos de confianza, se constata que todos los factores son estadísticamente significativos, a excepción del espectrómetro 2, ya que su intervalo de confianza incluye el valor 0. Por otro lado, el intercepto, el suelo urbano, el espectrómetro 3 y las interacciones suelo urbano:espectrómetro 2 y suelo urbano:espectrómetro 3, exhiben diferencias significativas en comparación con la referencia establecida, que corresponde al suelo rural y el espectrómetro 1.

# Testamos las hipótesis  
anova\_modelo <- Anova(modelo, type = "III", test = "Chisq")  
print(anova\_modelo)

## Analysis of Deviance Table (Type III Wald chisquare tests)  
##   
## Response: ratio  
## Chisq Df Pr(>Chisq)   
## (Intercept) 47.6697 1 5.044e-12 \*\*\*  
## suelo 12.5961 1 0.0003866 \*\*\*  
## espectrometro 5.9247 2 0.0516970 .   
## suelo:espectrometro 18.4324 2 9.942e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

La especificación de la ANOVA de “tipo III” implica la atribución de variabilidad a cada término, teniendo en cuenta todos los demás términos, incluidas las interacciones que involucran dicho término. Esta aproximación resulta útil cuando se presentan interacciones y se busca evaluar la contribución de cada variable a la variabilidad explicada, considerando las otras variables y sus interacciones.

Cabe destacar que, dado que no se han verificado las suposiciones de normalidad y homocedasticidad, se prefiere utilizar la prueba de Chi-cuadrado en lugar de la prueba F.

En cuanto a los resultados, se observa que tanto el suelo como la interacción suelo:espectrofotómetro son significativas (valor de p < 0.05), mientras que el espectrofotómetro por sí mismo no lo es (valor de p > 0.05). A continuación, se detalla la interpretación de estos hallazgos:

* Suelo significativo: Se constatan diferencias estadísticamente significativas en la proporción de 14N a 15N entre los suelos urbanos y rurales (valor de p < 0.05, rechazo de H0), después de considerar los demás efectos en el modelo. En resumen, existe evidencia estadística que respalda la afirmación de que el tipo de suelo tiene un impacto significativo en la proporción de nitrógeno.
* Espectrómetro no significativo: No se dispone de suficiente evidencia estadística para sostener que existen diferencias en la proporción de 14N a 15N entre los distintos espectrofotómetros (valor de p > 0.05, aceptación de H0), después de tomar en cuenta los otros efectos. En resumen, las variaciones observadas en la proporción de nitrógeno no pueden atribuirse de manera significativa a las diferencias entre los espectrofotómetros utilizados. Es posible que la no significancia se deba a que el espectrofotómetro 3 no difiere, a pesar de que el 2 sí lo hace según el intervalo de confianza.
* Interacción suelo-espectrómetro significativa: Se evidencia que la influencia del espectrofotómetro en la proporción de nitrógeno depende del tipo de suelo (valor de p < 0.05, rechazo de H0). En otras palabras, la diferencia en la proporción de 14N a 15N entre los espectrofotómetros puede variar según si el suelo es urbano o rural. Esta interacción indica que el efecto del espectrofotómetro no es constante en todos los suelos y viceversa.

En resumen, los resultados sugieren que el tipo de suelo y la interacción entre el tipo de suelo y el espectrofotómetro son factores significativos en la variabilidad de la proporción de 14N a 15N, mientras que el espectrofotómetro por sí mismo no tiene un efecto significativo en esta proporción, aunque este patrón parece no aplicar al espectrofotómetro 2.

## 1.5 Comparaciones múltiples

Dado que el factor suelo consta únicamente de dos niveles, carece de fundamento llevar a cabo una comparación múltiple. A pesar de que el espectrómetro no muestra significancia de manera general, hemos observado que el espectrómetro 2 presenta un significado estadístico; procederemos a verificar estos resultados. Por último, examinaremos también la interacción suelo:espectrómetro.

#### Para la interacción:

# install.packages("emmeans")  
library(emmeans)

## Warning: package 'emmeans' was built under R version 4.3.2

library(knitr)

## Warning: package 'knitr' was built under R version 4.3.2

# Realizamos las comparaciones múltiples  
emmeans\_result <- emmeans(modelo, pairwise ~ suelo:espectrometro)  
  
# Ajustamos p-valores usando la corrección de Bonferroni  
adjusted\_emmeans <- summary(emmeans\_result, infer = c(TRUE, TRUE), adjust = "bonferroni")  
  
# Imprimimos los valores ajustados  
print(adjusted\_emmeans)

## $emmeans  
## suelo espectrometro emmean SE df lower.CL upper.CL t.ratio p.value  
## rural espectrometro\_1 6.89 0.999 12 3.74620 10.042 6.904 0.0001  
## urbano espectrometro\_1 1.88 0.999 12 -1.26569 5.031 1.885 0.5031  
## rural espectrometro\_2 6.09 0.999 12 2.94315 9.239 6.100 0.0003  
## urbano espectrometro\_2 -2.94 0.999 12 -6.08548 0.211 -2.942 0.0740  
## rural espectrometro\_3 3.60 0.999 12 0.45026 6.746 3.604 0.0217  
## urbano espectrometro\_3 3.14 0.999 12 -0.00981 6.286 3.143 0.0509  
##   
## Degrees-of-freedom method: kenward-roger   
## Confidence level used: 0.95   
## Conf-level adjustment: bonferroni method for 6 estimates   
## P value adjustment: bonferroni method for 6 tests   
##   
## $contrasts  
## contrast estimate SE df lower.CL  
## rural espectrometro\_1 - urbano espectrometro\_1 5.012 1.41 12 -0.141  
## rural espectrometro\_1 - rural espectrometro\_2 0.803 1.41 12 -4.350  
## rural espectrometro\_1 - urbano espectrometro\_2 9.832 1.41 12 4.679  
## rural espectrometro\_1 - rural espectrometro\_3 3.296 1.41 12 -1.857  
## rural espectrometro\_1 - urbano espectrometro\_3 3.756 1.41 12 -1.397  
## urbano espectrometro\_1 - rural espectrometro\_2 -4.209 1.41 12 -9.362  
## urbano espectrometro\_1 - urbano espectrometro\_2 4.820 1.41 12 -0.333  
## urbano espectrometro\_1 - rural espectrometro\_3 -1.716 1.41 12 -6.869  
## urbano espectrometro\_1 - urbano espectrometro\_3 -1.256 1.41 12 -6.409  
## rural espectrometro\_2 - urbano espectrometro\_2 9.029 1.41 12 3.876  
## rural espectrometro\_2 - rural espectrometro\_3 2.493 1.41 12 -2.660  
## rural espectrometro\_2 - urbano espectrometro\_3 2.953 1.41 12 -2.200  
## urbano espectrometro\_2 - rural espectrometro\_3 -6.536 1.41 12 -11.689  
## urbano espectrometro\_2 - urbano espectrometro\_3 -6.076 1.41 12 -11.228  
## rural espectrometro\_3 - urbano espectrometro\_3 0.460 1.41 12 -4.693  
## upper.CL t.ratio p.value  
## 10.165 3.549 0.0600  
## 5.956 0.569 1.0000  
## 14.985 6.962 0.0002  
## 8.449 2.334 0.5670  
## 8.909 2.660 0.3120  
## 0.944 -2.980 0.1721  
## 9.973 3.413 0.0772  
## 3.437 -1.215 1.0000  
## 3.897 -0.889 1.0000  
## 14.181 6.393 0.0005  
## 7.646 1.765 1.0000  
## 8.106 2.091 0.8769  
## -1.383 -4.628 0.0087  
## -0.923 -4.302 0.0154  
## 5.613 0.326 1.0000  
##   
## Degrees-of-freedom method: kenward-roger   
## Confidence level used: 0.95   
## Conf-level adjustment: bonferroni method for 15 estimates   
## P value adjustment: bonferroni method for 15 tests

