Modelos ANOVA

Iván Gracia Baquedano

Trabajo de Fin de Grado de Matemáticas
Universidad de Zaragoza

Directora del trabajo: Ana Pérez Palomares
4 de septiembre de 2023

Abstract

The objective of this work is to provide an overview of the different ANOVA models that exist and the circumstances in which they are used. The research is structured into a brief introduction and five chapters.

The first section serves as the bedrock, introducing fundamental concepts of linear models. Topics cover the model's structure, assumptions, and estimation techniques, facilitating a clear understanding of the subsequent models.

Advancing, the second section delves into one-factor ANOVA models, an essential tool for comparing means between groups in statistical analysis. We will explore how to design and structure experiments, identifying the factor of interest.

The third section expands the analysis to encompass two-factor ANOVA models, which allow the exploration of interactions between two independent variables. This segment explores the complexities of such models, including main effects, interactions, and interpretation of results. We will also provide some indications of how these models can be extended to more factors.

In the fourth section, the study delves into the domain of random-effects ANOVA models, explaining how the incorporation of random factors can contribute to the total variability of the dataset. We will discuss different models that use these random factors, such as the mixed or nested models.

The fifth and final section presents a study of real data analysis, applying the results seen in the previous chapters and drawing conclusions about the studies.

Throughout the study, an initial exploration of these modeling techniques is presented, covering theoretical foundations and practical applications.

Índice general

Introducción

El Análisis de Varianza, comúnmente conocido como ANOVA, fue desarrollado por Fisher en 1930 y constituye una potente técnica estadística empleada para investigar y comparar las diferencias entre las medias de dos o más grupos dentro de un conjunto de datos. Su aplicación abarca una amplia gama de disciplinas. Por ejemplo, en el ámbito médico, resulta de gran interés estudiar la respuesta de una variable clínica en diferentes tratamientos asignados a una serie de pacientes. Sin embargo, su utilidad se extiende más allá de la medicina y se aplica en otras áreas de las ciencias, como en estudios de mercado en economía, en investigación cuantitativa en las ciencias sociales, en química y muchas más. El ANOVA se fundamenta en principios estadísticos sólidos y proporciona una estructura sistemática para evaluar si las diferencias observadas entre grupos son estadísticamente significativas o simplemente resultado del azar. El modelo ANOVA más sencillo compara el comportamiento de una variable (continua) entre dos grupos de individuos, es decir, se compara las medias para dos muestras independientes. Este caso tan sencillo, que se estudia en todos los cursos básicos de Estadística, puede extenderse a multitud de modelos que recogen distintas maneras de realizar los diseños experimentales. La extensión inmediata es considerar más de dos grupos, pero podemos plantearnos añadir otra agrupación a los individuos y que esta pueda tener algún tipo de interacción con la considerada inicialmente. En el ejemplo de los tratamientos a una serie de pacientes, se debería añadir la variable género y plantear si hay diferencias en la respuesta entre hombres y mujeres o incluso si las posibles diferencias entre el tratamiento aplicado no es el mismo para hombres que para mujeres. Las extensiones de los modelos no sólo las encontramos en la manera de incluir los factores que pueden influir en la variable de interés sino también en las hipótesis que se deben cumplir para aplicar una técnica estadística determinada; por ello la bibliografía existente en este tema es muy extensa.

La manera más frecuente de abordar el estudio estadístico de modelos ANOVA es tratarlos como una subclase específica dentro de la familia más general de modelos lineales. Los modelos lineales, en esencia, buscan establecer una relación entre una o más variables independientes (predictoras) y una variable dependiente (respuesta) a través de una función lineal. En este sentido, el ANOVA se convierte en una forma particular de modelo lineal que se enfoca en la comparación de medias entre grupos. En el ejemplo mencionado de los pacientes con distintos tratamientos, podemos plantearnos si la edad o distintos características antropométricas pueden afectar también a la variable respuesta.

A medida que el análisis estadístico ha evolucionado, se han desarrollado variantes más complejas de los modelos lineales, dando lugar a los modelos de regresión generalizados que permiten el tratamiento de variables aleatorias como predictoras o relaciones no lineales con la variable de interés. Estos modelos ofrecen mayor flexibilidad y capacidad para abordar una variedad de relaciones entre variables. A su vez, los modelos ANOVA han evolucionado para manejar diseños más complejos, como los modelos ANOVA de factores aleatorios y ANOVA mixtos, donde los factores pueden ser tanto fijos como aleatorios o los modelos ANCOVA en el que se incluyen además variables cuantitativas como variables predictoras. En resumen, los modelos ANOVA y los modelos lineales están estrechamente relacionados, siendo los modelos ANOVA una aplicación específica de los principios generales de los modelos lineales. Esta relación demuestra cómo las técnicas estadísticas evolucionan y se adaptan para abordar diferentes tipos de preguntas de investigación y diseños experimentales, proporcionando a los analistas una caja de herramientas versátil y robusta para el análisis de datos en una variedad de contextos.

El objetivo del presente trabajo es mostrar los modelos ANOVA básicos y dar una visión general de modelos más complejos. El enfoque se ha basado en la teoría de modelos lineales (asignatura optativa en el grado de Matemáticas), por lo que en el Capítulo 1 se ha incluido un resumen de los tópicos más importantes. Los Capítulos 2 y 3 tratan modelos con un factor y dos factores fijos, respectivamente y en el Capítulo 4 se incluye los modelos con un único factor aleatorio y se muestra, sin abordarlos exhaustivamente, alguna extensión de este tipo de modelos que tienen una gran aplicación en el análisis de medidas repetidas. Finalmente, el Capítulo 5 incluye el análisis de datos reales que muestra un ejemplo práctico de lo previamente estudiado.

Capítulo 1

MODELO LINEAL

1.1. Planteamiento del modelo

Daremos inicio explorando los modelos lineales, un tema que se encuentra abundantemente cubierto en los libros especializados en este campo. Un ejemplo de ello es: [1, Capítulo 1] o [4, Capítulo 2].
Sea Y una variable aleatoria que fluctúa alrededor de un valor desconocido η, esto es $Y = \eta + \varepsilon$ donde ε es el error. Supongamos que η toma valores distintos de acuerdo con diferentes situaciones experimentales según el modelo lineal $\eta = \beta_1 x_1 + \cdots + \beta_m x_m$, donde β_i son parámetros desconocidos y x_i son valores conocidos, cada uno de los cuales ilustra situaciones experimentales diferentes. En general se tienen n observaciones de la variable Y. Diremos que $y_1,\ y_2, \cdots,\ y_n$, siguen un modelo lineal si

$$y_i = x_{i1}\beta_1 + \cdots + x_{im}\beta_m + \varepsilon_i \qquad\qquad i = 1,\ 2, \cdots,\ n,$$

donde $\beta_1,\ \beta_2, \cdots,\ \beta_m$, son los parámetros; los errores ε_i son desviaciones que se comportan como variables aleatorias independientes e idénticamente distribuidas siguiendo una distribución normal de media cero y varianza constante σ^2, es decir, $N(0, \sigma^2)$. La hipótesis de varianza constante se denomina homocedasticidad. La expresión del modelo lineal en su forma matricial es

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \rightarrow \mathbf{Y} = \mathbf{X}\beta + \varepsilon.$$

La matriz del modelo $\mathbf{X}$, también llamada matriz de diseño, tiene dimensiones $n \times m$ $(n > m)$, donde n es el número de observaciones y m es el número de variables predictoras en el modelo. La primera columna suele contener unos para representar el término independiente (β_1). Las otras columnas contienen los valores de las variables predictoras. Otra hipótesis habitual es que el rango de la matriz de diseño es m, es decir, de rango máximo ya que coincide con el número de parámetros. En este contexto, se dice que el modelo está ajustado, lo que implica que no existe colinealidad entre las variables predictoras. En otras palabras, cada variable predictora contribuye de manera única y no redundante al modelo.
Por tanto, el vector de observaciones $\mathbf{Y}$ sigue una distribución normal n-variante de media $\mathbf{X}\beta$ y matriz de varianzas covarianzas $\sigma^2\mathbf{I}_n$, es decir, $N_n(\mathbf{X}\beta, \sigma^2\mathbf{I}_n)$.

Ejemplo 1.1. El modelo lineal más simple consiste en relacionar una variable aleatoria Y con una variable controlable x no aleatoria, de modo que las observaciones de Y verifiquen

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad\qquad i = 1,\ 2, \cdots,\ n,$$

Se dice que Y es la variable a explicar o dependiente y x es la variable explicativa, por ejemplo Y es la respuesta de un fármaco a una dosis x.

1.2. Estimación de los parámetros

En esta sección, vamos a abordar inicialmente la definición de los estimadores de los parámetros. Posteriormente, exploraremos teoremas relacionados con sus propiedades, los cuales se encuentran expuestos en mayor detalle en el [1, Capítulo 2] y en [4, Capítulo 2], junto con demostraciones más exhaustivas.
La estimación de los parámetros $\beta = (\beta_1,\ \beta_2, \cdots,\ \beta_m)'$ se hace con el criterio de los mínimos cuadrados. Se trata de hallar el conjunto de valores de los parámetros $\hat{\beta} = (\hat{\beta}_1,\ \hat{\beta}_2, \cdots,\ \hat{\beta}_m)'$ que minimicen la siguiente suma de cuadrados.

$$\varepsilon'\varepsilon = (\mathbf{Y}-\mathbf{X}\beta)'(\mathbf{Y}-\mathbf{X}\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - \cdots - x_{im}\beta_m)^2.$$

La estimación es $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Si la matriz $(\mathbf{X}'\mathbf{X})^{-1}$ es no singular, existen varios métodos para llevar a cabo esta estimación. Uno de estos métodos se presenta en el anexo 6.1, y para obtener información sobre otros métodos, véase [4, Capítulos 3 y 4].
Es sencillo ver que por las hipótesis del modelo que $\hat{\beta}$ es insesgado para β y que tiene como matriz de covarianzas $\text{var}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Una vez estimado β, falta estimar la varianza del modelo σ^2. Para ello definimos la variación no explicada, a través de los residuos del modelo, $\text{e}= (\mathbf{Y}-\mathbf{X}\hat{\beta})$, como $\text{VNE} = \mathbf{e}'\mathbf{e} = (\mathbf{Y}-\mathbf{X}\hat{\beta})'(\mathbf{Y}-\mathbf{X}\hat{\beta})$. El estimador de la varianza σ^2 es $\hat{\sigma}^2 = VNE/(n-m)$ que es un estimador insesgado de la varianza. Las propiedades de los estimadores y los estadísticos que aparecen en los contrastes de un modelo lineal, se deducen de las propiedades de combinaciones lineales y de formas cuadráticas de vectores normales. En el siguinte lema resumiremos las más importantes.

Lema 1.1. *Sea* $\mathbf{Y} \sim N_n(\Theta,\Sigma)$*, con* Θ *un n-vector de medias y* Σ *una matriz de dimensión* $n \times n$ *definida positiva. Consideremos* $\mathbf{A_{p\times n}}$ *de rango* $p < n$*, entonces*

1. $\mathbf{AY} \sim N_p(\mathbf{A}\Theta, \mathbf{A}\Sigma\mathbf{A}')$.
2. $(\mathbf{Y}-\Theta)'\Sigma^{-1}(\mathbf{Y}-\Theta)$ *sigue una distribución chi-cuadrado con n grados de libertad, es decir,* χ^2_n
3. *Si* $\mathbf{A}$ *es idempotente* $(\mathbf{Y}-\Theta)'\mathbf{A}\Sigma^{-1}\mathbf{A}'(\mathbf{Y}-\Theta) \sim \chi^2_p$.
4. $E(\mathbf{Y}'\mathbf{AY}) = tr(\mathbf{A}\Sigma) + \Theta'\mathbf{A}\Theta$*, donde tr denota la traza de una matriz.*
5. *Sea* $\mathbf{B}$ *una matriz* $k \times n$ *tal que* $\mathbf{AB}' = 0$*, entonces* $\mathbf{AY}$ *y* $\mathbf{BY}$ *son variables aleatorias independientes.*

Demostración: Véase anexo 6.1.
Una breve observación sobre el apartado 3 es que este es un resultado general en relación a formas cuadráticas y no requiere que la variable $\mathbf{Y}$ siga una distribución normal.
A continuación vamos a mostrar las propiedades más importantes de los estimadores.

Teorema 1.1. *Algunas propiedades de los estimadores*

1. $\hat{\beta} \sim N_m(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$.
2. $\hat{\sigma}^2$ *es insesgado para* σ^2.
3. $(n-m)\hat{\sigma}^2/\sigma^2 = VNE/\sigma^2 \sim \chi^2_{n-m}$.
4. $\hat{\beta}$ *y* $\hat{\sigma}^2$ *son independientes.*
5. $(\hat{\beta}-\beta)'\mathbf{X}'\mathbf{X}(\hat{\beta}-\beta)/\sigma^2 \sim \chi^2_m$.
6. $(\hat{\beta}_i - \beta_i)/\hat{\sigma}\sqrt{d_{ii}}$ *sigue una distribución t de Student con n-m grados de libertad, que denotaremos* t_{n-m}*. Definimos como* d_{ii} *a los elementos de la diagonal* $(\mathbf{X}'\mathbf{X})^{-1}$.
7. *El intervalo de confianza de* β_i *a nivel* $100(1-\alpha)\,\%$ *es*

$$\hat{\beta}_i \pm t_{n-m,1-\alpha/2}\hat{\sigma}\sqrt{d_{ii}},$$

donde $t_{n-m;1-\alpha/2}$ *denota el cuantil* $1-\alpha/2$ *de la distribución t.*

Demostración. Durante la demostración, vamos a utilizar la proyección ortogonal $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Para ver las propiedades de esta matriz consultar el anexo 6.2.

1. Se deduce de la expresión de $\hat{\beta}$ y del apartado 1 del Lema 1.1.

2. $VNE = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}) = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P})'(\mathbf{I}_n - \mathbf{P})\mathbf{Y} = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y}$. Aplicando el apartado 4 del lema 1.1 obtenemos que $E(VNE) = E(\mathbf{Y}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y}) = \sigma^2 tr(\mathbf{I}_n - \mathbf{P}) + (\mathbf{X}\beta)'(\mathbf{I}_n - \mathbf{P})(\mathbf{X}\beta)$. Es inmediato comprobar que $(\mathbf{I}_n - \mathbf{P})(\mathbf{X}\beta) = 0$, Además, debido a las propiedades de la matriz $\mathbf{P}$, es posible afirmar que $tr(\mathbf{I}_n - \mathbf{P}) = n - m$, de lo que se deduce que $E(VNE) = \sigma^2(n-m)$ y por tanto el apartado 2 del presente teorema.

