PAC6

Maria Lucas Gascón

2024-01-17

# Problema 1

Cargamos los datos del problema.

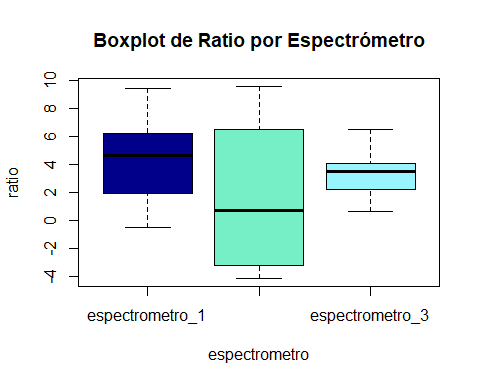
# Seteamos el directorio de trabajo  
setwd("D:/Antiguos estudios/MASTER2/Sem3/Diseño/PAC6/PAC6")  
  
# Importamos los datos  
data <- read.csv("dades\_problema1\_pec6.csv", dec = ",", header = TRUE, sep = ";")  
  
# Marcamos los factores  
data$espectrometro <- as.factor(data$espectrometro)  
data$suelo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

## 1.1 Análisis descriptivo de los datos

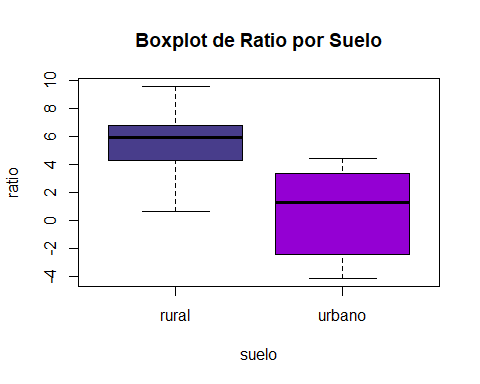
# Primero examinamos la variable ratio  
summary(data$ratio)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -4.1356 0.9882 3.5102 3.1112 5.7919 9.5738

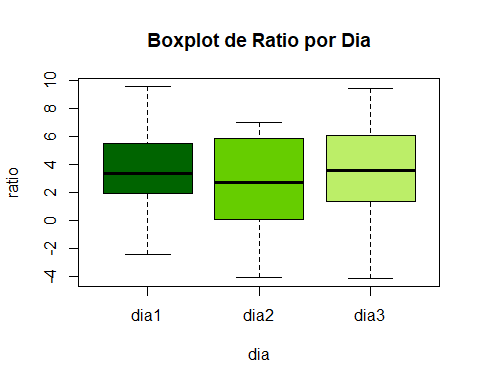
# Box plot de ratio por espectrometro  
boxplot(ratio ~ espectrometro, data = data, col = c("blue4", "aquamarine2", "cadetblue1"), main = "Boxplot de Ratio por Espectrómetro")



# 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")



## 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

# 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

Especificar que queremos la ANOVA de “tipo III” implica que se atribuye la variabilidad a cada término después de tener en cuenta todos los demás términos, incluidas las interacciones que involucran ese término. Es útil cuando se tienen interacciones y se busca evaluar la contribución de cada variable a la variabilidad explicada, considerando las otras variables y sus interacciones.

Además, como no se han comprobado las suposiciones (normalidad y homocedasticidad), es preferible usar la prueba Chi-cuadrado en lugar de F.