Como podemos ver, las siguientes combinaciones resultan significativas:

# Extraemos los contrastes y p-valores  
contrast\_df <- as.data.frame(adjusted\_emmeans$contrasts)  
  
# Filtramos por p-valores < 0.05  
significant\_contrasts <- contrast\_df[contrast\_df$p.value < 0.05, c("contrast", "p.value")]  
  
  
# Formateamos la tabla para imprimir  
formatted\_table <- data.frame(  
 Contrast = significant\_contrasts$contrast,  
 P\_Value = sprintf("%.4f", significant\_contrasts$p.value)  
)  
  
# Imprimimos la tabla  
library(knitr)  
kable(formatted\_table, col.names = c("Contraste", "P-Valor"), format = "pandoc")

| Contraste | P-Valor |
| --- | --- |
| rural espectrometro\_1 - urbano espectrometro\_2 | 0.0002 |
| rural espectrometro\_2 - urbano espectrometro\_2 | 0.0005 |
| urbano espectrometro\_2 - rural espectrometro\_3 | 0.0087 |
| urbano espectrometro\_2 - urbano espectrometro\_3 | 0.0154 |

Eso significa que:

* Hay evidencia significativa para afirmar que la proporción de nitrógeno en el suelo rural es diferente cuando se utiliza el espectrómetro\_1 en comparación con el suelo urbano cuando se utiliza el espectrómetro\_2.
* Hay evidencia significativa para afirmar que la proporción de nitrógeno en el suelo rural es diferente cuando se utiliza el espectrómetro\_2 en comparación con el suelo urbano cuando también se utiliza el espectrómetro\_2.
* Hay evidencia significativa para afirmar que la proporción de nitrógeno en el suelo urbano es diferente cuando se utiliza el espectrómetro\_2 en comparación con el suelo rural cuando se utiliza el espectrómetro\_3.
* Hay evidencia significativa para afirmar que la proporción de nitrógeno en el suelo urbano es diferente cuando se utiliza el espectrómetro\_2 en comparación con el mismo suelo urbano cuando se utiliza el espectrómetro\_3.

Estos valores de p bajos indican que, tras ajustar para múltiples comparaciones, las diferencias observadas en estas combinaciones específicas son estadísticamente significativas. Para las demás combinaciones, no se detectan diferencias significativas.

Es relevante subrayar la importancia de realizar correcciones para mitigar el error de tipo 1, especialmente dado el gran número de comparaciones efectuadas. En este caso, se ha implementado la corrección de Bonferroni.

#### Para el espectrómetro:

# Instalar el paquete emmeans si aún no lo has hecho  
# install.packages("emmeans")  
  
# Cargar el paquete  
library(emmeans)  
  
# Calcular las medias marginales ajustadas (LSMeans)  
lsmeans\_modelo <- emmeans(modelo, ~espectrometro)

## NOTE: Results may be misleading due to involvement in interactions

# Comparaciones múltiples entre niveles de espectrómetros  
comparaciones <- pairs(lsmeans\_modelo)  
  
# Ver los resultados  
summary(comparaciones)

## contrast estimate SE df t.ratio p.value  
## espectrometro\_1 - espectrometro\_2 2.81 0.999 12 2.816 0.0385  
## espectrometro\_1 - espectrometro\_3 1.02 0.999 12 1.022 0.5780  
## espectrometro\_2 - espectrometro\_3 -1.79 0.999 12 -1.794 0.2130  
##   
## Results are averaged over the levels of: suelo   
## Degrees-of-freedom method: kenward-roger   
## P value adjustment: tukey method for comparing a family of 3 estimates

Se ratifica la existencia de una disparidad significativa entre el espectrómetro 2 y el espectrómetro 1, dado que se obtiene un valor de 0.03, lo cual conduce al rechazo de la hipótesis nula (H0) que sostiene que no existen diferencias entre ambos. Por otro lado, se constata la ausencia de discrepancias entre el espectrómetro 1 y el espectrómetro 3, así como entre el espectrómetro 2 y el espectrómetro 3 (valor de p > 0.05). Este resultado ya se reflejaba en el resumen del modelo, aunque la comparación entre el espectrómetro 2 y el espectrómetro 3 no estaba detallada.

# Problema 2

Cargamos los datos del problema.

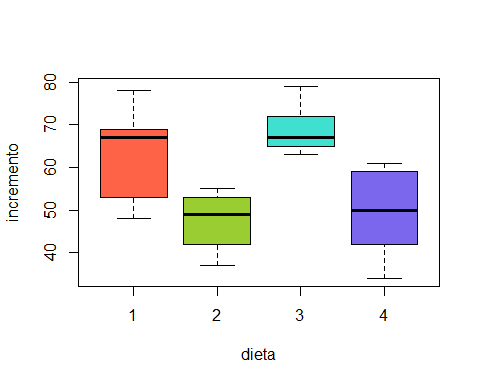
# Seteamos el directorio de trabajo  
setwd("D:/Antiguos estudios/MASTER2/Sem3/Diseño/PAC6/PAC6")  
  
# Importamos los datos  
data <- read.csv("dades\_problema2\_pec6.csv", dec = ",", header = TRUE, sep = ";")  
  
# Marcamos los factores  
data$dieta <- as.factor(data$dieta)

## 2.1 Análisis descriptivo de los datos

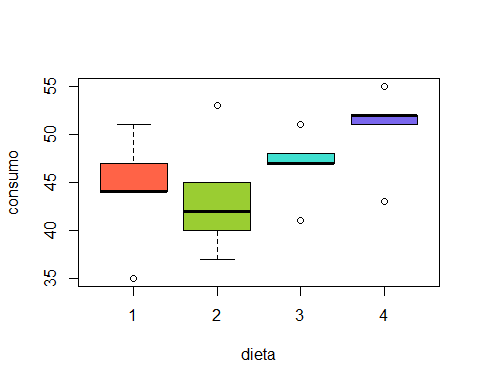
# Realiza un resumen descriptivo  
summary(data)

## dieta consumo incremento   
## 1:5 Min. :35.00 Min. :34.00   
## 2:5 1st Qu.:42.75 1st Qu.:48.75   
## 3:5 Median :47.00 Median :57.00   
## 4:5 Mean :46.25 Mean :57.15   
## 3rd Qu.:51.00 3rd Qu.:67.00   
## Max. :55.00 Max. :79.00