3. Dado que $(\mathbf{I}_n - \mathbf{P})(\mathbf{X}\beta) = 0$, resulta posible expresar $VNE = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{I}_n - \mathbf{P})(\mathbf{Y} - \mathbf{X}\beta)$. Podemos aplicar el apartado 3 del Lema 1.1. al vector $(\mathbf{Y} - \mathbf{X}\beta)$, cuya matriz de varianzas-covarianzas es $\sigma^2\mathbf{I}_n$ y se obtiene el resultado.

4. Es inmediato, por el apartado 5 del Lema 1.1. ya que $\mathbf{X}\hat{\beta} = \mathbf{PY}$ y $\mathbf{Y} - \mathbf{X}\hat{\beta} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y}$ son independientes debido que $\mathbf{P}$ es idempotente. Por tanto $\mathbf{X}\hat{\beta}$ y VNE lo son y de aquí se obtiene el resultado.

5. Se sigue del apartado 2 del lema 1.1 ya que, $(\hat{\beta} - \beta)'\mathbf{X}'\mathbf{X}(\hat{\beta} - \beta)/\sigma^2 = (\hat{\beta} - \beta)'\text{var}(\hat{\beta})^{-1}(\hat{\beta} - \beta) \sim \chi^2_m$.

6. Dado que $(\hat{\beta}i - \beta_i)/(\sigma\sqrt{dii}) \sim N(0,1)$ y $\hat{\sigma}^2$ es independiente de $\hat{\beta}$ según el apartado 4 del teorema actual, podemos utilizar la definición de la distribución t de Student (véase anexo 6.1) para obtener el resultado deseado.

7. Inmediato por 6.

□

A partir de este momento denotaremos el cuantil α de una distribución K como K_α.

1.3. Descomposición de la varianza

Un elemento importante en el estudio de un modelo lineal es la descomposición de la variabilidad total, que se define como $VT = \sum_{i=1}^n (y_i - \bar{y})^2$, donde $\bar{y}$ representa la media muestral de $\mathbf{Y}$. La VT representa la cantidad total de la variabilidad en los datos y se divide en explicada y no explicada (residual), como vamos a ver a continuación.
Definimos $\hat{y}_i = x_i\hat{\beta}$, es decir, $\hat{y}_i$ es el valor de la función de ajuste para un valor prefijado de las variables explicativas.

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^n [(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})].$$

Puesto que $(\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\hat{\beta}) = 0$, tenemos que $\bar{y}\sum_{i=1}^n (y_i - \hat{y}_i) = 0$; $\sum_{i=1}^n (\hat{y}_i)(y_i - \hat{y}_i) = 0$.

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

$VE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ es la parte de la variabilidad que explica el modelo.
$VNE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ es la parte que no conseguimos explicar mediante el modelo y que ya se había definido para estimar la varianza del modelo.

1.4. Contrastes de hipótesis

En esta sección, nos adentraremos en la comparación de las hipótesis lineales más reconocidas, las cuales están disponibles en [1, Capítulo 3,4] y en [4, Capítulo 5]
Las hipótesis más habituales para contrastar en un modelo lineal son combinaciones lineales de los parámetros, es decir, del tipo $\mathbf{A}\beta = c$ donde $\mathbf{A_{p\times m}}$ es una matriz de rango $p < m$ y c es un vector de p parámetros. Como casos particulares, tenemos el contraste global de regresión que contrasta si las variables explicativas tienen o no un efecto significativo sobre la variable resultado. Otro caso particular es si cada coeficiente β_i es nulo, que sirve para evaluar si un coeficiente de regresión específico en un modelo lineal es significativamente diferente de cero. Para introducir el contraste general damos primero un lema de interés.

Lema 1.2. *Bajo la restricción H_0 : $\mathbf{A}\beta = c$ el estimador mínimo cuadrático de β es*

$$\hat{\beta}_H = \hat{\beta} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'](\mathbf{A}\hat{\beta} - c).$$

Demostración: Véase anexo 6.1.
Ahora estamos en condiciones de dar el contraste general y algún resultado asociado.

Teorema 1.2. *Sea $\mathbf{A_{p\times m}}$ una matriz de rango $p < m$ y c un vector de p componentes. Sea $\hat{\beta}_H$ la estimación de β imponiendo la restricción $H_0 : \mathbf{A}\beta = c$ y sea VNE_H la variabilidad no explicada en el modelo restringido con $\mathbf{A}\beta = c$. Entonces*

1. $VNE_H - VNE = (\mathbf{A}\hat{\beta} - c)'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\beta} - c)$.
2. $E(VNE_H - VNE) = p\sigma^2 + (\mathbf{A}\beta - c)'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\beta - c)$ *y bajo la restricción H_0 tenemos que $E((VNE_H - VNE)/p)$ es un estimador insesgado para σ^2.*
3. *Bajo la restricción H_0, tenemos que $(VNE_H - VNE)/\sigma^2 \sim \chi_p$.*
4. *$VNE_H - VNE$ y VNE son independientes.*
5. *Bajo la restricción H_0, tenemos que $F = \dfrac{(VNE_H - VNE))/p}{VNE/(n-m)}$ sigue una distribución F de de Snedecor con p y n-m grados de libertad, es decir, $F_{p,n-m}$.*

Por tanto, para el contraste H_0 : $\mathbf{A}\beta = c$ frente a H_1 : $\mathbf{A}\beta \neq c$, utilizaremos el estadístico F.

Demostración. 1.2

1. Vamos a demostrarlo para $c = 0$.

   $VNE = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P})\mathbf{Y} = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\beta}$, $VNE_H = (\mathbf{Y} - \mathbf{X}\hat{\beta}_H)'(\mathbf{Y} - \mathbf{X}\hat{\beta}_H) = \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\hat{\beta}_H + \hat{\beta}_H'\mathbf{X}'\mathbf{X}\hat{\beta}_H$.
   Considerando la igualdad $\mathbf{X}'\mathbf{X}\hat{\beta}_H = \mathbf{X}'\mathbf{Y} - (1/2)\mathbf{A}'\lambda$ que se presenta en la demostración del lema 1.2 y aplicando la restricción $\mathbf{A}\hat{\beta}_H = 0$, obtenemos la siguiente expresión para VNE_H, $VNE_H = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\beta}_H$. Por último, al emplear la igualdad del lema 1.2, concluimos que $VNE_H - VNE = (\hat{\beta}' - \hat{\beta}_H')\mathbf{X}'\mathbf{Y} = (\mathbf{A}\hat{\beta})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\mathbf{A}\hat{\beta})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\beta})$.

2. Del apartado anterior deducimos que la diferencia de las variabilidades es una forma cuadrática del vector $(\mathbf{A}\hat{\beta} - c)$ con $E(\mathbf{A}\hat{\beta} - c) = \mathbf{A}\beta - c$; $Var(\mathbf{A}\hat{\beta} - c) = \mathbf{A}\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'$. Así, aplicando el apartado 4 del lema 1.1, se deduce que $E(VNE_H - VNE) = (\mathbf{A}\beta - c)'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\beta - c) + tr([\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}\mathbf{A}\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}') = p\sigma^2 + (\mathbf{A}\beta - c)'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\beta - c)$.

3. Por el apartado 1 de este teorema $(VNE_H - VNE)/\sigma^2 = (\mathbf{A}\hat{\beta} - c)'[\mathbf{A}\text{var}(\hat{\beta})\mathbf{A}']^{-1}(\mathbf{A}\hat{\beta} - c)$, notar que $\mathbf{A}\hat{\beta} \sim N_q(c, \mathbf{A}\text{var}(\hat{\beta})\mathbf{A}')$ de donde se sigiue $VNE_H - VNE/\sigma^2 \sim \chi^2_p$ por el apartado 2 del lema 1.1.

4. Es inmediato porque $\hat{\beta}$ y VNE son independientes, tal como se demostró en el apartado 4 del teorema 1.1.

5. Se trata de una consecuencia inmediata de los apartados previos, una vez que hemos adquirido conocimiento sobre la definición de la distribución F, tal como se explica en detalle en el apéndice 6.1.

□

1.4.1. Contraste global de regresión

El contraste global nos va a permitir determinar si el modelo tiene algún poder predictivo sobre la variable respuesta, es decir, si hay evidencia estadística que apoye el uso del modelo lineal. Como se ha dicho, habitualmente el modelo tiene un término constante, β_1, y el resto de los parámetros están asociados a las variables explicativas del modelo. Por tanto, denotando $\beta = (\beta_1, \beta*)'$ la hipótesis para contrastar la existencia de relación lineal entre las variables predictoras y la respuesta es:

$$H_0 : \beta* = 0$$
$$H_1 : \beta* \neq 0$$

Vemos que la restricción H_0 puede ser considerada como un contraste general $\mathbf{A}\beta = c$ donde

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \text{ es } (m-1)\times m, \qquad c = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \text{ es } (m-1)\times 1.$$

Luego $VNE_H = (\mathbf{Y} - \mathbf{X}\hat{\beta}_H)'(\mathbf{Y} - \mathbf{X}\hat{\beta}_H) = \sum_{i=1}^n (y_i - \bar{y})^2$, que coincide con la variabilidad total del modelo inicial. $VNE = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}) = \sum_{i=1}^n (y_i - \hat{y}_i)^2$.
Por el apartado 1.3 de la descomposición de la varianza tenemos que $VNE_H - VNE = VE$.
Por último utilizaremos el estadístico F para el contraste de H_0.

$$F = \frac{VE/(m-1)}{VNE/(n-m)} \sim F_{m-1,n-m}.$$

Rechazaremos H_0 a nivel $100\alpha\,\%$ si el valor observado del estadístico anterior es superior a $F_{(m-1,n-m);1-\alpha}$.

1.4.2. Contrastes de significación individual

Recordemos que un contraste de significación individual en un modelo lineal se refiere a una prueba estadística que se realiza para determinar si un coeficiente de regresión específico es significativamente diferente de cero. En otras palabras, evaluamos si una variable predictora en particular tiene un efecto significativo sobre la variable de respuesta. La hipótesis de contraste es:

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

Para realizar este contraste se puede utilizar el contraste lineal general, con $\mathbf{A}$ una vector fila con todo ceros salvo el valor 1 en el lugar del parámetro a contrastar, o bien podemos utilizar el estadístico 6 del teorema 1.1. Es sencillo ver que el estadístico utilizado en el test F es el cuadrado del test t, por lo que son equivalentes.
Rechazaremos H_0 a nivel $100\alpha\,\%$ si el valor observado del estadístico anterior es superior a $t_{n-m;1-\alpha/2}$.

Capítulo 2

MODELOS ANOVA DE UN FACTOR FIJO

2.1. Planteamiento del modelo

Antes de adentrarnos en los detalles de los modelos ANOVA de un factor fijo, es importante destacar que este tema tiene un trasfondo más extenso que podemos explorar en mayor profundidad en [3, Capítulo 5] o [4, Capítulo 10]. En esta sección, el propósito central es estudiar la influencia que diversos niveles de un factor (una variable categórica) tienen sobre una variable de respuesta (una variable continua). A modo de ejemplo, consideremos a un investigador que se encuentra desarrollando aditivos para aumentar la durabilidad de la gasolina. En este escenario, la durabilidad de la gasolina se denota como variable $\mathbf{Y}$ (continua), y los los diferentes aditivos serían los niveles de una variable categórica. Indiquemos por y_{ij} la réplica $j = 1,\ 2, \cdots,\ n_i$, en el nivel $i = 1,\ 2, \cdots,\ m$, donde n_i es el número de réplicas en el nivel i. El modelo que se adapta a este diseño es

$$y_{ij} = \mu_i + \varepsilon_{ij} \qquad i = 1,\ 2, \cdots,\ m;\ j = 1,\ 2, \cdots,\ n_i,$$

siendo μ_i los valores medios para los diferentes niveles del factor en consideración. En este modelo tenemos m niveles de un factor y para cada uno de ellos disponemos de una muestra aleatoria simple y_{ij} con distribución $N(\mu_i, \sigma^2)$. Además supondremos independencia entre los distintos niveles del factor, es decir, los errores del modelo ε_{ij} son variables independientes e idénticamente distribuidas como $N(0, \sigma^2)$. Podemos plantear el modelo anova de un factor fijo como un modelo lineal. En el ejemplo mencionado, la variable Y a explicar es la vida útil de la gasolina y las variables explicativas son los distintos aditivos. La expresión del modelo ANOVA de un factor en su forma matricial es

$$\begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{m1} \\ y_{m2} \\ \vdots \\ y_{mn_m} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{2n_2} \\ \vdots \\ \varepsilon_{m1} \\ \varepsilon_{m2} \\ \vdots \\ \varepsilon_{mn_m} \end{pmatrix} \rightarrow \mathbf{Y} = \mathbf{X}\mu + \varepsilon.$$

La matriz de diseño $\mathbf{X}_{N\times m}$ tiene rango m. Definimos $N = \sum_{i=1}^{m} n_i$.

2.2. Estimación y contrastes

Es inmediato al plantearlo como un modelo lineal, que las estimaciones de los parámetros μ_i son $\hat{\mu}_i = \bar{y}_{i.},\ i = 1,\ 2, \cdots,\ m$, donde $\bar{y}_{i.} = (1/n_i)\sum_{j=1}^{n_i} y_{ij}$; es decir para estimar la media en el grupo i, usamos la media muestral correspondiente a ese grupo.

2.2.1. Sumas de cuadrados

Si aplicamos la descomposición de la varianza vista en el capítulo 1.3, obtenemos las correspondientes sumas de cuadrados. La nomenclatura en el caso de los modelos ANOVA es diferente, asi la variación total la denotaremos como SC_T, suma de cuadrados total, que tiene la expresión

$$SC_T = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2,$$

donde $\bar{y}_{..} = (1/N)\sum_{i=1}^{m}\sum_{j=1}^{n_i} y_{ij}$.
La variabilidad explicada es SC_E, suma de cuadrados entre grupos, que tiene la expresión

$$SC_E = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2 = \sum_{i=1}^{m} n_i(\bar{y}_{i.} - \bar{y}_{..})^2,$$

mide la variabilidad entre grupos, ponderando por el tamaño de este. Finalmente la variabilidad no explicada que llamaremos SC_D;

$$SC_D = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2,$$

que recoge la variabilidad dentro de cada grupo. Así $SC_T = SC_E + SC_D$. Por último calculamos los valores esperados de cada forma cuadrática. Para SC_D sabemos por el teorema 1.1 que $E(SC_D) = (N-m)\sigma^2$. Para el cálculo de la esperanza de SC_E, podemos verla como una forma cuadrática de las variables $\bar{y}_{i.},\ i = 1,\ 2, \cdots,\ m$ de la siguiente manera. Definimos $\bar{\mathbf{Y}}_{.} = (\bar{y}_{1.},\ \bar{y}_{2.}, \cdots,\ \bar{y}_{m.})$ y $\mathbf{A} = \text{diag}(n_1,\ n_2, \cdots,\ n_m) - (1/N)\mathbf{1}_{n_i} \times \mathbf{1}'_{n_i}$ donde $\mathbf{1}'_{n_i} = (n_1,\ n_2, \cdots,\ n_m)$, entonces tenemos que

$$SC_E = \bar{\mathbf{Y}}_{.}\mathbf{A}\bar{\mathbf{Y}}_{.}. \tag{2.1}$$

Además, $\bar{y}_{i.}$ son independientes con media μ_i y varianza σ^2/n_i, $i = 1,\ 2, \cdots,\ m$, así, podemos aplicar el lema 1.1 obteniendo que $E(SC_E) = \sum_{i=1}^{m} n_i(E[\bar{y}_{i.}] - E[\bar{y}_{..}])^2 + (m-1)\sigma^2 = \sum_{i=1}^{m} n_i(\mu_i - \bar{\mu})^2 + (m-1)\sigma^2$ donde $\bar{\mu} = (1/m)\sum_{i=1}^{m}\mu_i$.