Como podemos ver en los resultados, el suelo y la interacción suelo:espectrofotometro son significativas (pv<0.05), aunque el espectrofotómetro por sí mismo no lo es (pv>0.05). Veamos en detalle lo que esto significa:

* Suelo Significativo: Hay diferencias estadísticamente significativas en la proporción de 14N a 15N entre los suelos urbanos y rurales (pv<0.05, se rechaza H0), después de tener en cuenta los otros efectos en el modelo. En otras palabras, hay evidencia estadística para afirmar que el tipo de suelo tiene un impacto significativo en la proporción de nitrógeno.
* Espectrómetro No Significativo: No hay evidencia estadística suficiente para afirmar que hay diferencias en la proporción de 14N a 15N entre los distintos espectrómetros (pv>0.05, se acepta H0), después de considerar los otros efectos. En resumen, las variaciones observadas en la proporción de nitrógeno no pueden atribuirse de manera significativa a las diferencias entre los espectrómetros utilizados.
* Interacción Suelo-Espectrómetro Significativa: La influencia del espectrómetro en la proporción de nitródeno depende del tipo de suelo. En otras palabras, la diferencia en la proporción de 14N a 15N entre los espectrómetros puede variar según si el suelo es urbano o rural. (pv<0.05, se rechaza H0). Esta interacción indica que el efecto del espectrómetro 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 espectrómetro son factores significativos en la variabilidad de la proporción de 14N a 15N, mientras que el espectrómetro por sí mismo no tiene un efecto significativo en esta proporción.

## 1.5 Comparaciones múltiples

Cómo el factor espectrómetro no mostraba diferencias significativas y el factor suelo sólo tiene 2 niveles; la única comparación múltiple que tiene sentido realizar es la de la interación.

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

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

library(knitr)  
  
# 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  
kable(formatted\_table, col.names = c("Contraste", "P-Valor"))

| 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 p-valores bajos indican que, después de ajustar para múltiples comparaciones, las diferencias observadas en estas combinaciones específicas son estadísticamente significativas. Para el resto de combinaciones no se observan diferencias significativas.

Cabe destacar la importancia de realizar una corrección para minimizar el error de tipo 1, ya que se estan realizando un gran número de comparaciones. En este caso se ha usado la corrección de Bonferroni.

# Problema 2

Cargamos los datos del problema.

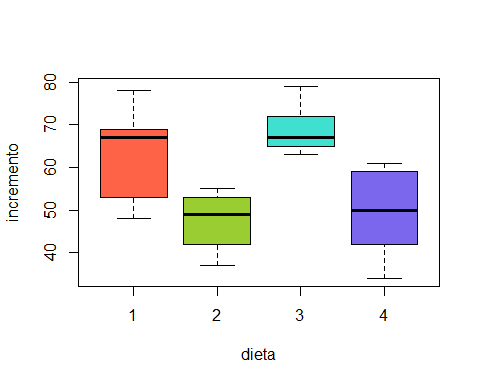
# Seteamos el directorio de trabajo  
setwd("D:/Antiguos estudios/MASTER2/Sem3/Diseño/PAC6/PAC6")  
  
# Importamos los datos  
data <- read.csv("dades\_problema2\_pec6.csv", dec = ",", header = TRUE, sep = ";")  
  
# Marcamos los factores  
data$dieta <- as.factor(data$dieta)

## 2.1 Análisis descriptivo de los datos

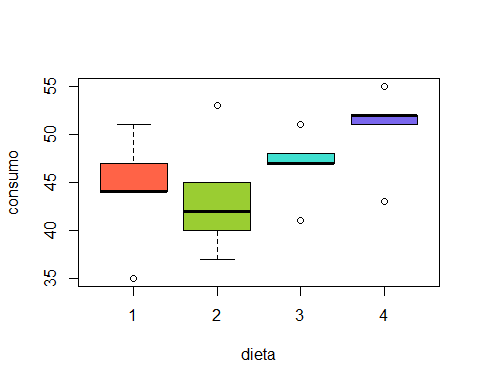
# Realiza un resumen descriptivo  
summary(data)

## 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)

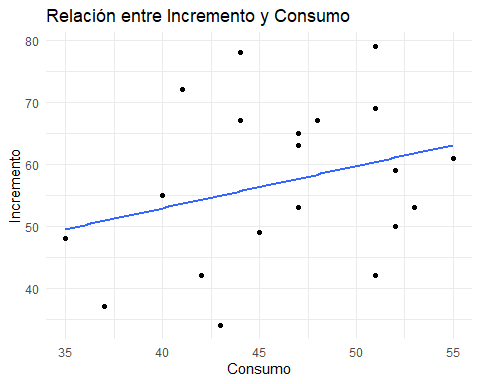


library(ggplot2)