# Gráfico relación consumo-incremento  
color\_dieta = c("tomato1", "olivedrab3", "turquoise", "slateblue2")  
  
# Boxplot incremento por dieta  
boxplot(incremento ~ dieta, data = data, col = color\_dieta)



# Boxplot consumo por dieta  
boxplot(consumo ~ dieta, data = data, col = color\_dieta)

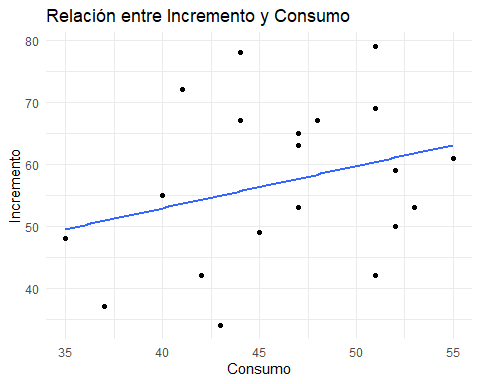


library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.3.2

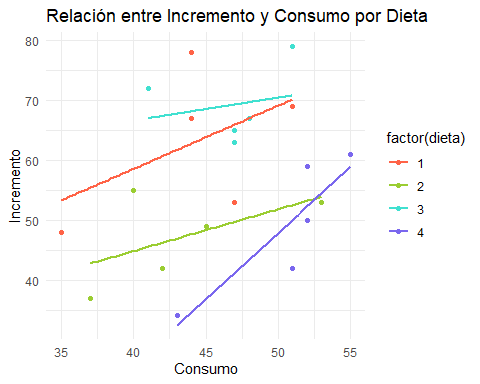
# Gráfico de dispersión Incremento vs Consumo sin dieta  
ggplot(data, aes(x = consumo, y = incremento)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) + # Añade la línea de tendencia general  
 labs(title = "Relación entre Incremento y Consumo",  
 x = "Consumo",  
 y = "Incremento") +  
 theme\_minimal()

## `geom\_smooth()` using formula = 'y ~ x'



# Gráfico de dispersión Incremento vs Consumo por dieta  
ggplot(data, aes(x = consumo, y = incremento, color = factor(dieta))) +  
 scale\_color\_manual(values = color\_dieta) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) + # Añade la línea de tendencia +   
 labs(title = "Relación entre Incremento y Consumo por Dieta",  
 x = "Consumo",  
 y = "Incremento") +  
 theme\_minimal()

## `geom\_smooth()` using formula = 'y ~ x'



Se evidencia una distribución equilibrada con 5 animales por cada tipo de dieta, lo que configura un diseño balanceado.

El análisis del primer boxplot, que representa la relación entre el incremento de peso y la dieta, sugiere variaciones significativas entre los grupos. Esto se infiere al observar diferencias en las alturas de las cajas y la línea central, que representa la media de cada grupo. Además, las distintas alturas y longitudes de las líneas de dispersión indican variabilidades divergentes entre los grupos.

En contraste, al examinar el segundo boxplot que relaciona el consumo de calorías con la dieta, se puede inferir si existen disparidades en el consumo calórico entre los grupos. Efectivamente, se observan diferencias notables entre los grupos.

Por otro lado, el gráfico que ilustra la relación entre el aumento de peso y el consumo calórico revela una correlación positiva general. A mayor consumo de calorías, se evidencia un mayor incremento de peso, y viceversa.

Al analizar esta relación de manera específica para cada dieta, se mantiene la correlación positiva, pero se aprecian pendientes distintas para cada tipo de dieta. Destaca que la dieta 4 exhibe la correlación positiva más pronunciada.

En conjunto, estos hallazgos contribuyen a una comprensión más profunda de las relaciones entre las variables, lo que facilita el ajuste preciso de los modelos de regresión.

## 2.2 Modelo Lineal

### Ajustando según el consumo calórico

El modelo lineal, teniendo en cuenta el consumo calórico, puede especificarse como:

Donde:

* es el aumento de peso del animal bajo la dieta .
* es el intercepto común para todas las dietas.
* es el intercepto específico para la dieta .
* es la pendiente de regresión, que representa el efecto del consumo calórico en el aumento de peso.
* es el consumo calórico del animal bajo la dieta .
* son errores experimentales aleatorios con distribución normal, media 0 y varianza . ().

Suposiciones adicionales:

* El coeficiente de regresión es el mismo para todos los grupos de tratamiento (dietas).
* Las dietas no influyen sobre el consumo calórico (covariante )

El modelo covariante propone utilizar una línea de regresión para cada tratamiento. Esta línea representa cómo la variable de interés cambia en relación con la covariable. Todas las líneas son similares en términos de la dirección de cambio, es decir, comparten la misma “pendiente” (). Esto significa que todos los tratamientos responden de manera similar a la covariable.

Sin embargo, hay diferencias entre los tratamientos, y estas diferencias se reflejan en los términos independientes o “ordenadas al origen” (). Cada tratamiento tiene su propia ordenada al origen (), que se suma a una ordenada al origen común (). Esto significa que aunque las líneas son paralelas, pueden tener alturas diferentes en el eje vertical, indicando que cada tratamiento puede tener un nivel inicial diferente en la variable de interés.

En resumen, el modelo utiliza líneas de regresión paralelas para representar cómo los diferentes tratamientos responden a la covariable, con la flexibilidad de tener diferentes niveles iniciales debido a los términos .

### Sin ajustar por consumo calórico

El modelo lineal, sin tener en cuenta el consumo calórico, puede especificarse como:

Donde:

* es el aumento de peso del animal bajo la dieta .
* es el efecto medio global.
* es el efecto adicional de la dieta en comparación con el promedio global.
* son errores experimentales aleatorios con distribución normal, media 0 y varianza . ().

En este caso, el modelo lineal es mucho más simple porque no se tiene en cuenta la covariable consumo. Es por ello que este modelo simple describe cómo las observaciones () se componen del nivel medio común (), la contribución única de cada tratamiento (), y la variabilidad no explicada ().

## 2.3 Efecto de la dieta

# Ajustando el consumo

# Ajustar el modelo lineal  
modelo\_con <- lm(incremento ~ dieta + consumo, data = data)  
  
summary(modelo\_con)

##   
## Call:  
## lm(formula = incremento ~ dieta + consumo, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.994 -4.639 -2.314 5.756 15.214   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15.7416 18.2280 0.864 0.40141   
## dieta2 -14.9446 5.2558 -2.843 0.01233 \*   
## dieta3 3.4201 5.3499 0.639 0.53228   
## dieta4 -20.6428 5.8478 -3.530 0.00303 \*\*  
## consumo 1.0692 0.4038 2.648 0.01826 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.294 on 15 degrees of freedom  
## Multiple R-squared: 0.6798, Adjusted R-squared: 0.5944   
## F-statistic: 7.96 on 4 and 15 DF, p-value: 0.001192

En el caso de variables dummy que representan categorías (como las dietas), el término constante (o intercepto) es la estimación del valor medio de la variable de respuesta para la categoría de referencia. En este caso, la dieta 1 es la categoría de referencia. Entonces, el término constante representará la estimación del aumento de peso medio para la dieta 1. Las otras variables dummy (dieta 2, dieta 3 y dieta 4) indicarán cómo difieren las otras dietas en comparación con la dieta 1. En resumen, se puede considerar que la dieta 1 es equivalente al intercepto en este contexto específico.

Una vez explicado esto, podemos ver en el resumen del modelo que la estimación para la dieta 1 (intercepto) es de 15.74, y para la dieta 2 es de -14.94. Esto quiere decir que la dieta 1 hace que el peso incremente más que con la dieta 2. No debemos confundir que el estimado de la dieta 2 sea negativo con que disminuya el peso, si no que aumenta pero en menor medida que con la dieta 1, que es con la que se está comparando. Esto concuerda con la exploración inicial de los datos, en la que la dieta 1 mostraba mayor incremento de peso que la dieta 2.