2.2.2. Contraste de Hipótesis

Uno de los objetivos fundamentales de este modelo es evaluar si los distintos niveles del factor tienen un impacto significativo en la variable $\mathbf{Y}$. En otras palabras, estamos investigando si el valor promedio de $\mathbf{Y}$ permanece constante a lo largo de los diferentes niveles del factor. Así, el contraste fundamental en estos modelos es

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_m$$

$$H_1 : \mu_i \neq \mu_j;\ \text{para algún } i \neq j \in (1,\ 2, \cdots,\ m).$$

Si H_0 es cierta, el modelo se reduce a $y_{ij} = \mu + \varepsilon_{ij}$. La nueva estimación mínimo cuadrática del parámetro μ es $\hat{\mu}_i = \bar{y}_{..}$, y por tanto la la variabilidad no explicada en el modelo con la restricción es la suma de cuadrados total SC_T. Así, la expresión del estadístico es la análoga al contraste de significación global de un modelo lineal, es decir,

$$F = \frac{SC_E/(m-1)}{SC_D/(N-m)} \sim F_{m-1,N-m}.$$

La tabla ANOVA de un solo factor es una herramienta que resume los aspectos fundamentales del análisis de varianza. Estas tablas serán recurrentes a lo largo de los diversos modelos que examinaremos. Para simplificar, utilizaremos la abreviatura "g.l." para referirnos a los grados de libertad asociados con cada suma de cuadrados.

Tipo de grupo	Suma de cuadrados	g.l	Cuadrados Medios	Test F
Entre grupos	$SC_E = \sum_{i=1}^{m} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$	$m-1$	$SC_E/(m-1)$	$\frac{SC_E/(m-1)}{SC_D/(N-m)}$
Dentro de grupos	$SC_D = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2$	$N-m$	$SC_D/(N-m)$	
Total	$SC_T = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\cdot\cdot})^2$	$N-1$		

Cuadro 2.1: Tabla ANOVA de un factor.

2.3. Un modelo equivalente

Podemos considerar una parametrización distinta del modelo, tomando $\alpha_i = \mu_i - \mu$ donde $\mu = \bar{\mu} = (1/m)\sum_{i=1}^{m} \mu_i$

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad i = 1,\ 2, \cdots,\ m;\ j = 1,\ 2, \cdots,\ n_i,$$

notar que por como hemos definido α_i, tenemos $\sum_{i=1}^{m} \alpha_i = 0$. Esta restricción es necesaria para determinar el cálculo de los $m+1$ parámetros en un modelo de rango m.
Nuestra nueva parametrización depende de los parámetros $(\mu, \alpha_1, \alpha_2, \cdots, \alpha_{m-1})$

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

La matriz de diseño reducida $\mathbf{X}$ vuelve a tener rango máximo m. La estimación mínimo cuadrática de los parámetros es $\hat{\mu} = \bar{y}_{\cdot\cdot}$; $\hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$. Por último el test de hipótesis quedaría de la siguiente forma

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$$

$$H_1 : \alpha_i \neq 0;\ \text{para algún } i \in (1,\ 2, \cdots,\ m).$$

La ventaja de este modelo es la interpretación, ya que α_i es la diferencia de cada media a la media global y es muy útil en modelos con más factores. A partir de aquí se continua análogamente al primer modelo.

2.4. Comparaciones múltiples de medias

Cuando realizamos un análisis de varianza (ANOVA) de un factor y encontramos una diferencia estadísticamente significativa entre al menos dos grupos, es posible que deseemos realizar pruebas adicionales para identificar qué pares de grupos son significativamente diferentes entre sí. Para comparar si las parejas de medias dos a dos son iguales, podemos construir intervalos de confianza para cada pareja

y ver si el 0 pertenece o no al intervalo. El problema de este método al igual que al realizar varios contrastes de manera simultánea es que no se alcanza el nivel de significación deseado con los intervalos o en el caso de los contrastes, se superaría el nivel de significación requerido. Se trata por tanto de asegurar que los intervalos de manera simultánea contengan los verdaderos parámetros a un nivel fijo. Para ello hay varios métodos que se exponen a continuación. Para obtener información más detallada sobre estos métodos, así como explorar posibles alternativas, consultar [3, Capítulo 5]

1. **Bonferroni:** En el caso de las m medias hay un total de $m(m-1)/2 = p$ pares diferentes a comparar. Si $A_j,\ j = 1,\ 2, \cdots,\ p$, denota el evento de que el j-ésimo par de medias se declaren iguales, al realizar p pruebas de contraste de hipótesis, surge la posibilidad de cometer un error de tipo I. Este tipo de error se produce cuando rechazamos una hipótesis nula que en realidad es cierta. La probabilidad de incurrir en este error se puede expresar como: $P(A_1^c \cup A_2^c \cup \cdots \cup A_p^c)$ y esta probabilidad puede ser muy alta y no ser el nivel de significación deseado. Para tener el nivel de significación conjunto igual a α se utiliza la siguiente desigualdad: $P(A_1^c \cup A_2^c \cup \cdots \cup A_p^c) \leq P(A_1^c) + P(A_2^c) + \cdots + P(A_p^c)$ y se toma como nivel de signifcación de cada comparación, α/p, para obtener el nivel conjunto deseado.
   Bonferroni proporciona un enfoque conservador y sólido para controlar el error de tipo I y mantener un nivel de confianza en tus conclusiones. Sin embargo, debido a su naturaleza conservadora, la corrección de Bonferroni puede no detectar efectos significativos incluso si existen en realidad.

2. **Scheffé:** Scheffé proporciona intervalos de confianza simultáneos para todas las posibles parejas. El intervalo de confianza a nivel $100(1-\alpha)\,\%$ para una pareja de medias $\mu_i,\ \mu_j$ viene dado por:

$$(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) \ \pm \ s[(m-1)F_{(m-1,n-m),\alpha}[\frac{1}{n_i} + \frac{1}{n_j}]]^{1/2},$$

   donde $s^2 = 1/(n-m)\sum_{i=1}^{m}(n_i-1)s_i^2$ con $s_i^2 = \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2/(n_i-1)$. Permite contrastar simultáneamente la igualdad de medias de todos los pares a un nivel especificado. Una de las principales ventajas de la prueba de Scheffé es su robustez. Funciona bien incluso cuando las varianzas entre grupos no son iguales o cuando se tienen tamaños de muestra desiguales. La prueba de Scheffé te permite realizar comparaciones específicas que son de interés para tu investigación pero tiende a dar intervalos menos precisos que otros métodos si no hay muchas comparaciones múltiples.

3. **Tukey:** La prueba de Tukey compara todas las combinaciones posibles de pares de grupos y proporciona intervalos de confianza ajustados para las diferencias entre las medias. Es apropiada cuando se tienen muestras de tamaño similar y varianzas homogéneas entre los grupos.
   Dadas las medias independientes $\bar{y}_{i\cdot},\ i = 1,\ 2, \cdots,\ m$, con un tamaño de muestra común $n_i = n_0;\ i = 1,\ 2, \cdots,\ m$, donde $n = mn_0$ y $s^2 = SC_D/n_0$ el rango estudentizado es dado por

$$q_{m,n-m} = \max_{1 \leqslant i \leqslant m; i \neq j} |\bar{y}_{i\cdot} - \bar{y}_{j\cdot}|/s/\sqrt{n_0}.$$

   Los valores críticos de $q_{m,n-m}$ se denotan por $q_{m,n-m,\alpha}$ donde $P[q_{m,n-m} \leqslant q_{m,n-m,\alpha}] = 1-\alpha$. Para todos los p pares de medias , un intervalo de confianza de $100(1-\alpha)\,\%$ viene dado por

$$(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) - q_{m,n-m,\alpha}s/\sqrt{n_0} \leqslant (\mu_i - \mu_j) \leqslant (\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) + q_{m,n-m,\alpha}s/\sqrt{n_0}.$$

   Si el intervalo no incluye el valor cero, concluimos que $\mu_i,\ \mu_j$ son significativamente diferentes. Si el número de observaciones no es igual, remplazamos $s/\sqrt{n_0}$ por $1/2[\frac{1}{n_i} + \frac{1}{n_j}]]^{1/2}$. La prueba de Tukey controla el error de tipo I de manera efectiva al mantener un nivel global de significación. Además compara todas las combinaciones de pares de grupos, lo que significa que te proporciona una imagen completa de las diferencias significativas entre los grupos.

2.5. Homogeneidad de las varianzas

La homogeneidad de varianzas es una de las suposiciones clave en el análisis de varianza (ANOVA), ya que si no se cumple, puede afectar la validez de los resultados y llevar a conclusiones incorrectas. Se pueden utilizar pruebas estadísticas formales para evaluar la homogeneidad de varianzas, como la prueba de Levene o la prueba de Bartlett. La hipótesis nula de estos estadísticos es que las varianzas σ_i^2 correspondientes a cada nivel de factor $i \in (1,\ 2, \cdots,\ m)$ son iguales, es decir,

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_m^2$$
$$H_1 : \sigma_i^2 \neq \sigma_j^2 \text{ para algún } i,\ j \in (1,\ 2, \cdots,\ m).$$

1. **Prueba de Barlett:** Esta prueba asume que los datos siguen una distribución normal. El estadístico $M/(c+1)$ puede ser utilizado para llevar a cabo el test H_0 donde
   $M = \sum_{i=1}^{m}(n_j - 1)\ln(s^2/s_i^2), \qquad c = (1/(3(m-1))[\sum_{i=1}^{m}(1/(n_j-1)) - 1/\sum_{i=1}^{m}(n_j-1)].$
   Este estadístico sigue una distribución χ^2_{m-1}.

2. **Prueba de Levene** Se utiliza cuando los datos no cumplen con la suposición de normalidad o cuando se necesita una prueba más robusta frente a datos atípicos. En esta prueba se definen nuevas variables $Z_{ij} = |y_{ij} - \bar{y}_{i\cdot}|$. Luego se lleva a cabo un procedimiento ANOVA utilizando Z_{ij} en lugar de Y_{ij} y se realiza el contraste de igualdad de medias para estas nuevas variables. Aunque los Z_{ij} no son mutuamente independientes ni están distribuidos de manera idéntica, y no están normalmente distribuidos, esta prueba funciona razonablemente bien.

Para explorar pruebas adicionales de homogeneidad de varianza, así como métodos ANOVA no paramétricos que son utilizados cuando la suposición de homogeneidad de varianza no se cumple, véase [3, Páginas 425-429].

Capítulo 3

MODELOS ANOVA DE DOS FACTORES FIJOS

3.1. Planteamiento del modelo

En esta sección, exploraremos los modelos ANOVA de dos factores fijos. Los resultados que presentaremos a continuación se basan principalmente en [3, Capítulo 5.2.5] o [4, Capítulo 9.2].
Si la relación entre una variable Y depende de dos variables cualitativas, entonces se requiere un modelo de dos factores. Suponemos que el primer factor A tiene I niveles mientras que el segundo factor B tiene J niveles. Tomando como base el ejemplo previo, la variable Y es la vida útil de la gasolina, el factor A son los distintos tipos de aditivos que se consideran y el factor B son las diferentes marcas de las gasolinas. Queremos ver si puede haber diferencias en la vida media útil de la gasolina según el aditivo que lleve y si la marca de la gasolina también influye en esa vida media. Además, puede ocurrir que el cambio producido en la vida media al considerar dos aditivos distintos sea diferente según la marca, es decir, que la diferencia de la vida media de la gasolina en dos aditivos sea diferente en la marca 1 que en la 2. En este caso hay una interacción entre ambos factores, por lo que el modelo que se planteará deberá recoger esta interacción. Por último vamos a suponer que hay el mismo número de observaciones K para cada combinación posible de factores; este modelo se donomina balanceado. Matemáticamente, un modelo ANOVA de dos factores se puede expresar de la siguiente manera:

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \qquad i = 1,\ 2, \cdots,\ I;\ j = 1,\ 2, \cdots,\ J;\ k = 1,\ 2, \cdots,\ K, \tag{3.1}$$

donde para los distintos valores de $i,\ j,\ k$, ε_{ijk} son variables independientes con la misma distribución según $N(0, \sigma^2)$ y μ_{ij} son constantes desconocidas. En términos de las variables observables, tenemos que para i y para cada j, y_{ijk}, $k = 1,\ 2, \cdots,\ K$, es una muestra aleatoria simple de una variable $N(\mu_{ij}, \sigma^2)$, donde hay independencia entre las observaciones de los distintos niveles de los factores. Con este planteamiento el modelo (3.1) se puede escribir como

$$\mathbf{Y} = \mathbf{X}\mu + \varepsilon.$$

Donde $\mathbf{X}$ es una matriz de dimensión $N \times IJ$ con $N = I \times J \times K$ y rango máximo $I \times J$.
A continuación, presentaremos las definiciones de las medias que emplearemos a lo largo de este capítulo.

$$\bar{y}_{i.} = 1/(JK)\sum_{j,k} y_{ijk}\ ,\quad \bar{y}_{.j} = 1/(IK)\sum_{i,k} y_{ijk}\ ,\quad \bar{y}_{ij.} = 1/(K)\sum_{k} y_{ijk}\ ,\quad \bar{y}_{..} = 1/(IJK)\sum_{i,j,k} y_{ijk}\ ,$$

$$\bar{\mu}_{i.} = (1/J)\sum_{j} \mu_{ij}\ ,\quad \bar{\mu}_{.j} = (1/I)\sum_{i} \mu_{ij}\ ,\quad \bar{\mu}_{..} = (1/IJ)\sum_{i,j} \mu_{ij}\ .$$

