Índice general

Tema 1

Análisis de la varianza: MANOVA

1.1. El modelo

El análisis multivariante de la varianza (MANOVA) es una generalización a $p > 1$ del análisis de la varianza (ANOVA). Supongamos que tenemos n observaciones independientes de p variables $Y_1, \ldots, Y_p$ obtenidas de forma experimental. La matriz de datos por tanto será:

$$\begin{pmatrix} y_{11} & y_{21} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix} = [\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_p]$$

donde $\tilde{y}_j = (y_{1j}, y_{2j}, \ldots, y_{pj})'$ son las n observaciones, independientes, de la variable Y_j, que se supondremos siguen un modelo lineal univariante $\tilde{y}_j = X\beta_j + e_j$.

El modelo lineal multivariante será de la forma

$$Y = XB + E$$

donde X es la matriz de diseño:

$$X = \begin{pmatrix} x_{11} & x_{21} & \ldots & x_{1m} \\ x_{21} & x_{22} & \ldots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nm} \end{pmatrix}$$

B es la matriz de parámetros

$$B = \begin{pmatrix} \beta_{11} & \beta_{21} & \ldots & \beta_{1p} \\ \beta_{21} & \beta_{22} & \ldots & \beta_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{m1} & \beta_{m1} & \ldots & \beta_{mp} \end{pmatrix}$$

y E es la matriz de desviaciones aleatorias:

$$E = \begin{pmatrix} e_{11} & e_{21} & \dots & e_{1p} \\ e_{21} & e_{22} & \dots & e_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ e_{n1} & e_{n1} & \dots & e_{np} \end{pmatrix}$$

Las matrices Y y X son conocidas y las filas de E son independientes $\mathcal{N}_p(0, \Sigma)$.

1.2. Estimación de parámetros

En el modelo descrito en la sección anterior tenemos $m \times p$ de la matriz parámetros de regresión B a estimar así como la matriz de covarianzas Σ. En este caso se puede demostrar que el estimador de mínimos cuadrados de B es $\hat{B}$ tal que minimiza la traza:

$$Tr\left(\hat{E}'\hat{E}\right) = Tr[(Y - X\hat{B})'(Y - X\hat{B})],$$

siendo $\hat{E} = Y - X\hat{B}$.

La matriz de residuos es la matriz $\mathbf{R}_0 = (R_0(i,j))$ de orden $p \times p$,

$$\mathbf{R}_0 = \hat{E}'\hat{E} = (Y - X\hat{B})'(Y - X\hat{B})$$

donde $R_0(i,j)$ es la suma de cuadrados residuales del modelo univariante $\tilde{y}_j = X\beta_j + e_j$.

Teorema 1.1. *Consideremos el modelo de regresión multivariante $Y = XB + E$, siendo*

$$Y = \begin{pmatrix} y_1' \\ \vdots \\ y_n' \end{pmatrix}, E = \begin{pmatrix} e_1' \\ \vdots \\ e_n' \end{pmatrix}$$

verificando

- $E[Y] = XB$, es decir, $E[E] = 0$
- $cov(y_i) = cov(e_i) = \Sigma$, donde y_i' son las filas de Y y e_i' son las filas de E.
- $cov(y_i, y_j) = cov(e_i, e_j) = 0$, para $i \neq j$.

Entonces las estimaciones máximo verosímiles de los parámetros de regresión de B verifican las ecuaciones normales

$$X'X\hat{B} = X'Y$$

y son

$$\hat{B} = (X'X)^{-1}X'Y,$$

cuando el diseño es de rango máximo $r = rango(X) = m$ y por

$$\hat{B} = (X'X)^{-}X'Y,$$

cuando $r < m$. El estimador $\hat{B}$ minimiza la traza $tr(\hat{E}'\hat{E})$ asi como el determinante $det(\hat{E}'\hat{E})$. También $\hat{B}$ es un estimador insesgado de B.