### Sin ajustar el consumo

# Ajustar el modelo lineal  
modelo\_sin <- lm(incremento ~ dieta, data = data)  
  
# Resumen del modelo  
summary(modelo\_sin)

##   
## Call:  
## lm(formula = incremento ~ dieta, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.20 -6.45 1.30 6.45 15.00   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 63.000 4.351 14.480 1.3e-10 \*\*\*  
## dieta2 -15.800 6.153 -2.568 0.0206 \*   
## dieta3 6.200 6.153 1.008 0.3286   
## dieta4 -13.800 6.153 -2.243 0.0394 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.729 on 16 degrees of freedom  
## Multiple R-squared: 0.5301, Adjusted R-squared: 0.4419   
## F-statistic: 6.016 on 3 and 16 DF, p-value: 0.006055

En este caso, el estimado para la dieta 1 es de 63, mientras que para la dieta 2 es de -15.8. Como mencionamos previamente, esto indica que la dieta 1 está asociada con un mayor aumento de peso en comparación con la dieta 2.

Al comparar el estimado de la dieta 1 ajustando el modelo con y sin la covariable, notamos una diferencia sustancial en la estimación (15.74 vs 63). Esta disparidad probablemente se debe a la omisión del efecto del consumo calórico en el modelo sin covariable.

Como hemos observado en la exploración de datos, existe una relación positiva entre la covariable (consumo) y la variable de respuesta. Al no incluir esta covariable en el modelo, parte de la variabilidad debida al consumo se atribuye erróneamente a la variable dieta. Esto se refleja en un estimado más alto para la dieta 1, que sirve como referencia.

Al incorporar el consumo como una covariable en el modelo, estamos intentando aislar el efecto específico de la dieta en el aumento de peso, controlando o ajustando por las diferencias en el consumo entre las dietas. Este enfoque proporciona una estimación más precisa del efecto de la dieta, eliminando el posible sesgo en la relación entre la covariable y la variable de respuesta.

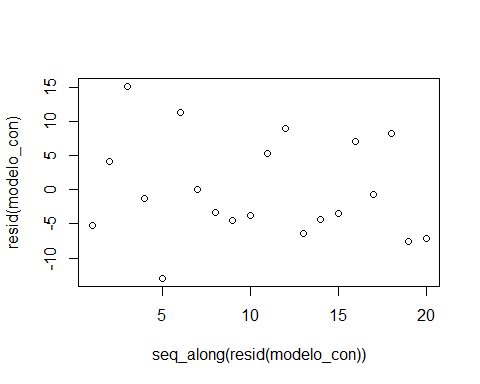
En resumen, al incluir el consumo como covariable, se mejora la capacidad del modelo para discernir entre los efectos de la dieta y el consumo en el aumento de peso, lo que resulta en estimaciones más ajustadas y potencialmente más precisas.

En resumen, al incluir el consumo como covariable, se está mejorando la capacidad del modelo para discernir entre los efectos de la dieta y el consumo en el aumento de peso, y, por lo tanto, se obtienen estimaciones más ajustadas y potencialmente más precisas.

## 2.4 Diagnosis de suposiciones

#### Independencia de los residuos

# Gráfico de residuos vs. orden de observaciones  
plot(resid(modelo\_con) ~ seq\_along(resid(modelo\_con)))



library(lmtest)

## Warning: package 'lmtest' was built under R version 4.3.2

## Loading required package: zoo

##   
## Attaching package: 'zoo'

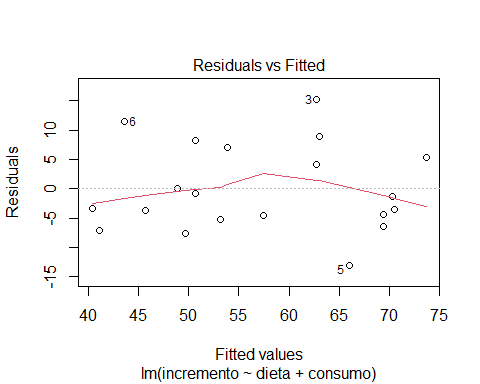
## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

# Prueba de Durbin-Watson para autocorrelación de los residuos  
dwtest(modelo\_con)

##   
## Durbin-Watson test  
##   
## data: modelo\_con  
## DW = 2.1341, p-value = 0.3188  
## alternative hypothesis: true autocorrelation is greater than 0

#### Homocedasticidad

plot(modelo\_con, which=1)



# Prueba de Breusch-Pagan para homocedasticidad  
library(lmtest)  
bptest(modelo\_con)

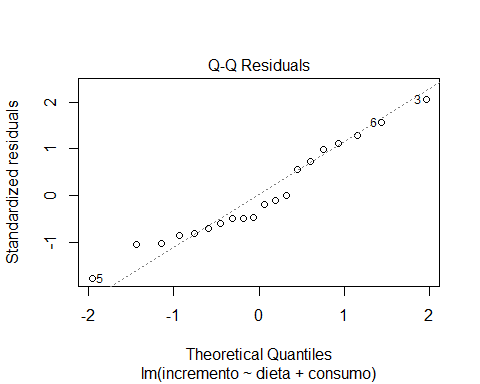
##   
## studentized Breusch-Pagan test  
##   
## data: modelo\_con  
## BP = 2.9601, df = 4, p-value = 0.5645

La dispersión de los residuos es constante a medida que cambian los valores ajustados. Vemos una dispersión uniforme de puntos alrededor de cero, indicando homocedasticidad.

Por otro lado, la prueba de Breusch-Pagan testa la H0 de homocedasticidad. Como obtenemos un pvalor alto (mayor que 0.05) sugiere que no hay suficiente evidencia para rechazar la hipótesis nula, indicando homocedasticidad.

#### Normalidad

plot(modelo\_con, which=2)



shapiro.test(modelo\_con$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: modelo\_con$residuals  
## W = 0.95099, p-value = 0.3824

Tanto a partir del QQ-Plot (puntos alrededor de la diagonal), como el test de Shapiro-Wilks (p-valor > 0,05), concluimos que los residuos son normales.

## 2.5 Diferencias entre dietas

Para determinar entre qué dietas hay diferencias significativas, se deben realizar pruebas de comparaciones múltiples.

# Instalar y cargar el paquete multcomp  
# install.packages("multcomp")  
library(multcomp)

## Loading required package: mvtnorm

## Warning: package 'mvtnorm' was built under R version 4.3.2

## Loading required package: survival

## Loading required package: TH.data

## Loading required package: MASS

##   
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':  
##   
## geyser

# Realizar comparaciones múltiples con el método de Tukey  
comp <- glht(modelo\_con, linfct = mcp(dieta = "Tukey"))  
  
# Resumen de las comparaciones con ajuste de pvalores  
resumen\_comp <- summary(comp)  
pvalores\_ajustados <- p.adjust(resumen\_comp$test$pvalues, method = "bonferroni")  
resumen\_comp$test$pvalues\_ajustados <- pvalores\_ajustados  
  
# Mostrar el resumen con pvalores ajustados  
print(resumen\_comp)

##   
## Simultaneous Tests for General Linear Hypotheses  
##   
## Multiple Comparisons of Means: Tukey Contrasts  
##   
##   
## Fit: lm(formula = incremento ~ dieta + consumo, data = data)  
##   
## Linear Hypotheses:  
## Estimate Std. Error t value Pr(>|t|)   
## 2 - 1 == 0 -14.945 5.256 -2.843 0.05349 .   
## 3 - 1 == 0 3.420 5.350 0.639 0.91727   
## 4 - 1 == 0 -20.643 5.848 -3.530 0.01419 \*   
## 3 - 2 == 0 18.365 5.423 3.387 0.01893 \*   
## 4 - 2 == 0 -5.698 5.998 -0.950 0.77807   
## 4 - 3 == 0 -24.063 5.466 -4.403 0.00249 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## (Adjusted p values reported -- single-step method)

Se han realizado comparaciones múltiples por el método Tukey, y después de ajustar los pvalores por Bonferroni, vemos que existen diferencias significativas entre las dietas 4 y 1, las dietas 3 y 2 y por último las dietas 4 y 3.