En este modelo la estimación de μ_{ij} son $\hat{\mu}_{ij} = \bar{y}_{ij.},\ i = 1,\ 2, \cdots,\ I;\ j = 1,\ 2, \cdots,\ J$, por tanto, la variabilidad no explicada en el modelo es $SCR = \sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ijk} - \bar{y}_{ij.})^2$ que sabemos, por el teorema 1.1

del Capítulo 1, que (normalizada por la varianza) se distribuye como $\chi_{IJ(K-1)}$ (notar que $IJ(K-1)$ es el número de observaciones menos el número de parámetros). La cuestión ahora es ver qué combinaciones lineales μ_{ij} son de interés en este tipo de modelos. Una de las primeras cuestiones que se pueden plantear es si el efecto de un factor varía según los niveles del otro factor. Por lo tanto, si no hay interacción se verifica: $\mu_{i_1 j} - \mu_{i_2 j}$ no depende de j, $\forall\, i_1,\ i_2 \in i = 1,\ 2, \cdots,\ I$; $\forall\ j \in 1,\ 2, \cdots,\ J$. Así, la ausencia de interacción se formula como

$$\mu_{i_1 j} - \mu_{i_2 j} = \bar{\mu}_{i_1 .} - \bar{\mu}_{i_2 .} \qquad \forall\, i_1,\ i_2 \in i = 1,\ 2, \cdots,\ I;\ \forall\ j \in 1,\ 2, \cdots,\ J,$$

equivalentemente a

$$\mu_{ij} = \bar{\mu}_{i.} + \bar{\mu}_{.j} - \bar{\mu}_{..} \tag{3.2}$$

Obsérvese que obtendriamos esta misma condición si partiéramos del segundo factor en lugar del primero. Así, la condición (3.2) es la condición que indica la ausencia de interacción entre ambos factores. Supongamos ahora que se verifica la condición (3.2), la siguiente hipótesis que nos planteamos es si un factor no afecta a la variable respuesta, es decir, para cada j $\mu_{ij} = \bar{\mu}_{.j}$; $\forall\, i \in 1,\ 2, \cdots,\ I$, lo cual es equivalente por (3.2) a $\bar{\mu}_{i.} - \bar{\mu}_{..} = 0$; $\forall\, i \in 1,\ 2, \cdots,\ I$. De la misma manera para el otro factor. Si representásemos gráficamente las medias de dos modelos, uno sin interacción y el otro con ella, obtendríamos algo similar a la siguiente situación:

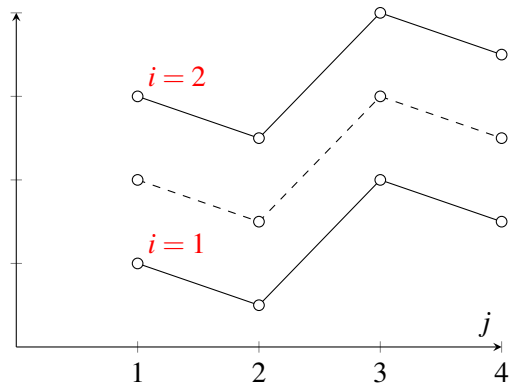

Figura 3.1: Sin interacción

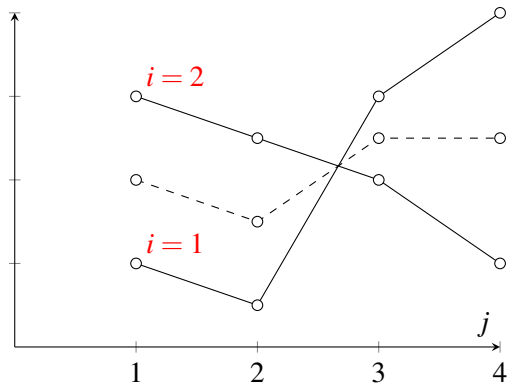

Figura 3.2: Con interacción

Una vez que hemos identificado los efectos que deseamos estudiar, podemos llevar a cabo una reparametrización del modelo de la siguiente manera:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \qquad i = 1,\ 2, \cdots,\ I;\ j = 1,\ 2, \cdots,\ J;\ k = 1,\ 2, \cdots,\ K, \tag{3.3}$$

donde $\mu = \bar{\mu}_{..}$ es la media global común a los dos factores; $\alpha_i = \bar{\mu}_{i.} - \bar{\mu}_{..}$ el efecto del primer factor; $\beta_j = \bar{\mu}_{.j} - \bar{\mu}_{..}$ el efecto del segundo y $(\alpha\beta)_{ij} = \mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \bar{\mu}_{..}$ es la interacción entre ambos factores. Obsérvese que $\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{i,j}^{I,J} (\alpha\beta)_{ij} = 0$. Esta parametrización tiene una matriz $\mathbf{X}$ de rango no completo, pero se utiliza para una mejor interpretación. Hay autores que tratan el modelo con esta parametrización basándose en los resultados de modelos lineales con matrices de diseño singulares (véase [5, Capítulo 6.4]).
Los contrastes de interés con respecto a esta reparametrización son los siguientes:

1. H^{AB}; ${H_0}^{AB} : (\alpha\beta)_{ij} = 0;\ \forall\, i,\ j.$
   ${H_1}^{AB} : (\alpha\beta)_{ij} \neq 0;$ para algún $i,\ j.$
2. H^{A}; ${H_0}^{A} : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$
   ${H_1}^{A} : \alpha_i \neq 0$ para algún $i \in (1,\ 2, \cdots,\ I).$

3. H^B; ${H_0}^B : \beta_1 = \beta_2 = \cdots = \beta_J = 0$
   ${H_1}^B : \beta_j \neq 0$ para algún $j \in (1,\ 2, \cdots,\ J)$.

H^A (o H^B) contrasta si el efecto de A es cero (o el de B). Pero notar que si H^{AB} ha sido rechazada entonces, $\alpha_i = 0$ no es esquivalente a que $\mu_{ij} = \bar{\mu}_{.j}\ \forall\ i,\ j$, que sería realmente la hipótesis de que el primer factor no influye.

3.2. Estimación y contrastes

Es inmediato que bajo el modelo (3.3) la estimación mínimo cuadrática de los parámetros son $\hat{\mu} = \bar{y}_{ij.}$; $\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$; $\hat{\beta}_j = \bar{y}_{.j} - \bar{y}_{..}$; $\widehat{(\alpha\beta)}_{ij} = \bar{y}_{ij.} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}$. En el modelo (3.1) obtenemos directamente del modelo lineal general que una estimación insesgada para σ^2 es $\hat{\sigma}^2 = SCR/(IJ)(K-1) = \sum_{i,j,k}(y_{ijk} - \bar{y}_{ij.})^2/(IJ)(K-1) \quad \forall\ i,\ j,\ k$.

3.2.1. Suma de Cuadrados

Para entender mejor y deducir las distribuciones y los grados de libertad de las sumas de cuadrados, vamos a plantear cuál es la variabilidad no explicada en el modelo lineal sujeto a cada hipótesis que aparecen en este tipo de modelos, es decir,

$$\min_{\alpha_i, \beta_j, (\alpha\beta)_{ij}} \sum_{i,j,k} (y_{ijk} - \mu - \alpha_i - \beta_j - (\alpha\beta)_{ij})^2,$$

con $\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{i,j}^{IJ} (\alpha\beta)_{ij} = 0$, bajo H_{AB}, H_A, H_B . Para ello consideremos la siguiente descomposición fundamental de la suma de cuadrados, véase [5, Proposición 6.2] o [10, Sección 9.2.2] para una demostración completa.

$$\sum_{i,j,k} (y_{ijk} - \mu - \alpha_i - \beta_j - (\alpha\beta)_{ij})^2 = \sum_{i,j,k} (y_{ijk} - \mu - \alpha_i - \beta_j - (\alpha\beta)_{ij} + \bar{y}_{..} - \bar{y}_{..})^2 =$$

$$\sum_{i,j,k} (\bar{y}_{..} - \mu)^2 + \sum_{i,j,k} (y_{ijk} - \alpha_i - \beta_j - (\alpha\beta)_{ij} - \bar{y}_{..} + \bar{y}_{i.} - \bar{y}_{i.})^2 =$$

$$\sum_{i,j,k} (\bar{y}_{..} - \mu)^2 + \sum_{i,j,k} (\bar{y}_{i.} - \bar{y}_{..} - \alpha_i)^2 + \sum_{i,j,k} (y_{ijk} - \beta_j - (\alpha\beta)_{ij} - \bar{y}_{i.} + \bar{y}_{j.} - \bar{y}_{j.})^2 =$$

$$\sum_{i,j,k} (\bar{y}_{..} - \mu)^2 + \sum_{i,j,k} (\bar{y}_{i.} - \bar{y}_{..} - \alpha_i)^2 + \sum_{i,j,k} (\bar{y}_{j.} - \bar{y}_{..} - \beta_j)^2 + \sum_{i,j,k} (\bar{y}_{ij.} - \bar{y}_{i.} - \bar{y}_{j.} + \bar{y}_{..} - (\alpha\beta)_{ij})^2 + \sum_{i,j,k} (y_{ijk} - \bar{y}_{ij.})^2.$$

Por tanto, de la igualdad anterior vemos que VNE en el modelo con $(\alpha\beta)_{ij} = 0;\ \forall\ i,\ j$, se alcanza en:

$$VNE_{H_{AB}} = K \sum_{i,j} (\bar{y}_{ij.} - \bar{y}_{i.} - \bar{y}_{j.} + \bar{y}_{..})^2 + \sum_{i,j,k} (y_{ijk} - \bar{y}_{ij.})^2 = SC_{AB} + SCR.$$

Análogamente tenemos para H_A y H_B respectivamente que

$$VNE_{H_A} = JK \sum_{i} (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{ijk} (y_{ijk} - \bar{y}_{ij.})^2 = SC_A + SCR.$$

$$VNE_{H_B} = IK \sum_{i} (\bar{y}_{j.} - \bar{y}_{..})^2 + \sum_{ijk} (y_{ijk} - \bar{y}_{ij.})^2 = SC_B + SCR.$$

Donde $SC_A = JK \sum_i (\bar{y}_{i.} - \bar{y}_{..})^2$ es la suma de cuadrados debido al factor A, $SC_B = IK \sum_i (\bar{y}_{j.} - \bar{y}_{..})^2$ es la suma de cuadrados debido al factor B, $SC_{AB} = K \sum_{i,j} (\bar{y}_{ij.} - \bar{y}_{i.} - \bar{y}_{j.} + \bar{y}_{..})^2$ es la suma de cuadraos debido a la interacción entre factores y $SCR = \sum_{ijk} (y_{ijk} - \bar{y}_{ij.})^2$ es la suma de cuadrados residual.

Además, es inmediato que la suma de cuadrados total es

$$SCT = \sum_{i,j,k} (y_{ijk} - \bar{y}_{..})^2 = SC_A + SC_B + SC_{AB} + SCR. \tag{3.4}$$

Finalmente vamos a calcular las esperanzas de estas formas cuadráticas. Aplicando la fórmula del lema 1.1 para el cálculo de esperanzas de formas cuadráticas y teniendo en cuenta que $Var(\bar{y}_{.}) = 1/(IKJ)\sigma^2$, tenemos que $E(SC_A) = JK\sum_i(\mu_{i.} - \bar{\mu})^2 + (I-1)\sigma^2 = JK\sum_i \alpha_i^2 + (I-1)\sigma^2$. Análogamente con SC_B, $E(SC_B) = IK\sum_j \beta_j^2 + (J-1)\sigma^2$. En el término asociado a la interacción, la matriz que define la forma cuadrática tiene como traza $IJ - I - J + 1 = (I-1)(J-1)$ y por tanto $E(SC_{AB}) = K\sum_{ij}(\alpha\beta)_{ij} + \sigma^2(J-1)(I-1)$. Notar que todos ellos tienen la misma esperanza cuando los efectos de los factores son nulos.

3.2.2. Contraste de Hipótesis

Las expresiones de los estadísticos y sus distribuciones para los contrastes se deducen de manera inmediata de las sumas de cuadrados obtenidas en el párrafo anterior y del teorema 1.2 del capítulo 1. En la práctica, toda esta información se muestra en forma de tabla como aparece a continuación.

Tipo de grupo	Suma de cuadrados	g.l	Cuadrados Medios	Test F
Factor A	$SC_A = JK\sum_i(\bar{y}_{i.} - \bar{y}_{..})^2$	$I-1$	$SC_A/(I-1)$	$\frac{SC_A/(I-1)}{SCR/(IJ)(K-1)}$
Factor B	$SC_B = IK\sum_j(\bar{y}_{.j} - \bar{y}_{..})^2$	$J-1$	$SC_B/(J-1)$	$\frac{SC_B/(J-1)}{SCR/(IJ)(K-1)}$
Interacción	$SC_{AB} = K\sum_{i,j}(\bar{y}_{ij.} - \bar{y}_{i.} - \bar{y}_{j.} + \bar{y}_{..})^2$	$(I-1)(J-1)$	$SC_{AB}/(I-1)(J-1)$	$\frac{SC_{AB}/(I-1)(J-1)}{SCR/(IJ)(K-1)}$
Residuo	$SCR = \sum_{ijk}(y_{ijk} - \bar{y}_{ij.})^2$	$(IJ)(K-1)$	$SCR/(IJ)(K-1)$	
Total	$SCT = \sum_{i,j,k}(y_{ijk} - \bar{y}_{..})^2$	$IJK-1$		

Cuadro 3.1: Tabla ANOVA de dos factores con interacción

3.3. Modelo no balanceado

Como se ha comentado al principio del capítulo, un modelo balanceado es aquél en el que el número de observaciones en cada combinación de niveles de los factores es el mismo. En otro caso diremos que es un diseño o modelo no balanceado. Vamos a dar algunas indicaciones sobre este caso, pero para un análisis más completo véase [6, Capítulo 9].

Las diversas sumas de cuadrados vistas en 3.2.1 son aditivas, y representan una descomposición ortogonal del vector $\mathbf{Y} = y_{ijk}$. Esta descomposición ortogonal hace que las sumas de cuadrados que aparecen en el numerador de los estadísticos considerados no dependan del modelo de partida, es decir, si queremos contrastar si el efecto del factor A es significativo, no importa si partimos del modelo con o sin interacción en el cálculo de la suma de cuadrados del numerador. En el caso de un modelo no balanceado esto no es así, la descomposición vista en 3.2.1 no se verifica. Esta no ortogonalidad hace que haya diferentes formas de contrastar los efectos de los factores utilizando distintas sumas de cuadrados. Para ello, vamos a definir $SCR(A,B,AB)$ a la suma de cuadrados residual que se ha considerado hasta ahora, es decir, con el factor A, factor B y la interacción AB. Definimos $SCR(A,AB)$ ($SCR(B,AB)$) a la suma de cuadrados residual considerando sólo el factor A (B) y la interacción en el modelo; $SCR(A,B)$ es la suma de cuadrados residual con los dos factores sin la interacción; $SCR(A)$ y $SCR(B)$ son las sumas cuando se incluye solo el factor A o el B en el modelo y finalmente $SCR(1)$ denotará la suma de cuadrados residual cuando sólo se introduce el término constante.