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

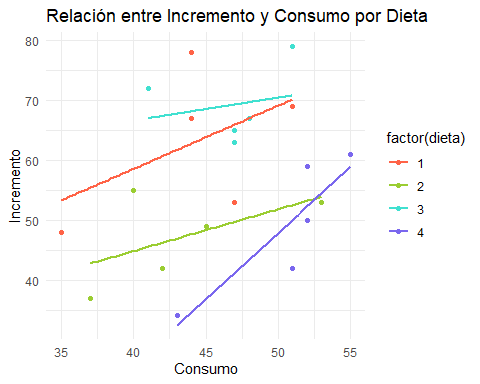
# Gráfico de dispersión Incremento vs Consumo sin dieta  
ggplot(data, aes(x = consumo, y = incremento)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE) + # Añade la línea de tendencia general  
 labs(title = "Relación entre Incremento y Consumo",  
 x = "Consumo",  
 y = "Incremento")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

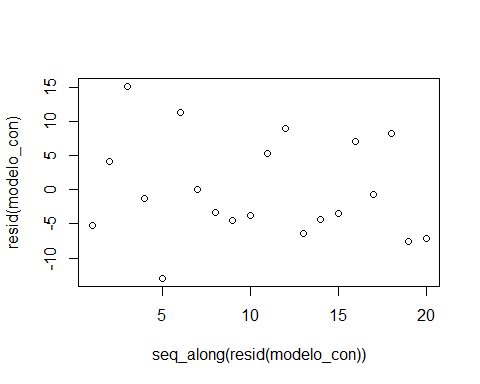
En resumen, al incluir el consumo como covariable, se mejora la capacidad del modelo para discernir entre los efectos de la dieta y el consumo en el aumento de peso, lo que resulta en estimaciones más ajustadas y potencialmente más precisas.

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'

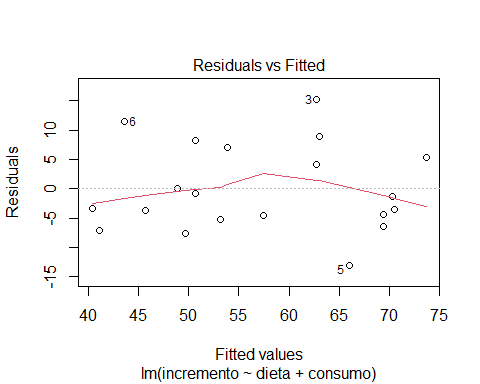
## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

# 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)

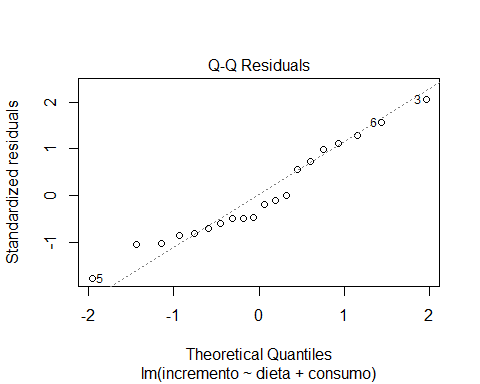
##   
## studentized Breusch-Pagan test  
##   
## data: modelo\_con  
## BP = 2.9601, df = 4, p-value = 0.5645

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

## 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 0.05343 .   
## 3 - 1 == 0 3.420 5.350 0.639 0.91725   
## 4 - 1 == 0 -20.643 5.848 -3.530 0.01424 \*   
## 3 - 2 == 0 18.365 5.423 3.387 0.01900 \*   
## 4 - 2 == 0 -5.698 5.998 -0.950 0.77811   
## 4 - 3 == 0 -24.063 5.466 -4.403 0.00263 \*\*  
## ---  
## 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.