# Problema 3

Cargamos los datos del problema.

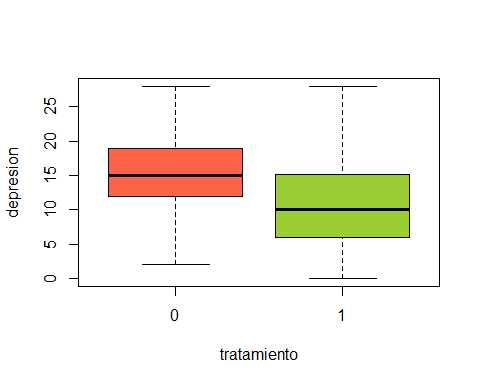
# Seteamos el directorio de trabajo  
setwd("D:/Antiguos estudios/MASTER2/Sem3/Diseño/PAC6/PAC6")  
  
# Importamos los datos  
data <- read.csv("dades\_problema3\_pec6.csv", dec = ".", header = TRUE, sep = ",")  
  
# Marcamos los factores  
data$tratamiento <- as.factor(data$tratamiento)  
data$paciente <- as.factor(data$paciente)  
data$tiempo <- as.factor(data$tiempo)

## 3.1 Análisis descriptivo

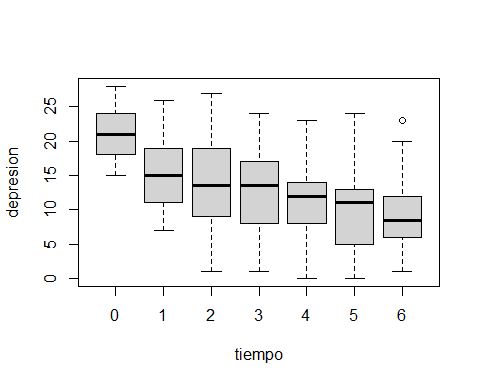
# Realiza un resumen descriptivo  
summary(data)

## tratamiento paciente depresion tiempo  
## 0:175 1 : 7 Min. : 0.00 0:50   
## 1:175 2 : 7 1st Qu.: 8.00 1:50   
## 3 : 7 Median :13.00 2:50   
## 4 : 7 Mean :13.21 3:50   
## 5 : 7 3rd Qu.:18.00 4:50   
## 6 : 7 Max. :28.00 5:50   
## (Other):308 6:50

color\_tratamiento = c("tomato1", "olivedrab3")  
  
# Boxplot depresión por dieta  
boxplot(depresion ~ tratamiento, data = data, col = color\_tratamiento)

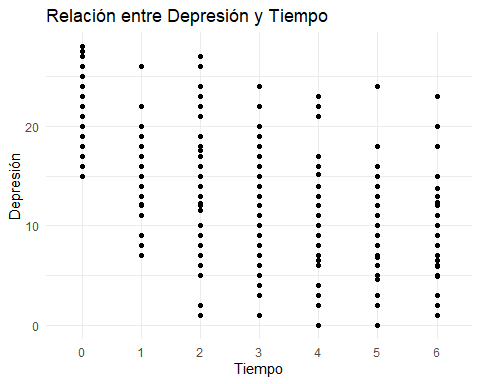


# Boxplot depresión por tiempo  
boxplot(depresion ~ tiempo, data = data)



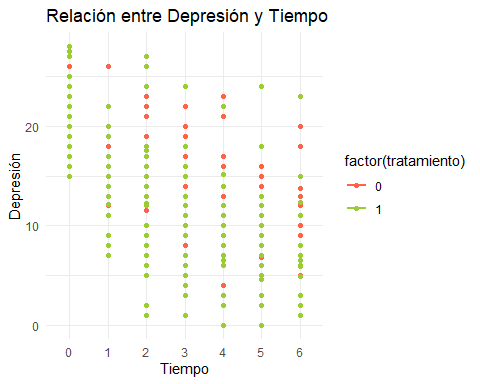
library(ggplot2)  
  
# Gráfico de dispersión depresión vs tiempo  
ggplot(data, aes(x = tiempo, y = depresion)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) + # Añade la línea de tendencia general  
 labs(title = "Relación entre Depresión y Tiempo",  
 x = "Tiempo",  
 y = "Depresión") +  
 theme\_minimal()

## `geom\_smooth()` using formula = 'y ~ x'



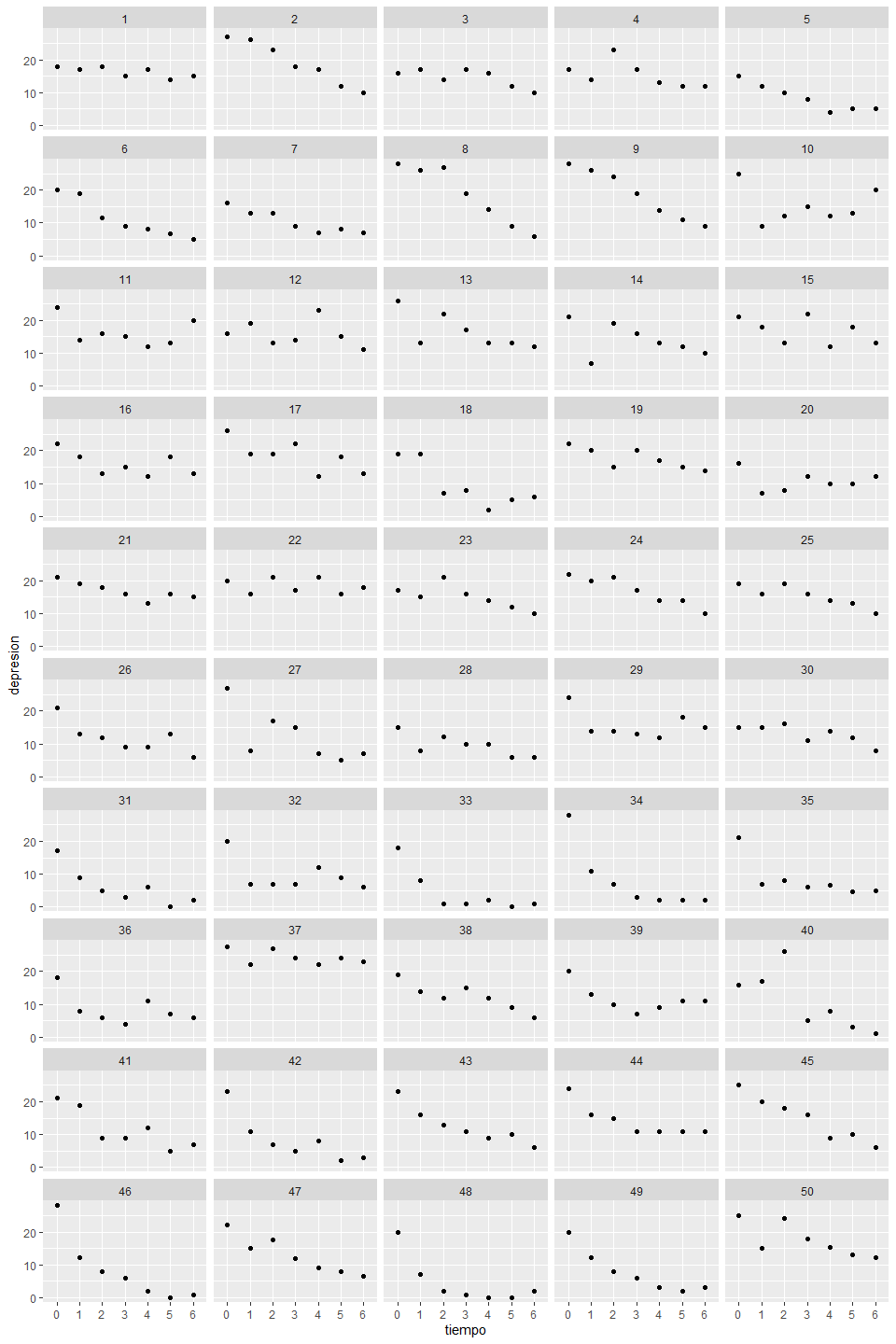
# Gráfico de dispersión depresión vs tiempo con tratamiento  
ggplot(data, aes(x = tiempo, y = depresion, color = factor(tratamiento))) +  
 scale\_color\_manual(values = color\_tratamiento) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) + # Añade la línea de tendencia +   
 labs(title = "Relación entre Depresión y Tiempo",  
 x = "Tiempo",  
 y = "Depresión") +  
 theme\_minimal()