Hay 3 tipos de sumas de cuadrados para un modelo anova con dos factores fijos, estas sumas se han denominado tradicionalmente de tipo I, II y III. En la siguiente tabla, aparece un resumen de los tipos de sumas de cuadrados en el contraste del efecto de cada uno de los factores y de la interacción.

Término	Sumas de cuadrados tipo I	Sumas de cuadrados tipo II	Sumas de cuadrados tipo III
A	$SCR(1) - SCR(A)$	$SCR(B) - SCR(A,B)$	$SCR(B,AB) - SCR(A,B,AB)$
B	$SCR(A) - SCR(A,B)$	$SCR(A) - SCR(A,B)$	$SCR(A,AB) - SCR(A,B,AB)$
AB	$SCR(A,B) - SCR(A,B,AB)$	$SCR(A,B) - SCR(A,B,AB)$	$SCR(A,B) - SCR(A,B,AB)$

Cuadro 3.2: Tabla de las sumas de cuadrados en el caso no balanceado.

1. Tipo I o método secuencial: como vemos las sumas de cuadrados se van calculando añadiendo uno a uno los términos del modelo. Este método tiene el inconveniente de que depende del orden en el que se incluyen los factores.

2. Tipo II o método jerárquico o parcialmente secuencial: en este modelo las sumas de cuadrados de cada uno de los factores se calculan incluyendo en el modelo el otro factor sin la interacción.

3. Tipo III o método marginal: en este tipo de sumas, se considera el término de la interacción en las sumas de cuadrados para ambos factores.

Como vemos la suma de cuadrados residual para la interacción es la misma para los 3 tipos. La manera más aceptada de proceder es utilizar el método II en el caso de ausencia de interacción y el método III en otro caso.

3.4. Extensión del modelo

En las secciones anteriores han sido tratados los diseños de uno y dos factores y se ha estudiado cómo descomponer adecuadamente la variabilidad. Los diseños en los que intervienen tres o más factores pueden estudiarse también descomponiendo adecuadamente la variabilidad total

$$SC_T = \sum (y_{ij\cdots m} - \bar{y})^2,$$

en diferentes sumas de cuadrados, más una suma de cuadrados residual. Veamos cómo debe procederse para un diseño de cuatro factores que indicaremos a, b, c y d con A, B, C y D niveles respectivamente, y suponiendo que las interacciones de cuarto orden son cero. El modelo sería:

$$y_{ijkm} = \mu + \alpha_i^A + \alpha_j^B + \alpha_k^C + \alpha_m^D + \alpha_{ij}^{AB} + \alpha_{ik}^{AC} + \alpha_{im}^{AD} + \alpha_{jk}^{BC} + \alpha_{jm}^{BD} + \alpha_{km}^{CD} + \alpha_{ijm}^{ABD} + \alpha_{ikm}^{ACD} + \alpha_{jkm}^{BCD} + \varepsilon_{ijkr}$$

$i = 1,\ 2, \cdots,\ a;\ j = 1,\ 2, \cdots,\ b;\ k = 1,\ 2, \cdots,\ c;\ m = 1,\ 2, \cdots,\ d$ y siendo y_{ijkm} la respuesta para los niveles $i,\ j,\ k,\ m$ de $A,\ B,\ C,\ D$; μ la media general; $\alpha_i^A,\ \alpha_j^B,\ \alpha_k^C,\ \alpha_k^D$ los efectos principales de $A,\ B,\ C,\ D$; $\alpha_{ij}^{AB},\ \alpha_{ik}^{AC}, \cdots,\ \alpha_{km}^{CD}$ las interacciones de orden dos entre los factores $A,\ B,\ C,\ D$; $\alpha_{ijk}^{ABC},\ \alpha_{ijm}^{ABD}, \alpha_{ikm}^{ACD},\ \alpha_{jkm}^{BCD}$ las interacciones de orden tres entre los factores $A,\ B,\ C,\ D$; ε_{ijkm} la desviación aleatoria con distribución $N(0, \sigma^2)$. Por último, es necesario aplicar la restricción de que la suma de los parámetros α en relación a todos sus subíndices sea igual a cero. A modo de ejemplo

$$\sum_j \alpha_{jkm}^{BCD} = \sum_k \alpha_{jkm}^{BCD} = \sum_m \alpha_{jkm}^{BCD} = 0.$$

Las sumas de cuadrados y sus correspondientes grados de libertad se encuentran en el cuadro 6.1, del anexo A. Las descomposiciones ortogonales para el modelo de 4 factores, así como para cualquier otro número de factores, puede programarse por ordenador siguiendo el algoritmo propuesto por Hartley[9, Capítulo 20]. La principal dificultad de estos diseñños es la gran cantidad de observaciones necesarias, de modo que en la práctica no se consideran diseños con más de cuatro factores. En algunos casos se puede suponer que las interacciones altas son nulas y estimar el resto de parámetros. Esta es la propuesta de los diseños en cuadrados latinos y greco-latinos que permiten estimar los efectos principales con el mínimo de observaciones (véase [8, Capítulo 6]).

Capítulo 4

MODELOS ANOVA CON FACTORES ALEATORIOS

En modelos con efectos fijos se supone que se tiene un número fijo de valores en la variable factor y que de cada uno de ellos se ha elegido una muestra de individuos. Sin embargo en el análisis de ANOVA de un factor aleatorio, el conjunto de niveles del factor constituye una muestra de un conjunto más grande de posibles valores. Por ejemplo, podríamos estar interesados en estudiar cómo diferentes maestros influyen en el rendimiento de los estudiantes en un examen, considerando que los maestros son una muestra aleatoria de una población más grande de maestros.

4.1. Modelo ANOVA con un único factor aleatorio.

En este caso el objetivo no es comparar las medias de los niveles, sino estimar la varianza del factor aleatorio y comprender cómo contribuye a la variabilidad total de la variable de resultado. El modelo se escribe:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad i = 1,\ 2, \cdots,\ m;\ j = 1,\ 2, \cdots,\ n_i,$$

donde m es el número de niveles considerados y n_i el número de observaciones en el nivel i; μ es constante y es la media global de la variable de interés; α_i y ε_{ij} son variables aleatorias independientes e idénticamente distribuidas siguiendo distribuciones $N(0,\sigma_\alpha^2)$ y $N(0,\sigma_\varepsilon^2)$, respectivamente. Además α_i y ε_{ij} son independientes entre si. Los tres parámetros del modelo, que se suponen fijos pero desconocidos, son $\mu,\ \sigma_\alpha^2,\ \sigma_\varepsilon^2$.

$$E(y_{ij}) = \mu, \qquad \mathrm{Var}(y_{ij}) = \mathrm{Var}(\alpha_i) + \mathrm{Var}(\varepsilon_{ij}) = \sigma_y^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2,\ \forall\ i,\ j.$$

$$\mathrm{Cov}(y_{ij}, y_{ik}) = E[(\alpha_i + \varepsilon_{ij})(\alpha_i + \varepsilon_{ik})] = \sigma_\alpha^2,\ j \neq k, \qquad \mathrm{Cov}(y_{ij}, y_{lk}) = 0\ i \neq l,\ j \neq k.$$

Es muy importante notar que ahora las y_{ij} no son independientes. Fijado i, y_{i1}, $y_{i2}, \cdots$, y_{in_i}, tienen en común el término aleatorio. Luego la matriz de varianzas-covarianzas Σ de $\mathbf{Y}$ ya no es diagonal, sino que tiene una estructura llamada simetría compuesta, es decir, es diagonal por bloques (véase 6.1). Por lo tanto $\mathbf{Y}$ está normalmente distribuido con media constante μ y matriz de varianzas covarianzas Σ, es decir, $\mathbf{Y} \sim N(\mu, \Sigma)$.

En primer lugar establecemos la misma notación que en el capítulo 2, es decir, $\bar{y}_{..}$, $\bar{y}_{i.}$, $\bar{\varepsilon}_{i.}$, $\bar{\varepsilon}_{..}$ y $\bar{\alpha}_{.}$, $i = 1,\ 2, \cdots,\ m$, que son las nuevas medias correspondientes y de nuevo $N = \sum_{i=1}^{m} n_i$. La descomposición de la variabilidad total vista en el apartado 2.2.1 sigue siendo válida ya que no se ha utilizado ninguna de las hipótesis del modelo sino sólo la definición de las medias consideradas. Así, tenemos

$$SCT = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 = N\sum_{i=1}^{m}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 = SC_E + SC_D.$$

En primer lugar, observemos que no podemos deducir las esperanzas de cada suma de cuadrados a partir de los modelos anteriores, dado que la hipótesis de independencia de y_{ij} ya no se satisface. Podemos utilizar la expresión (2.1) y así aplicar el lema 1.1, aunque antes es conveniente examinar la matriz de varianzas-covarianzas de $\bar{y}_{i.}$.

$$\mathrm{Var}(\bar{y}_{i.}) = \mathrm{Var}(1/n_i \sum_{i=1}^{n_i} y_{ij}) = 1/n_i^2(\sigma_y^2 n_i + (n_i-1)n_i\sigma_\alpha^2) = (\sigma_y^2/n_i) + (n_i-1)/n_i\sigma_\alpha^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2/n_i.$$

$i = 1,\ 2, \cdots,\ m$. Además, $\bar{y}_{i.},\ i = 1,\ 2, \cdots,\ m$ son independientes con media μ, por lo que

$$E(SC_E) = E[\bar{\mathbf{Y}}_.]\mathbf{A}E[\bar{\mathbf{Y}}_.] + \mathrm{tr}(\mathbf{A}\Sigma) = \sigma_\varepsilon^2 \mathrm{tr}(\mathbf{A}/\mathrm{diag}(1/n_1,\ 1/n_2, \cdots,\ 1/n_m)) + \sigma_\alpha^2 \mathrm{tr}(\mathbf{A}) = \sigma_\varepsilon^2 \sum_{i=1}^{m}(1 - n_i/N)$$

$$+ \sigma_\alpha^2 N - \sigma_\alpha^2 \sum_{i=1}^{m} n_i^2/N = \sigma_\varepsilon^2(m - \sum_{i=1}^{m} n_i/N) + \sigma_\alpha^2(N^2 - \sum_{i=1}^{m} n_i^2)/N = \sigma_\varepsilon^2(m-1) + \sigma_\alpha^2(N^2 - \sum_{i=1}^{m} n_i^2)/N.$$

Así, denotando como $n_0 = (N^2 - \sum_{i=1}^{m} n_i^2)/N$, tenemos que $E(SC_E) = (m-1)\sigma_\varepsilon^2 + n_0\sigma_\alpha^2$. De manera más sencilla se puede ver que $E(SC_D) = N(m-1)\sigma_\varepsilon^2$. Resolviendo el sistema resultante de igualar los cuadrados medios con los cuadrados esperados podemos proporcionar los estimadores insesgados de las varianzas $\sigma_\varepsilon^2,\ \sigma_\alpha^2$ como

$$\hat{\sigma}_\varepsilon^2 = SC_D/(N-m), \quad \hat{\sigma}_\alpha^2 = (1/n_0)[SC_E/(m-1) - SC_D/(N-m)].$$

Obsérvese que $\hat{\sigma}_\alpha^2$ puede ser negativo, por lo que a pesar de ser insesgado no suele utilizarse. Hay distintos métodos que proponen estimadores para σ_α^2 con mejores propiedades, véase [2, Capítulo 3.1] o [7, Capítulo 11.4]. Como hemos comentado, el objetivo principal es ver si la varianza del factor aleatorio contribuye a la variabilidad total de la variable resultado, para ello vamos a plantear el siguiente contraste de hipótesis

$$H_0 : \sigma_\alpha^2 = 0$$
$$H_1 : \sigma_\alpha^2 \neq 0$$

Aunque las propiedades de las formas cuadráticas en cuanto a la independencia y distribución no se pueden deducir del modelo lineal general, sí se puede realizar el contraste. Obsérvese que bajo H_0, las variables $\alpha_i = 0,\ i = 1,\ 2, \cdots,\ m$, por lo que el modelo se reduce a $y_{ij} = \mu + \varepsilon_{ij},\ i = 1,\ 2, \cdots,\ m;\ j = 1,\ 2, \cdots,\ n_i$, y hemos visto en el capítulo 2 que el estadístico (bajo H_0) es

$$F = \frac{SC_E/(m-1)}{SC_D/(N-m)} \sim F_{m-1,N-m}.$$

Tipo de grupo	Suma de cuadrados	g.l	Cuadrados Medios	Test F
Tratamientoss	$SC_E = N\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2$	$m-1$	$SC_E/(m-1)$	$\frac{SC_E/(m-1)}{SC_D/(N-m)}$
Error	$SC_D = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$	$N-m$	$SC_D/(N-m)$	
Total	$SC_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$	$N-1$		

Cuadro 4.1: Tabla ANOVA de un factor aleatorio

4.2. Extensión de los modelos anteriores

Como hemos visto en el apartado anterior la inclusión de un factor aleatorio indica que los niveles son una posible muestra de una población. Este enfoque tiene una gran aplicación en los diseños con medidas repetidas de los mismos individuos. Supongamos que se quieren aplicar dos tratamientos distintos a una serie de individuos, pero se aplican ambos tratamientos a cada uno de ellos. El objetivo sería ver si hay cambio en alguna variable de interés (nivel de mejora por ejemplo) entre los tratamientos. Aunque los tratamientos son los niveles del factor, no podriamos utilizar el modelo descrito en el capítulo 2 ya que las observaciones de la variable **Y** en los tratamientos no son independientes. Para ello deberiamos introducir un factor aleatorio que recogiera la correlación entre las medidas de los mismos individuos y un factor fijo que fuese los distntos tratamientos. Supongamos ahora, que medimos en hombres y mujeres una variable y lo hacemos en varios instantes de tiempo para cada individuo. Queremos saber si la respuesta es igual en hombres que en mujeres. También tendríamos medidas repetidas pero los individuos del primer nivel (hombres) del factor son distintos a los del segundo nivel (mujeres), así, no sería aplicable el modelo con un factor fijo y uno aleatorio. Como vemos, hay muchas situaciones y muchas formas de diseñar un experimento, por lo que la variedad de modelos es elevada. A continuación mostramos algunos modelos de interés con la tabla ANOVA correspondiente. Para una lectura en profundidad de estos modelos véase [8, Capítulo 8] o [4, Capítulo 11].

4.2.1. Modelo ANOVA de un factor fijo y un factor aleatorio

Este enfoque también es conocido como modelo mixto, ya que combina factores de naturaleza fija y aleatoria. Este modelo correspondería al ejemplo de los tratamientos y los pacientes, donde los tratamientos (**A**) son un factor fijo y los pacientes (**B**) es un factor aleatorio. Teniendo en cuenta la interacción y considerando tamaños muestrales iguales en cada cruce de factores, es decir balanceado, el modelo se escribe:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \qquad i = 1,\ 2, \cdots,\ I;\ j = 1,\ 2, \cdots,\ J;\ k = 1,\ 2, \cdots,\ K,$$

donde μ y α_i son constantes con $\sum_i^I \alpha_i = 0$; β_j, $(\alpha\beta)_{ij}$ y ε_{ijk} son variables aleatorias independientes e idénticamente distribuidas siguiendo distribuciones $N(0,\sigma_\beta^2)$, $N(0,\sigma_{\alpha\beta}^2)$ y $N(0,\sigma_\varepsilon^2)$, respectivamente. Además β_i,$(\alpha\beta)_{ij}$ y ε_{ijk} son indendientes entre si. Los parámetros del modelo, que se suponen fijos pero desconocidos, son μ, α_i, σ_β^2, $\sigma_{\alpha\beta}^2$, σ_ε^2. Dado que estamos trabajando con un modelo balanceado, podemos utilizar la descomposición de la variabilidad total (3.3). Sin embargo, dado que la matriz de varianzas-covarianzas de **Y** ya no es diagonal, las esperanzas de las sumas de cuadrados de los términos aleatorios cambian, siendo en este caso:

$$E(SC_A) = (I-1)\sigma_\varepsilon^2 + \sum_{i,j,k} \alpha_i^2, \qquad E(SC_B) = \sigma_\varepsilon^2 + IK\sigma_\beta^2 + K\sigma_{\alpha\beta}^2, \qquad E(SC_{AB}) = \sigma_\varepsilon^2 + K\sigma_{\alpha\beta}^2.$$

En este diseño los contrastes de interés serían los siguientes:

$$H_0{}^A : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0, \qquad H_0{}^B : \sigma_\beta^2 = 0, \qquad H_0{}^{AB} : \sigma_{\alpha\beta}^2 = 0.$$

Para contrastar estas hipótesis, empleamos los estadísticos F que se encuentran en la tabla ANOVA (3.1). En el ejemplo de los distintos tratamientos que se aplican a una serie de pacientes, el modelo se debe plantear sin interacción. Así el segundo contraste significaría que no hay correlación en las mediciones repetidas de los pacientes, es decir, que las observaciones de un mismo paciente se pueden considerar independientes y el primer contraste que no hay efecto de los distintos tratamientos en los pacientes. Cuando se aplica a medidas repetidas, solo el primer contraste suele ser de interés, ya que siempre se supone que las observaciones repetidas de un individuo están correladas.

4.2.2. Modelo ANOVA de factores anidados

Un diseño anidado en el contexto de un análisis de varianza (ANOVA) implica que uno o más factores se encuentran anidados dentro de otros factores. Esto significa que todos los niveles de un factor no están presentes en los niveles de otro factor y esto puede complicar la estructura del diseño experimental. Decimos que el factor B está anidado en el A, cuando cada nivel del factor B aparece asociado a un único nivel del factor A. Así, en un modelo mixto, es decir, un modelo anidado mixto, tendría la siguiente expresión:

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk} \qquad i = 1,\ 2, \cdots,\ I;\ j = 1,\ 2, \cdots,\ J;\ k = 1,\ 2, \cdots,\ K,$$

donde μ y α_i son constantes con $\sum_i^I \alpha_i = 0$; $\beta_{j(i)}$ y ε_{ijk} son variables aleatorias independientes e idénticamente distribuidas siguiendo distribuciones $N(0, \sigma_\beta^2)$ y $N(0, \sigma_\varepsilon^2)$, respectivamente.
La descomposición fundamental de la variabilidad total ya no es como se describe en (3.4), ya que la interpretación en este caso es distinta. En esta situación, la descomposición se presenta de la siguiente manera:

$$SCT = \sum_{i,j,k}(y_{ijk} - \bar{y}_{..})^2 = JK\sum_i(\bar{y}_{i.} - \bar{y}_{..})^2 + K\sum_{i,j,k}(\bar{y}_{ij} - \bar{y}_{i.})^2 + \sum_{ijk}(y_{ijk} - \bar{y}_{ij.})^2 = SC_A + SC_{B|A} + SCR.$$

Los valores esperados de la suma de cuadrados son:

$$E(SC_A) = (I-1)(\sigma_\varepsilon^2 + K\sigma_\beta^2) + \sum_{i,j,k}\alpha_i^2, \quad E(SC_{B|A}) = I(J-1)(K\sigma_\beta^2 + \sigma_\varepsilon^2), \quad E(SCR) = IJ(K-1)\sigma_\varepsilon^2.$$

En este diseño los contrastes de interés serían los siguientes:

$$H_0^A : H_0{}^A : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0, \qquad H_0^B : \sigma_\beta^2 = 0.$$

Tipo de grupo	Suma de cuadrados	g.l	Cuadrados Medios	Test F
Factor A	$SC_A = JK\sum_i(\bar{y}_{i.} - \bar{y}_{..})^2$	$I-1$	$SC_A/(I-1)$	$\frac{SC_A/(I-1)}{SCR/(IJ)(K-1)}$
Factor B	$SC_{B(A)} = K\sum_j(\bar{y}_{.j} - \bar{y}_{..})^2$	$I(J-1)$	$SC_{B(A)}/I(J-1)$	$\frac{SC_{B(A)}/I(J-1)}{SCR/(IJ)(K-1)}$
Residuo	$SCR = \sum_{ijk}(y_{ijk} - \bar{y}_{ij.})^2$	$(IJ)(K-1)$	$SCR/(IJ)(K-1)$	
Total	$SCT = \sum_{i,j,k}(y_{ijk} - \bar{y}_{..})^2$	$IJK-1$		

Cuadro 4.2: Tabla ANOVA de dos factores aleatorios anidados.

Este modelo se puede aplicar al ejemplo que se ha comentado en la introducción. El objetivo es saber si una variable clínica tiene un comportamiento diferente en hombres y mujeres y para ello se realiza en cada paciente varias mediciones. En este caso el sexo es un factor fijo y los pacientes son los niveles del factor aleatorio. Así, para cada nivel de la variable sexo, por ejemplo hombre, tenemos tantos niveles del factor aleatorio como individuos haya en el nivel, es decir, el número de hombres. Lo mismo para el nivel mujer. Para cada individuo $j(i)$, tendremos K medidas repetidas. Un estudio de estos modelos en el caso no balanceado o modelos con más factores puede encontrarse en [7, Capítulo 15].

Capítulo 5

ANÁLISIS DE DATOS CON MODELOS ANOVA

En este capítulo, vamos a analizar un conjunto de datos reales, utilizando principalmente los modelos vistos en los capítulos 2 y 3. También aplicaremos los modelos con un factor aleatorio para el estudio de medidas repetidas. Los datos provienen de la Encuesta Nacional de Examen de Salud y Nutrición (NHANES), que puedes encontrar en este enlace. NHANES es un programa de investigación diseñado para evaluar el estado de salud y nutrición tanto en adultos como en niños en los Estados Unidos. Esta encuesta se distingue por su enfoque combinado de entrevistas y exámenes físicos, lo que proporciona una visión integral de la salud de la población. Como era de esperar, la información recopilada en la encuesta es muy amplia y se encuentra dispersa en diversos archivos. Para ello fue necesario llevar a cabo un proceso de selección de las variables a analizar. Las variables que se han considerado para el ejemplo son:
Presión diastólica (BPXDI1) medida en milímetros de mercurio (mmHg) como variable respuesta medida en momentos diferentes para cada individuo, considerando el instante 0 como el instante base de la medida. La raza del individuo (RIDRETH3) considerando los siguientes valores: 1-Mexicanos americanos; 2-Hispánicos; 3-No hispánicos blancos; 4-No hispánicos negros; 5-No hispánicos multirraciales; 6-No hispánicos asiáticos. La variable sexo (RIAGENDR) de los participantes: 1-Hombres; 2-Mujeres. El análisis de los modelos se ha realizado con el programa R versión 4.3.1. El código utilizado se muestra en el anexo 6.2.

1-Estudio de la presión diastólica según la raza

El objetivo es determinar si hay diferencias significativas en la presión diastólica entre seis grupos de individuos con diferentes orígenes. Nuestra variable de interés, representada como $\mathbf{Y}$, corresponde a la presión, mientras que las distintas razas son los niveles de un un factor fijo. Por tanto, analizaremos este problema mediante una análisis anova con un factor fijo.
Previo a la realización del ANOVA, verificamos si se cumplen los supuestos del modelo. Primero, realizamos la prueba de Shapiro-Wilk para evaluar la normalidad de las observaciones $\mathbf{Y}$ en las diversas razas. Dado que los p-valores son todos mayores que el nivel de significación ($\alpha = 0.05$), concluimos que las observaciones de cada raza siguen una distribución normal. Luego, efectuamos la prueba de Levene para evaluar la homogeneidad de las varianzas entre los grupos. Una vez más, el p-valor (0.7387) supera el nivel de significación α , lo que nos lleva a concluir que las variabilidades entre los grupos son comparables y homogéneas.
Una vez confirmada la adecuación de los supuestos, es prudente construir un gráfico de cajas (Fig 5.2) que permita una visión inicial de los resultados. Además, procederemos a realizar un análisis descriptivo que nos permita profundizar en la comprensión de los datos.

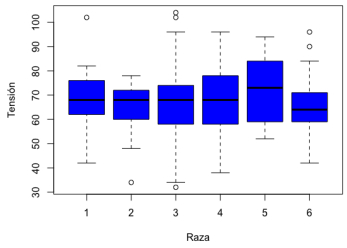

Figura 5.1: Diagrama de cajas.

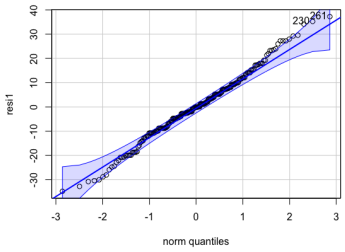

Figura 5.2: Gráfico Q-Q.

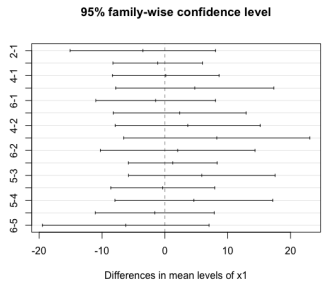

Figura 5.3: Prueba de Tukey.

Grupo	n	Media	Desviación típica	Mediana	Mínimo	Máximo
1	40	67,9	11,85	68	42	102
2	15	64,4	12,19	68	34	78
3	103	66,78	13,77	68	32	104
4	41	68,05	14,15	68	38	96
5	12	72,67	14,53	73	52	94
6	27	66,44	12,12	64	42	96

Cuadro 5.1: Descriptivos.

Una vez disponemos de los estimadores de los parámetros ($\hat{\mu} = 67{,}70$; $\hat{\alpha}_1 = 0{,}19$; $\hat{\alpha}_2 = -3{,}30$; $\hat{\alpha}_3 = -0{,}93$; $\hat{\alpha}_4 = 0{,}34$; $\hat{\alpha}_5 = 4{,}96$), nos planteamos el contraste de hipótesis. Recordemos que la hipótesis nula sostiene que las medias de la presión entre las diferentes razas son idénticas. Por otro lado, la hipótesis alternativa sugiere que al menos una de estas medias es distinta. Para abordar esta cuestión, utilizaremos la prueba F derivada de la tabla ANOVA.

Tipo de grupo	Suma de cuadrados	g.l	Cuadrados Medios	Test F	p-valor
Entre grupos	$SC_E = 557$	5	111.4	0,63	0.677
Dentro de grupos	$SC_D = 41060$	232	177		
Total	$SC_T = 41617$	237			

Cuadro 5.2: Tabla ANOVA de un factor.

Respecto al p-valor asociado al estadístico F, observamos que su valor es (0.667). Dado que este valor es mayor que el nivel de significación α, no disponemos de suficiente evidencia para descartar la hipótesis nula. En otras palabras, nuestras conclusiones apuntan a que no hay diferencias significativas entre las medias de los grupos analizados.

Después de haber realizado las estimaciones con el modelo, es crucial evaluar los residuos, que representan las discrepancias entre los valores observados y los valores pronosticados por el modelo. Los gráficos de residuos suelen mostrar si hay algún problema de independencia entre las observaciones o si la distribución no es normal. Una herramienta efectiva para llevar a cabo esta evaluación es el gráfico Q-Q (Quantile-Quantile) de los residuos (Fig 5.2). En este tipo de representación gráfica, se comparan los cuantiles observados de los residuos con los cuantiles que se esperarían bajo una distribución normal. Como los puntos en el gráfico se disponen de manera aproximada en una línea recta, esto sugiere que los residuos siguen una distribución normal.

Si bien hemos aceptado la hipótesis nula en el análisis ANOVA y hemos concluido que no hay diferencias significativas entre los grupos, aún es importante explorar si hay diferencias sutiles o tendencias en las medias que puedan no haber alcanzado significancia estadística debido a limitaciones en el tamaño de la muestra u otras razones. Para ello vamos a llevar a cabo la prueba de Tukey para calcular interva-

los de confianza al $100(1-\alpha)\%$. Al observar que todos los intervalos contienen el valor cero (Fig 5.3), podemos concluir que las medias entre los grupos son similares y no existen diferencias significativas entre ellos. Esta conclusión refuerza el resultado obtenido al aceptar la hipótesis nula en el análisis ANOVA.

Análisis de la presión diastólica según la raza y el sexo.

El objetivo de este estudio es discernir si existen disparidades significativas en la presión diastólica en seis grupos de individuos con diversas combinaciones de género y raza. La variable de interés **Y** sigue siendo la presión, mientras que el factor **A** se relaciona con la raza y el factor **B** con el género. El modelo adecuado para este problema es un modelo anova con dos factores no balanceado.
Una vez más, comenzaremos nuestro estudio realizando la validación de las suposiciones. En este caso, la normalidad de la variable **Y** será evaluada a través del análisis de los residuos del modelo (Fig 5.5). No obstante, en esta ocasión, vamos a llevar a cabo una vez más la prueba de homogeneidad de varianza. Esta repetición se justifica por la inclusión de un nuevo factor en el análisis. Al observar en la prueba de Levene que el valor de p obtenido (0.2325) es superior al nivel de significación α , podemos considerar las varianzas equiparables y homogéneas.
De manera similar al caso de ANOVA con un factor, es sensato construir un diagrama de cajas (Fig 5.4) que otorgue un panorama inicial de los resultados. Además, realizaremos un análisis descriptivo (véase cuadro 6.2) para profundizar en la comprensión de los datos.