## `geom\_smooth()` using formula = 'y ~ x'



ggplot(data, aes(x = tiempo, y = depresion)) +   
 geom\_smooth(method = "lm", level = 0.95) +  
 geom\_point() +   
 facet\_wrap(~ paciente, nrow = 10, ncol = 5) +  
 theme(  
 )

## `geom\_smooth()` using formula = 'y ~ x'



## 3.2 Modelo lineal

x

Donde:

* es la medida de depresión para la i-ésima paciente en la j-ésima visita.
* es la variable indicadora para el tratamiento (1 si es activo, 0 si es placebo). Es decir, la “i” en no indica el indice del tratamiento si no el número de paciente, el propio “” es indicador de tratamiento.
* es el tiempo de la j-ésima visita para la i-ésima paciente.
* es la intercepción fija del modelo.
* es el efecto fijo del tratamiento.
* es el efecto fijo del tiempo.
* representa la interacción entre el tratamiento y el tiempo.
* es el término aleatorio que modela las diferencias individuales en las intercepciones.
* es el término aleatorio que modela las diferencias individuales en las pendientes respecto al tiempo.
* es el término de error.

Cabe destacar que i=50 porque hay 50 pacientes y j=7 porque se mide la depresión en 7 visitas.

Este modelo puede expresarse en notación matricial de la siguiente manera:

Dimensiones:

* N=50×7=350N=50×7=350: Número total de observaciones.
* k: Número de tratamientos.
* p: Número de efectos fijos no relacionados con el tratamiento.
* q: Número de efectos aleatorios.

Ahora, definimos las matrices y vectores según estas dimensiones:

* : Vector de todas las medidas de depresión, de dimensión 350×1350×1.
* : Matriz de regresión de efectos fijos, de dimensión 350×(k+p)350×(k+p).
* : Vector de coeficientes de efectos fijos, de dimensión (k+p)×1(k+p)×1.
* : Matriz de regresión de efectos aleatorios, de dimensión 350×q350×q.
* : Vector de efectos aleatorios, de dimensión q×1q×1.
* : Vector de errores, de dimensión 350×1350×1.

Particionado para cada elemento, quedaría como:

$Y = $

## 3.3 Implementación del modelo

### Intercepciones aleatorias

Para ajustar este modelo, especificamos que depresión es la variable respuesta y tratamiento y tiempo son variables independientes. Con el término “tratamiento \* tiempo”, también incluimos la interacción entre el tratamiento y el tiempo, ya que podría haber efectos diferenciados del tiempo según el tipo de tratamiento. Por otro lado, “(1|paciente)” especifica que estamos modelando una intercepto aleatorio para cada paciente. Esto tiene en cuenta la variabilidad entre los pacientes.

# install.packages("lme4", dependencies=TRUE)  
# install.packages("Matrix", dependencies=TRUE)  
library(lme4)  
  
# Define el modelo con intercepciones aleatorias  
modelo1 <- lmer(depresion ~ tratamiento \* tiempo + (1|paciente), data = data)  
  
# Muestra un resumen del modelo  
summary(modelo1)

## Linear mixed model fit by REML ['lmerMod']  
## Formula: depresion ~ tratamiento \* tiempo + (1 | paciente)  
## Data: data  
##   
## REML criterion at convergence: 1930.7  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -2.5085 -0.5478 -0.0607 0.5080 4.1131   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## paciente (Intercept) 11.94 3.456   
## Residual 11.90 3.450   
## Number of obs: 350, groups: paciente, 50  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 20.8800 0.9766 21.379  
## tratamiento1 0.6184 1.3812 0.448  
## tiempo1 -4.1200 0.9758 -4.222  
## tiempo2 -4.0584 0.9758 -4.159  
## tiempo3 -5.3200 0.9758 -5.452  
## tiempo4 -7.9224 0.9758 -8.119  
## tiempo5 -8.4472 0.9758 -8.657  
## tiempo6 -9.4488 0.9758 -9.683  
## tratamiento1:tiempo1 -4.6880 1.3800 -3.397  
## tratamiento1:tiempo2 -5.3664 1.3800 -3.889  
## tratamiento1:tiempo3 -7.0584 1.3800 -5.115  
## tratamiento1:tiempo4 -4.7484 1.3800 -3.441  
## tratamiento1:tiempo5 -5.6656 1.3800 -4.106  
## tratamiento1:tiempo6 -5.5396 1.3800 -4.014

##   
## Correlation matrix not shown by default, as p = 14 > 12.  
## Use print(x, correlation=TRUE) or  
## vcov(x) if you need it

### Intercepciones y pendientes aleatorias

Para ajustar este modelo, especificamos que depresión es la variable respuesta y tratamiento y tiempo son variables independientes. Con el término “tratamiento \* tiempo”, también incluimos la interacción entre el tratamiento y el tiempo, ya que podría haber efectos diferenciados del tiempo según el tipo de tratamiento. Por otro lado, “(tiempo | paciente)” especifica que estamos modelando una intercepción aleatoria y una pendiente aleatoria para cada paciente.

El término (tiempo | paciente) implica que cada paciente puede tener su propia intercepción (efecto inicial) y pendiente (efecto en el tiempo) en relación con la variable de respuesta.

remove.packages(“Matrix”) remove.packages(“lme4”) install.packages(“lme4”, type = “source”) library(lme4)

library(lme4)  
  
# Define el modelo con intercepciones y pendientes aleatorias  
modelo2 <- lmer(depresion ~ tratamiento \* tiempo + (tratamiento | paciente), data = data)

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :  
## unable to evaluate scaled gradient

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :  
## Model failed to converge: degenerate Hessian with 1 negative eigenvalues

# TRATAMIENTO O TIEMPO HOLA??  
  
# Muestra un resumen del modelo  
summary(modelo2)

## Linear mixed model fit by REML ['lmerMod']  
## Formula: depresion ~ tratamiento \* tiempo + (tratamiento | paciente)  
## Data: data  
##   
## REML criterion at convergence: 1927.7  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -2.5323 -0.5856 -0.0679 0.5284 4.1159   
##   
## Random effects:  
## Groups Name Variance Std.Dev. Corr  
## paciente (Intercept) 7.262 2.695   
## tratamiento1 2.674 1.635 0.76  
## Residual 11.902 3.450   
## Number of obs: 350, groups: paciente, 50  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 20.8800 0.8755 23.849  
## tratamiento1 0.6184 1.3812 0.448  
## tiempo1 -4.1200 0.9758 -4.222  
## tiempo2 -4.0584 0.9758 -4.159  
## tiempo3 -5.3200 0.9758 -5.452  
## tiempo4 -7.9224 0.9758 -8.119  
## tiempo5 -8.4472 0.9758 -8.657  
## tiempo6 -9.4488 0.9758 -9.683  
## tratamiento1:tiempo1 -4.6880 1.3800 -3.397  
## tratamiento1:tiempo2 -5.3664 1.3800 -3.889  
## tratamiento1:tiempo3 -7.0584 1.3800 -5.115  
## tratamiento1:tiempo4 -4.7484 1.3800 -3.441  
## tratamiento1:tiempo5 -5.6656 1.3800 -4.106  
## tratamiento1:tiempo6 -5.5396 1.3800 -4.014

##   
## Correlation matrix not shown by default, as p = 14 > 12.  
## Use print(x, correlation=TRUE) or  
## vcov(x) if you need it

## optimizer (nloptwrap) convergence code: 0 (OK)  
## unable to evaluate scaled gradient  
## Model failed to converge: degenerate Hessian with 1 negative eigenvalues

### Comparación de modelos

#### Criterios de información

# AIC y BIC  
modelo1\_AIC <- AIC(modelo1)  
modelo1\_BIC <- BIC(modelo1)  
modelo2\_AIC <- AIC(modelo2)  
modelo2\_BIC <- BIC(modelo2)  
  
# Creamos un df  
resultados\_df <- data.frame(  
 Model = c("Modelo con intercepciones aleatorias", "Modelo con intercepciones y pendientes aleatorias"),  
 AIC = c(modelo1\_AIC, modelo2\_AIC),  
 BIC = c(modelo1\_BIC, modelo2\_BIC)  
)  
  
library(knitr)  
kable(resultados\_df, format = "pandoc")

| Model | AIC | BIC |
| --- | --- | --- |
| Modelo con intercepciones aleatorias | 1962.722 | 2024.449 |
| Modelo con intercepciones y pendientes aleatorias | 1963.715 | 2033.157 |

El AIC (Criterio de Información de Akaike) y el BIC (Criterio de Información Bayesiano), son medidas que evalúan la bondad de ajuste de un modelo, teniendo en cuenta su complejidad. En general, modelos con valores más bajos de AIC y BIC son considerados preferibles.

El AIC penaliza la complejidad del modelo menos que el BIC. Cuanto menor sea el valor de AIC, mejor se considera el modelo en términos de ajuste y parsimonia.

El BIC penaliza la complejidad del modelo de manera más fuerte que el AIC. Similar al AIC, un valor más bajo de BIC indica un mejor equilibrio entre ajuste y complejidad.

Según esta información, parece ser que el modelo con intercepciones aleatorias describe bien el modelo sin necesidad de tanta complejidad.

#### Comparación de modelos anidados

# Comparación de modelos usando likelihood ratio test  
anova(modelo1, modelo2)

## refitting model(s) with ML (instead of REML)

## Data: data  
## Models:  
## modelo1: depresion ~ tratamiento \* tiempo + (1 | paciente)  
## modelo2: depresion ~ tratamiento \* tiempo + (tratamiento | paciente)  
## npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)  
## modelo1 16 1981.9 2043.7 -974.97 1949.9   
## modelo2 18 1982.8 2052.2 -973.40 1946.8 3.1327 2 0.2088

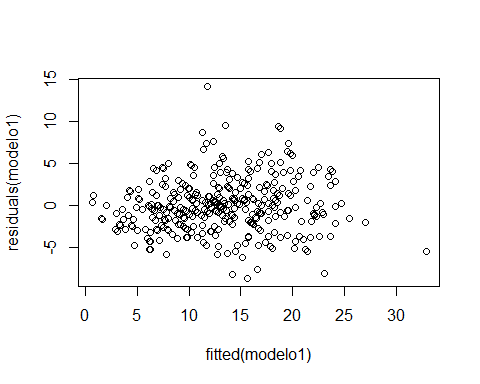
Según la prueba de razón de la verosimilitud, parece ser que no se muestran diferencias significativas entre los dos modelos (pvalor > 0.05). Siendo este el caso, es preferible escoger el modelo más sencillo, ya que explica igual de bien los datos que el modelo más complejo.

Un valor de p menor que un umbral (por ejemplo, 0.05) indicaría que el modelo más complejo es preferible.

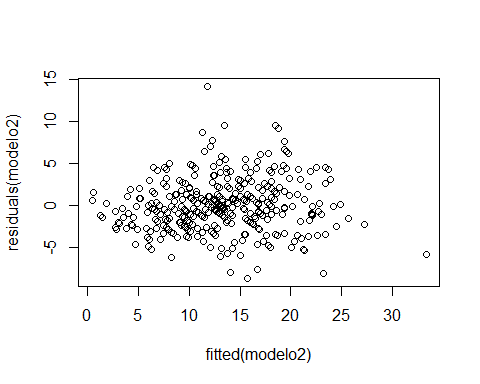
Cabe destacar que aunque los valores de AIC y BIC que nos proporciona la función anova son ligeramente distintos a los calculados a mano por la manera de calcularlos. Pero nos indican el mismo resultado, que es que parece ser que el modelo simple es mejor que el modelo complejo.

#### Diagnóstico de residuos

# Gráfico de residuos para el modelo1  
plot(residuals(modelo1) ~ fitted(modelo1))



# Gráfico de residuos para el modelo1  
plot(residuals(modelo2) ~ fitted(modelo2))



En ninguno de los dos modelos se observan patrones inadecuados de los residuos, así que en ese sentido ambos son correctos.Una vez más, puesto que los dos son correctos el más sencillo es preferible.

Con todo esto concluimos que el modelo con intercepciones aleatorias es el más adecuado, ya que su ajuste es bueno y es más sencillo que el modelo con intercepciones y pendientes aleatorias.