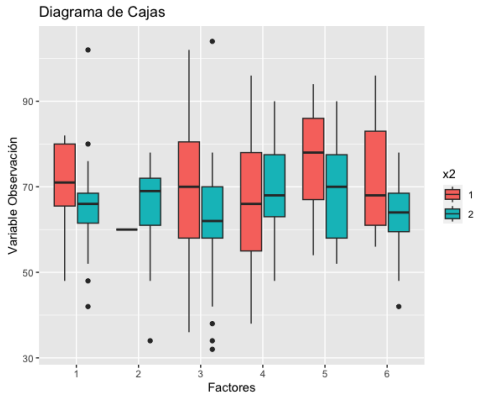

Figura 5.4: Diagrama de cajas.

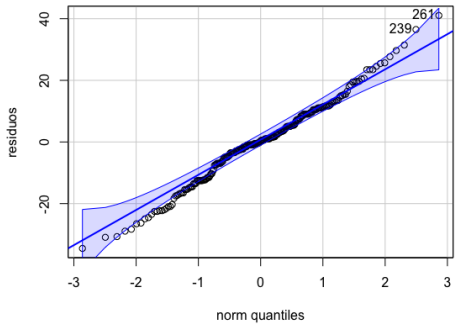

Figura 5.5: Gráfico Q-Q.

Una vez disponemos de los estimadores de los parámetros ($\hat{\mu} = 67{,}47$; $\hat{\alpha}_1 = 0{,}43$; $\hat{\alpha}_2 = -5{,}11$; $\hat{\alpha}_3 = -0{,}73$; $\hat{\alpha}_4 = 0{,}46$; $\hat{\alpha}_5 = 5{,}20$, $\hat{\beta}_1 = 1{,}63$; $(\alpha\beta)_{11} = 0{,}77$; $(\alpha\beta)_{21} = -3{,}99$; $(\alpha\beta)_{31} = 2{,}17$; $(\alpha\beta)_{41} = -3{,}25$; $(\alpha\beta)_{51} = 1{,}70$), nos planteamos los contrastes de hipótesis. En este caso, tenemos tres contrastes a estudiar: H^{AB}, H^{A} y H^{B}. Para abordar estas cuestiones, emplearemos las pruebas F derivadas de la tabla ANOVA de dos factores tipo II y III.

Tipo de grupo	Suma de cuadrados	g.l	Cuadrados Medios	Test F	p-valor
Raza	$SC_A = 428$	5	85,6	0,5013	0.775125
Género	$SC_B = 1404$	1	1404	8,2172	0,004541
Raza:Género	$SC_{AB} = 1039$	5	207,8	1,2158	0.302491
Residuo	$SCR = 38617$	226	170.87		
Total	$SCT = 41488$	237			

Cuadro 5.3: Tabla ANOVA de dos factores tipo II

Tipo de grupo	Suma de cuadrados	g.l	Cuadrados Medios	Test F	p-valor
Raza	$SC_A = 494$	5	98,8	0,5784	0.7166
Género	$SC_B = 213$	1	213	1,2492	0,2649
Raza:Género	$SC_{AB} = 1039$	5	1,2158	1,2158	0.302491
Residuo	$SCR = 38617$	226	170.87		
Total	$SCT = 40363$	237			

Cuadro 5.4: Tabla ANOVA de dos factores tipo III

Como la interacción no aparece, la tabla ANOVA adecuada será la de tipo II, en ella observamos que de los tres tests F realizados, solo uno arroja un valor p asociado (0.0045) que es menor que el nivel de significancción α. Este valor se refiere al factor género **B**, lo que indica que en este caso existe suficiente evidencia para rechazar la hipótesis nula y aceptar la alternativa. En otras palabras, podemos afirmar que los dos niveles de la variable sexo son significativamente diferentes entre sí. Por otro lado, no hay suficiente evidencia para rechazar la hipótesis nula en los tests H^{AB} y H^{A}, lo que sugiere que no hay diferencias significativas entre los niveles de la variable raza y que tampoco existe interacción entre el género y la raza. Una vez que hemos obtenido las estimaciones del modelo, hemos procedido a validar que los errores sigan una distribución normal (Figura 5.5). Por último, aunque la variable género solo tiene dos niveles y, por ende, conocemos las medias entre las que difieren, llevaremos a cabo la prueba de Tukey para calcular intervalos de confianza al$100(1-\alpha)\,\%$. Notamos que únicamente en el intervalo correspondiente a la variable sexo no se incluye el valor cero.

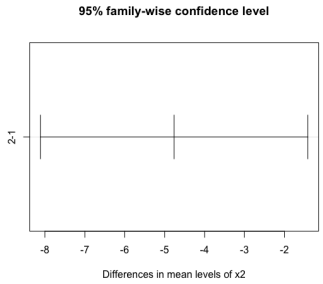

Figura 5.6: Tukey(**A**).

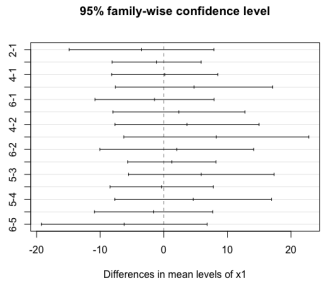

Figura 5.7: Tukey(**B**).

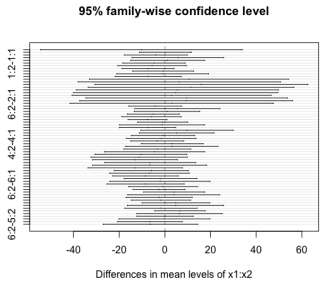

Figura 5.8: Tukey(**AB**).

En resumen, podemos concluir que la presión diastólica presenta diferencias significativas en función del género. Específicamente, se observa una media superior en los hombres. Por otro lado, la raza no ejerce una influencia significativa en la presión diastólica, y tampoco se detecta interacción entre el género y la raza.

Modelo ANOVA de medidas repetidas

El objetivo de este estudio es determinar si existen disparidades significativas en la presión diastólica entre seis grupos de individuos, pero en este caso, cada individuo será sometido a tres mediciones de la presión en momentos temporales consecutivos (BPXDI1, BPXDI2, BPXDI3). Nuestra nueva variable de interés, denotada como **Y**, recoge la presión diastólica en los tres instantes de tiempo. En este contexto, el factor fijo **A** representa la raza de los individuos, y se añade un factor aleatorio **B** que contiene el número de identificación de cada individuo. Dado que cada individuo realiza tres mediciones, su número de identificación aparecerá tres veces en el análisis. Por lo tanto, el modelo adecuado para abordar este problema es un modelo ANOVA aninado, como se ha comentado en el apartado 4.2.2. Este enfoque nos permitirá evaluar las diferencias en la presión diastólica entre grupos, teniendo en cuenta las tres mediciones.
Dado que esta muestra **Y** difiere de la que hemos utilizado anteriormente, es esencial que verifiquemos la suposición de normalidad. La normalidad de **Y** la verificaremos con los residuos del modelo (Fig 5.9). Veamos ahora un diagrama de cajas para ver como han variado las medidas de cada sujeto durante las tres mediciones (Fig 5.10).

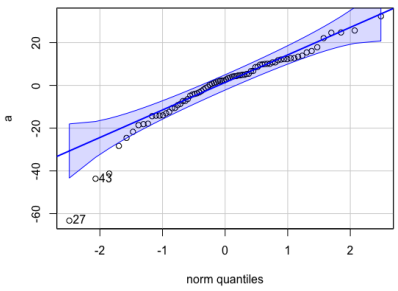

Figura 5.9: Gráfico Q-Q.

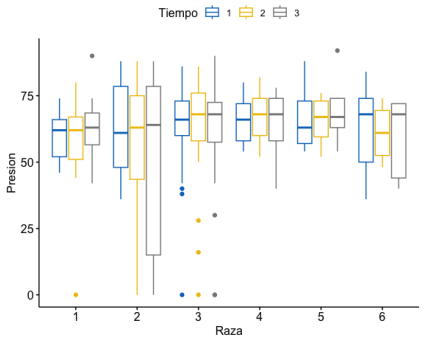

Figura 5.10: Diagrama de cajas.

En este caso, las hipótesis son las mismas que en el caso de un factor fijo. La hipótesis nula plantea que las medias de la presión entre las diferentes razas son idénticas, mientras que la hipótesis alternativa sugiere que al menos una de estas medias es distinta. Para abordar esta cuestión, utilizaremos la prueba F derivada de la tabla ANOVA de factores repetidos.

Tipo de grupo	Suma de cuadrados	g.l	Cuadrados Medios	Test F	p-valor
Raza	$SC_A = 1967$	5	393,4	0,514	0.7650
Residuos	$SC_R = 55117$	72	765,5		

Cuadro 5.5: Tabla ANOVA del factor fijo

Vemos que el p-valor es (0.7650), el cual es mayor que el nivel de significación α establecido. Por lo tanto, aceptamos la hipótesis nula y concluimos que la media de presión en las diferentes razas es la misma.

Capítulo 6

ANEXOS

6.1. Anexo A

Definición 6.1. Definiciones de algunas distribuciones de probabilidad comunes:

1. Si X_i son variables aleatorias independientes con distribución $N(0,1) \quad \forall i = 1,\ 2, \cdots,\ n$, entonces

$$Q = \sum_{i}^{n} X_i^2 \sim \chi_n^2$$

2. Si $U \sim N(0,1)$ y $V \sim \chi_n^2$ donde U y V son independientes, entonces

$$Y = \frac{U}{\sqrt{V/n}} \sim t_n$$

3. Si $A \sim \chi_a^2$ y $B \sim \chi_b^2$ donde A y B son independientes, entonces

$$F = \frac{A/a}{B/b} \sim F_{a,b}$$

Estas relaciones definen estas tres distribuciones.

Demostración. Vamos a demostrar la estimación y las propiedades de los parámetros $\hat{\beta}$.

1. Si desarrollamos la suma de cuadrados tenemos

   $\varepsilon'\varepsilon = \mathbf{Y}'\mathbf{Y} - 2\hat{\beta}\mathbf{X}'\mathbf{Y} + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta}$
   $\frac{\partial \varepsilon'\varepsilon}{\partial \beta} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\hat{\beta}$ y igualando a cero obtenemos la expresión $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

2. $E(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta = \beta$.

3. $\text{var}(\hat{\beta}) = \text{var}((\mathbf{X'X})^{-1}\mathbf{X'Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{var}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

□

Ejemplo de estimación de los parámetros β

Recordemos que la estimación de parámetros β requiere que $(\mathbf{X}'\mathbf{X})$ sea invertible, es decir que rango$\mathbf{X} = m$. ¿Qué sucede si $n < m$?. Entonces tenemos más columnas que filas, por lo que las columnas no pueden ser linealmente independientes , por lo tanto, rango$\mathbf{X} < m$ y $(\mathbf{X}'\mathbf{X})$ no es invertible. También podemos encontrar que $(\mathbf{X}'\mathbf{X})^{-1}$ no existe cuando $n > m$ si algunas de las columnas de $\mathbf{X}$ son

combinaciones lineales. Una solución a este problema es utilizar la regresión de crestas. La regresión de cresta es como la regresión de mínimos cuadrados ordinarios, excepto que agregamos un término de penalización para restringir el tamaño del parámetro. nuestro objetivo es minimizar $\mathbf{S}(\beta)$ con respecto a β.$\mathbf{S}(\beta) = (\mathbf{Y}-\mathbf{X}\beta)'(\mathbf{Y}-\mathbf{X}\beta)+\lambda\beta'\beta = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - \cdots - x_{im}\beta_m)^2 + \lambda\sum_{i=1}^{p}\beta_i^2$.
El segundo término $\lambda\beta'\beta$ es un término de penalización que penaliza los valores de β que son grandes. El parámetro λ es un parámetro de complejidad que controla la fuerza con la que se castigan los valores grandes de β. Consideramos $\lambda > 0$. El estimador de regresión de cresta, denominado $\hat{\beta}_C$, es el valor de β que minimiza $\mathbf{S}(\beta)$. Note que si λ es cero, esto es equivalente a la regresión de mínimos cuadrados y encontramos $\hat{\beta}_C = \hat{\beta}$. A medida que crece λ, crece la penalización por valores grandes de β. Cuando $\lambda \to \infty$ la solución óptima es tomar $\hat{\beta}_C = 0$. Podemos ver que el efecto del término de penalización es reducir las estimaciones de los parámetros hacia 0. Agregar un término de penalización a la suma de los cuadrados se llama regularización y es un enfoque muy poderoso para encontrar buenas estimaciones de parámetros en modelos sobreparametrizados.
La estimación que minimiza $\mathbf{S}(\beta)$ es $\hat{\beta}_C = (\mathbf{X}'\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{Y}$.
La demostración es similar a la $\hat{\beta}$. Vamos a ver el código en R-studio:
install.packages("glmnet") Instalamos los paquetes que vamos a usar.
library(glmnet)
n=150 Número de datos.
m=100 Número de variables explicativas.
mtrue=10 Número de variables que tienen algún efecto.
x=matrix(rnorm(n*m),n,m) matriz de diseño.
beta=rnorm(mtrue) Generamos valores aleatorios de los parámetros.
y= x[,1:mtrue] * beta + rnorm(n,0,5) Generamos las observaciones.
y= cresta= glmnet(x, y, alpha=0) Regresión de crestas.
plot(ridge, xvar="lambda") Gráfica de la regresión de crestas (Figura 6.1).

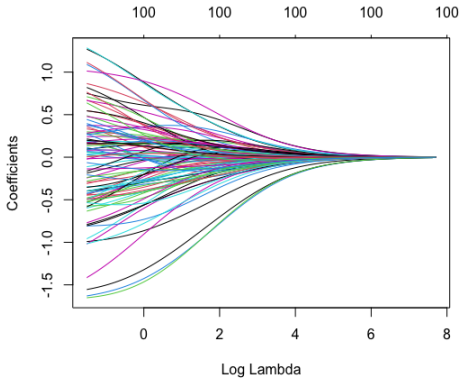

Figura 6.1: Valores de los estimadores $\hat{\beta}_C$ según λ crece.

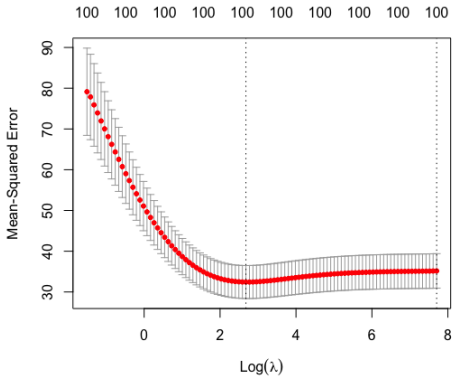

Figura 6.2: Predicciones de los errores según el valor de λ

.

Como ya sabíamos si $\lambda \to \infty$ los estimadores $\hat{\beta}_C \to 0$.
cvcresta=cv.glmnet(x,y, alpha=0) Validación de errores para encontrar el mejor valor de lambda.
plot(cvcresta)
cvcresta/lambda.min El valor de λ que minimiza el error es (14.59) (Figura 6.2).
En la representación gráfica de los errores, utilizamos el logaritmo de λ para una mejor apreciación visual. Al observar los resultados, notamos que el modelo exhibe los errores mas grandes en torno a valores negativos o cercanos a cero. Lo que se traduce en valores de λ próximos a cero. Cabe recordar que cuando el valor de λ era cero, el estimador del modelo $\hat{\beta}_C$ coincidía con el estimador del modelo de

mínimos cuadrados $\hat{\beta}$. Disponemos pues de evidencia que respalda que la estimación de parámetros del modelo de crestas es superior a la estimación de mínimos cuadrados.

Demostración. del lema 1.1

1. Transformaciones lineales de normales multivariantes.
2. Sea $\mathbf{Y} = \Sigma^{-1/2}(\mathbf{X}-\Theta)'$, entonces Y $\sim N_n(0,1)$, donde $\Sigma^{-1/2}$ es la matriz cuadrada de Σ^{-1}
   $\mathbf{Y}'\mathbf{Y} = (\mathbf{X}-\Theta)'\Sigma^{-1}(\mathbf{X}-\Theta) = \sum_{i=1}^{n} Y_i^2 \sim \sum_{i=1}^{n} N(0,1)^2 \sim \chi_n^2$.
3. Es evidente por los dos apartados anteriores.
4. Partiendo de la expresión $(\mathbf{Y}-\Theta)'\mathbf{A}(\mathbf{Y}-\Theta) = \mathbf{Y}'\mathbf{A}\mathbf{Y} - 2\Theta'\mathbf{A}\mathbf{Y} + \Theta'\mathbf{A}\Theta$, tenemos que
   $E(\mathbf{Y}'\mathbf{A}\mathbf{Y}) = E[\sum_{ij}(Y_i-\theta_i)a_{ij}(Y_j-\theta_j)] + 2E[\Theta'\mathbf{A}\mathbf{Y}] - \Theta'\mathbf{A}\Theta = (\sum_{ij} a_{ij}E[(Y_i-\theta_i)(Y_j-\theta_j)]) + \Theta'\mathbf{A}\Theta =$
   $\sum_{ij} a_{ij}\sigma_{ji} + \Theta'\mathbf{A}\Theta = tr(\mathbf{A}\Sigma) + \Theta'\mathbf{A}\Theta$.
5. $\text{cov}(\mathbf{A}\mathbf{Y}, \mathbf{B}\mathbf{Y}) = \mathbf{A}\text{var}(\mathbf{Y})\mathbf{B}' = 0$.

□

Definición 6.2. Tomando $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\beta} = \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y}$, definimos

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}',$$

donde $\mathbf{P}$ es la proyección ortogonal en $\Omega = \langle \mathbf{X} \rangle$. Verifica las siguientes propiedades.

1. $\mathbf{P}=\mathbf{P}', \mathbf{P}^2=\mathbf{P}$ es decir $\mathbf{P}$ es simétrica e idempotente.
2. $(\mathbf{I}_n - \mathbf{P})$ también es simétrica e idempotente.
3. $(\mathbf{I}_n - \mathbf{P})$ tiene $n-m$ valores propios con valor 1 y m con valor cero.
4. $tr(\mathbf{I}_n - \mathbf{P}) = \text{rango}(\mathbf{I}_n - \mathbf{P}) = n-m$.

Demostración. Recordemos que estamos en el caso en el que el rango de la matriz de diseño $\mathbf{X} = m$, y por tanto $\mathbf{P}$ tiene el mismo rango.

1. Inmediato por la definición de $\mathbf{P}$.
2. $(\mathbf{I}_n - \mathbf{P})^2 = \mathbf{I}_n - 2\mathbf{P} + \mathbf{P}^2 = \mathbf{I}_n - 2\mathbf{P} + \mathbf{P} = \mathbf{I}_n - \mathbf{P}$.
3. Es inmediato a partir de que
   $\mathbf{P}^2 = \mathbf{P}$, entonces $\mathbf{P}\mathbf{x} = \lambda\mathbf{x}$ con $\lambda \neq 0$, lo que implica que

   $$\lambda\mathbf{x} = \mathbf{P}\mathbf{x} = \mathbf{P}^2\mathbf{x} = \mathbf{P}(\mathbf{P}\mathbf{x}) = \mathbf{P}(\lambda\mathbf{x}) = \lambda(\lambda\mathbf{x}) = \lambda^2\mathbf{x},$$

   de manera que $\lambda^2 - \lambda = 0$. Luego los valores propios de $\mathbf{P}$ son la unidad tantas veces como indica el rango y el resto son ceros, ya que la suma de valores propios es el rango.
4. Por el apartado anterior, $\mathbf{I}_n - \mathbf{P}$ tiene $n-m$ valores propios 1 y el resto son cero. De aquí que $\text{tr}(\mathbf{I}_n - \mathbf{P}) = n-m$ ya que en matrices idempotantes la traza y el rango coinciden.

□

Demostración. Del lema 1.2 Vamos a utilizar los multiplicadores de Lagrange $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_m)'$, nuestro objetivo es minimizar $\mathbf{G}(\beta)$ con respecto a β

$$\mathbf{G}(\beta) = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) + (\beta'\mathbf{A}' - c')\lambda$$

$$\frac{\partial \mathbf{G}(\beta)}{\partial \beta} = -2\mathbf{X}'\mathbf{Y} + 2(\mathbf{X}'\mathbf{X})\beta + \mathbf{A}'\lambda.$$

En el punto $\hat{\beta}_H = \beta$ y igualando a cero.

$$\hat{\beta}_H = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} - 1/2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\lambda = \hat{\beta} - 1/2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\lambda.$$

Sin embargo $\mathbf{A}\hat{\beta}_H = c$ así que

$$c = \mathbf{A}\hat{\beta} - 1/2\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\lambda \rightarrow 1/2\lambda = [\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'](c - \mathbf{A}\hat{\beta}).$$

Sustituyendo λ en la ecuación de $\hat{\beta}_H$ completamos la demostración del lema.

□

Fuente	Suma de cuadrados	g.l
A	$\sum_{i,j,k,m}(y_{i...} - \bar{y})^2 = bcr\sum_i(y_{i..} - \bar{y})^2$	$a-1$
B	$\sum_{i,j,k,m}(y_{.j..} - \bar{y})^2$	$b-1$
C	$\sum_{i,j,k,m}(y_{..k.} - \bar{y})^2$	$c-1$
D	$\sum_{i,j,k,m}(y_{...m} - \bar{y})^2$	$d-1$
AB	$\sum_{i,j,k,m}(y_{ij..} - y_{i...} - y_{.j..} + \bar{y})^2$	$(a-1)(b-1)$
AC	$\sum_{i,j,k,m}(y_{i.k.} - y_{i...} - y_{..k.} + \bar{y})^2$	$(a-1)(c-1)$
AB	$\sum_{i,j,k,m}(y_{i..m} - y_{i...} - y_{...m} + \bar{y})^2$	$(a-1)(d-1)$
AB	$\sum_{i,j,k,m}(y_{.jk.} - y_{.j..} - y_{..k.} + \bar{y})^2$	$(b-1)(c-1)$
AB	$\sum_{i,j,k,m}(y_{.j.m} - y_{.j..} - y_{...m} + \bar{y})^2$	$(b-1)(d-1)$
AB	$\sum_{i,j,k,m}(y_{..km} - y_{..k.} - y_{..m.} + \bar{y})^2$	$(c-1)(d-1)$
ABC	$\sum_{i,j,k,m}(y_{ijk.} - y_{ij..} - y_{i.k.} - y_{.jk.} + y_{i...} + y_{.j..} + y_{..k.} - \bar{y})^2$	$(a-1)(b-1)(c-1)$
ABD	$\sum_{i,j,k,m}(y_{ij.m} - y_{ij..} - y_{i..m} - y_{.j.m} + y_{i...} + y_{.j..} + y_{...m} - \bar{y})^2$	$(a-1)(b-1)(d-1)$
ACD	$\sum_{i,j,k,m}(y_{i.km} - y_{i.k.} - y_{i..m} - y_{..km} + y_{i...} + y_{..k.} + y_{...m} - \bar{y})^2$	$(a-1)(c-1)(d-1)$
BCD	$\sum_{i,j,k,m}(y_{.jkm} - y_{.jk.} - y_{.j.m} - y_{..km} + y_{.j..} + y_{..k.} + y_{...m} - \bar{y})^2$	$(b-1)(c-1)(d-1)$
ABCD	$\sum_{i,j,k,m}(y_{ijkm} - y_{ijk.} - y_{ij.m} - y_{i.km} - y_{.jkm} + y_{ij..} + y_{i.k.} + y_{i..m} + y_{.jk.} + y_{.j.m} + y_{..km} - y_{i...} - y_{.j..} - y_{..k.} - y_{...m} + \bar{y})^2$	$(a-1)(b-1)(c-1)(d-1)$
SCT	$\sum_{ijkm}(y_{ijkm} - \bar{y})^2$	$abcd-1$

Cuadro 6.1: Descomposición ortogonal de la suma de cuadrados correspondiente a un diseño de cuatro factores.

Σ matriz de varianzas-covarianzas de $\mathbf{Y}$, en el modelo ANOVA de un factor aleatorio.

$$\Sigma = \begin{pmatrix} c & 0 & \cdots & 0 \\ 0 & c & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c \end{pmatrix} \qquad \text{con} \qquad c = \begin{pmatrix} \sigma_y^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_y^2 & \cdots & \sigma_\alpha^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \cdots & \sigma_y^2 \end{pmatrix}$$

6.2. Anexo B

Modelo ANOVA de un factor

install.packages("psych") install.packages(çar") Instalamos los paquetes que vamos a usar
library(psych) library(car)
y<- z/BPXDI1; x1<-z/RAZA Nombramos nuestras variables.
hist(y) Histograma de la variable **Y**.
by(y,x1,shapiro.test) Verificamos la normalidad por grupos de la variable **Y**.
leveneTest(y,x1) Homogeneidad de varianzas.
describeBy(y,x1, IQR=T) Descriptivos.
modeloanova <- aov(y x1); tablaanova <- summary(modeloanova); print(tablaanova) Tabla Anova.
modeloanova/coefficients Ver los estimadores del modelo.
resi1<-residuals(modeloanova); qqPlot(resi1) Residuos.
comparaciones <- TukeyHSD(modeloanova); print(comparaciones); plot(comparaciones) Prueba de Tukey.

Modelo ANOVA de dos factores

install.packages("psych"); install.packages(çar");install.packages(reshape");
install.packages("multcomp");install.packages("ggplot2");install.packages("pastecs")
Instalamos los paquetes que vamos a usar.
library(psych); library(car);library(reshape);library(multcomp);library(ggplot2);library(pastecs)
x2<-z/SEXO Nombramos nuestra nueva variable.
leveneTest(y,interaction(x1,x2), center=median) Homogeneidad de varianzas.
ggplot(z, aes(x = x1, y = y, fill = x2)) + geomboxplot() + labs(title = "Diagrama de Cajas", x = "Factores", y = "Variable Observación") Diagrama de cajas.
describeBy(y,group=interaction(x1,x2), IQR=T) Descriptivos 6.2.
modeloanova2 <- Anova(lm(y x1 + x2 + x1:x2), type=ÏI"); Tabla ANOVA de dos factores tipo II.
modeloanova3 <- Anova(lm(y x1 + x2 + x1:x2), type=ÏII"); Tabla ANOVA de dos factores tipo III.
summary(lm(y x1 + x2 + x1:x2)) Ver las estimaciones de los parámetros del modelo.
resi2<-residuals(modeloanova2); qqPlot(resi2) Gráfico Q-Q de los errores.
comparaciones2 <- TukeyHSD(modeloanova2); print(comparaciones2); plot(comparaciones2) Intervalos Turkey de confianza de $100(1-\alpha)\,\%$.

Grupo	Género	n	Media	Desviación típica	Mediana	Mínimo	Máximo
1	1	20	70,3	10,79	71	48	82
2	1	1	60	–	60	60	60
3	1	52	70,54	14,69	70	36	102
4	1	19	66,32	17,83	66	38	96
5	1	6	76	14,91	78	54	94
6	1	11	71,45	14,17	68	56	96
1	2	20	65,5	12,65	66	42	102
2	2	14	64,71	12,59	68	34	78
3	2	51	62,94	11,71	62	32	104
4	2	22	69,55	10,2	68	48	90
5	2	6	69,33	14,68	70	52	90
6	2	16	63	9,47	64	42	78

Cuadro 6.2: Descriptivos.

Modelo ANOVA de medidas repetidas

install.packages("psych"); install.packages(çar");install.packages(reshape");
install.packages("multcomp");install.packages("ggplot2");install.packages("pastecs");install.packages("WRS2");install.p
Instalamos los paquetes que vamos a usar
library(psych); library(car);library(reshape);library(multcomp);library(ggplot2);library(pastecs);library(WRS2);library(a
presion<-c(z/BPXDI1, z/BPXDI2, z/BPXDI3); raza<-c(z/RIAGENDR, z/RIAGENDR, z/RIAGENDR);
número<-c(z/NUMERO, z/NUMERO, z/NUMERO) Nombramos nuestras variables.
datos <- data.frame(Presión = presión, Raza = raza, Número = número) Creamamos un data frame con nuestras nuevas variables.
ggboxplot(datos, x= "Número", y= "Presión", color= Raza", palette="jco") Diagrama de cajas
modelo2<- aov (Presión Raza + Error (Número)) summary(modelo2) Modelo ANOVA de medidas repetidas.

Bibliografía

[1] DR. RICHARD WILKINSON, *Linear models*, University of Nottingham, Reino Unido.

[2] MARC.S PAOLELLA, *Linear Models and Time-Series Analysis,Regression, ANOVA, ARMA and GARCH*, University of Zurich, Switzerland (1991).

[3] J.D. JOBSON, *Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design*, Faculty of Business, University of Alberta, Canada (1991).

[4] F. CARMONA, *Modelos Lineales*, Departament d'Estadística, Universitat de Barcelona, España (2003).

[5] J. MONTANERO, *Modelos Lineales*, Universidad de Extremadura, España (2008).

[6] H. SCHEFFÉ, *The analysis of variance*, Wiley, New York (1959).

[7] H. SAHAI, M M OJEDA, *Analysis of Variance for Ramdom Models, Volume II: Unbalanced Data*, United States of America, (1959).

[8] D. PEÑA, *Estadística, Modelos y métodos, Volume II: Modelos lineales y series temporales*, Alianza Universidad Textos, (1989).

[9] H. O. HARTLEY, *Analysis of Variance. Mathematical Methods for Digital Computers*, (1962).

[10] G. A. F. SEBER, *Linear Regression Analysis*, Wiley, Nueva York (1962).