## 3.4 Factores significativos

summary(modelo1)

## Linear mixed model fit by REML ['lmerMod']  
## Formula: depresion ~ tratamiento \* tiempo + (1 | paciente)  
## Data: data  
##   
## REML criterion at convergence: 1930.7  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -2.5085 -0.5478 -0.0607 0.5080 4.1131   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## paciente (Intercept) 11.94 3.456   
## Residual 11.90 3.450   
## Number of obs: 350, groups: paciente, 50  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 20.8800 0.9766 21.379  
## tratamiento1 0.6184 1.3812 0.448  
## tiempo1 -4.1200 0.9758 -4.222  
## tiempo2 -4.0584 0.9758 -4.159  
## tiempo3 -5.3200 0.9758 -5.452  
## tiempo4 -7.9224 0.9758 -8.119  
## tiempo5 -8.4472 0.9758 -8.657  
## tiempo6 -9.4488 0.9758 -9.683  
## tratamiento1:tiempo1 -4.6880 1.3800 -3.397  
## tratamiento1:tiempo2 -5.3664 1.3800 -3.889  
## tratamiento1:tiempo3 -7.0584 1.3800 -5.115  
## tratamiento1:tiempo4 -4.7484 1.3800 -3.441  
## tratamiento1:tiempo5 -5.6656 1.3800 -4.106  
## tratamiento1:tiempo6 -5.5396 1.3800 -4.014

##   
## Correlation matrix not shown by default, as p = 14 > 12.  
## Use print(x, correlation=TRUE) or  
## vcov(x) if you need it

Como podemos ver, en el resumen del modelo no se muestran qué factores resultan significativos, pero podemos calcular los valores de confianzas usando un valor critico de la distribución de la t student. En este caso una valor de confianza de 0.05, implicaría un valor crítico de 1.96.

Intervalo de confianza = Estimado de coeficiente +- 1.96 \* Error estándard del coeficiente.

Para más comodidad, lo calculamos automáticamente con el paquete broom.

# install.packages("broom")  
library(broom)  
  
# Intervalos de confianza  
confint(modelo1)

## Computing profile confidence intervals ...

## 2.5 % 97.5 %  
## .sig01 2.724463 4.277201  
## .sigma 3.126764 3.669822  
## (Intercept) 18.991280 22.768720  
## tratamiento1 -2.052654 3.289454  
## tiempo1 -5.999869 -2.240131  
## tiempo2 -5.938269 -2.178531  
## tiempo3 -7.199869 -3.440131  
## tiempo4 -9.802269 -6.042531  
## tiempo5 -10.327069 -6.567331  
## tiempo6 -11.328669 -7.568931  
## tratamiento1:tiempo1 -7.346536 -2.029464  
## tratamiento1:tiempo2 -8.024936 -2.707864  
## tratamiento1:tiempo3 -9.716936 -4.399864  
## tratamiento1:tiempo4 -7.406936 -2.089864  
## tratamiento1:tiempo5 -8.324136 -3.007064  
## tratamiento1:tiempo6 -8.198136 -2.881064

Los factores significativos son aquellos en los que el intervalo de confianza no incluye el valor 0.

Podemos observar que sólo el intervalo de confianza del tratamiento1 incluye el valor 0. Es decir que el resto de factores (tiempo e interacción tiempo-tratamiento) son significativos, pero no el tratamiento.

Esto no es muy esperanzador para el estudio clínico, ya que parece ser que el tratamiento no tiene efecto sobre la depresión, si no que esta disminuye naturalmente a lo largo del tiempo en las pacientes independientemente del fármaco.

Sin embargo, al tener una interacción significativa parece ser que la relación entre el tratamiento y la respuesta de depresión cambia significativamente a lo largo del tiempo. En otras palabras, el efecto del tratamiento puede ser diferente en diferentes momentos del estudio. Esto podría estudiarse más detalladamente para esclarecerse mejor: Podría ser que el tratamiento no tenga un efecto inmediato, pero a medida que pasa el tiempo, se vuelva significativo. Esto podría sugerir que se necesita tiempo para que el tratamiento tenga un impacto en la respuesta de depresión. O bien que el tratamiento tenga un efecto significativo en algunos momentos específicos (por ejemplo, ciertos períodos de tiempo después de la administración), pero no en otros. O finalmente que el tratamiento afecte de manera diferente a la respuesta de depresión en distintos momentos del estudio, y estas variaciones a lo largo del tiempo se combinan de tal manera que el efecto general del tratamiento no es significativo.

Se podrían realizar más análisis post-hoc y análisis exploratorios para acabar de esclarecer estos efectos.

## 3.5 Modelo estimado en las pacientes 1 y 26

Tal y como hemos explicado antes:

Donde:

* es el intercepto.
* son los coeficientes asociados a las variables predictoras.
* es el efecto aleatorio para cada paciente.

Esta ecuación representa la relación entre la variable de respuesta (depresión) y las variables predictoras (tratamiento, tiempo, y la interacción entre tratamiento y tiempo) considerando tanto los efectos fijos como los efectos aleatorios.

Podemos extraer y estudiar los efectos aleatorios de los pacientes 1 y 26.

library(lme4)  
efectos\_aleatorios <- ranef(modelo1)$paciente  
  
efecto\_paciente1 <- efectos\_aleatorios[1, ]  
efecto\_paciente26 <- efectos\_aleatorios[26, ]  
  
cat("Paciente 1:", efecto\_paciente1)

## Paciente 1: 0.8949959

cat("\n")

cat("Paciente 26:", efecto\_paciente26)

## Paciente 26: 0.6120707

Para el paciente 1: El valor positivo sugiere que su respuesta en depresión es aproximadamente 0.895 unidades más alta de lo esperado basándose en el modelo global. En otras palabras, el paciente 1 experimenta una depresión más elevada de lo que se predice por el modelo para el conjunto de datos en su totalidad.

Para el paciente 26: El valor positivo de 0.612 sugiere que su respuesta en depresión es aproximadamente 0.612 unidades más alta de lo esperado basándose en el modelo global. En este caso, el paciente 26 también experimenta una depresión más elevada de lo predicho por el modelo.

Estos valores proporcionan información sobre la variabilidad individual entre los pacientes en términos de la respuesta a la depresión. Es importante tener en cuenta que estos efectos aleatorios representan la diferencia entre la respuesta real de cada paciente y la respuesta predicha por los efectos fijos del modelo global.

En resumen, un valor positivo indica que el paciente tiene una respuesta más alta de lo esperado, mientras que un valor negativo indicaría una respuesta más baja de lo esperado.

En cuanto a los efectos fijos o coeficientes:

summary(modelo1)

## Linear mixed model fit by REML ['lmerMod']  
## Formula: depresion ~ tratamiento \* tiempo + (1 | paciente)  
## Data: data  
##   
## REML criterion at convergence: 1930.7  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -2.5085 -0.5478 -0.0607 0.5080 4.1131   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## paciente (Intercept) 11.94 3.456   
## Residual 11.90 3.450   
## Number of obs: 350, groups: paciente, 50  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 20.8800 0.9766 21.379  
## tratamiento1 0.6184 1.3812 0.448  
## tiempo1 -4.1200 0.9758 -4.222  
## tiempo2 -4.0584 0.9758 -4.159  
## tiempo3 -5.3200 0.9758 -5.452  
## tiempo4 -7.9224 0.9758 -8.119  
## tiempo5 -8.4472 0.9758 -8.657  
## tiempo6 -9.4488 0.9758 -9.683  
## tratamiento1:tiempo1 -4.6880 1.3800 -3.397  
## tratamiento1:tiempo2 -5.3664 1.3800 -3.889  
## tratamiento1:tiempo3 -7.0584 1.3800 -5.115  
## tratamiento1:tiempo4 -4.7484 1.3800 -3.441  
## tratamiento1:tiempo5 -5.6656 1.3800 -4.106  
## tratamiento1:tiempo6 -5.5396 1.3800 -4.014

##   
## Correlation matrix not shown by default, as p = 14 > 12.  
## Use print(x, correlation=TRUE) or  
## vcov(x) if you need it

Como hemos explicado antes, tanto para el paciente 1 como para el 26, la depresión diminuye conforme el tiempo aumenta.