Demostración. Sea B_0 otro estimador de B, entonces

$$(Y - XB_0)'(Y - XB_0) = (Y - X\hat{B} + X\hat{B} - XB_0)'(Y - X\hat{B} + X\hat{B} - XB_0) =$$

$$R_0 + (X\hat{B} - XB_0)'(X\hat{B} - XB_0) + (Y - X\hat{B})'(X\hat{B} - XB_0) + (X\hat{B} - XB_0)'(Y - X\hat{B}) =$$

$$R_0 + (X\hat{B} - XB_0)'(X\hat{B} - XB_0),$$

puesto que $(Y - X\hat{B})'(X\hat{B} - XB_0) = (Y - X\hat{B})'X(\hat{B} - B_0) = 0$, al verificar $\hat{B}$ las ecuaciones normales. Por lo que

$$(Y - X\hat{B}_0)'(Y - X\hat{B}_0) = R_0 + M$$

con M una matriz $p \times p$ definida positiva. Por lo que la traza y el determinante de $(Y - XB_0)'(Y - XB_0)$ alanzarán el mínimo cuando $M = 0$, es decir para $B_0 = \hat{B}$. Por otro lado

$$E[\hat{B}] = (X'X)^{-1}X'E[Y] = (X'X)^{-1}(X'X)B = B$$

□

Teorema 1.2. *En las mismas condiciones que el teorema anterior, con $r = rango(X)$, se puede expresar la matriz de residuos como*

$$R_0 = Y'[I - X(X'X)^{-}X']Y$$

Una estimación centrada de la matriz de varianzas covarianzas Σ es

$$\hat{\Sigma} = \frac{R_0}{n - r}$$

Demostración. Sea

$$(Y - X\hat{B})'(Y - X\hat{B}) = Y'Y - Y'X\hat{B} - \hat{B}'X'Y + \hat{B}'X'X\hat{B}$$

al ser $\hat{B}'X'Y = \hat{B}'X'X\hat{B}$, resulta que

$$(Y - X\hat{B})'(Y - X\hat{B}) = Y'Y - Y'X\hat{B} = Y'Y - Y'X(X'X)^{-}X'Y = Y'[I - X(X'X)^{-}X')]Y$$

Tomando $T = [t_1 \ldots, t_r, t_{r+1}, \ldots, t_n]$ una matriz ortogonal tal que sus columnas forma una base ortonormal de R^n, de manera que las primeras r generen el mismo espacio $C_r(X)$ generado por las columnas de X. Por lo que el resto de las otras $n - r$ columnas serán ortogonales a $C_r(X)$, es decir,

$$t_i'X = * \quad \text{si } i \leq r$$

$$t_i'X = 0 \quad \text{si } i > r$$

donde * es un valor posiblemente no nulo.

Sea $Z = T'Y$, entonces

$$E[Z] = T'XB = \begin{bmatrix} \eta \\ 0 \end{bmatrix} \quad \begin{matrix} \text{r primeras filas} \\ \text{n-r primeras filas} \end{matrix}$$

Consideremos el residuo $\hat{E} = Y - X\hat{B}$. Al ser $X'(Y - X\hat{B}) = 0$, se deduce que $\hat{E}$ es ortogonal a X en el sentido de

$$T'\hat{E} = \begin{bmatrix} 0 \\ Z_{n-r} \end{bmatrix} \quad \begin{matrix} \text{r primeras filas} \\ \text{n-r primeras filas} \end{matrix}$$

donde Z_{n-r} es la matriz de orden $(n - r) \times p)$.

Pero

$$T'\hat{E} = T'Y - T'X\hat{B} = Z - \begin{bmatrix} * \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ Z_{n-r} \end{bmatrix},$$

es decir, las últimas n-r filas de Z y de $T'\hat{E}$ coinciden. Entonces, como $TT' = I$

$$R_0 = \hat{E}'\hat{E} = \hat{E}'TT'\hat{E} = [0, Z_{n-r}'] \begin{bmatrix} 0 \\ Z_{n-r} \end{bmatrix} = Z_{n-r}'Z_{n-r}$$

Con $Z_{n-r}' = [z_1, \ldots, z_{n-r}]$ donde $z_1', \ldots, z_{n-r}'$ son las filas independientes de Z_{n-r}. Entonces cada z_i es un vector de media cero y matriz de covarianzas Σ y por tanto $E[z_i, z_i'] = \Sigma$ y $Z_{n-r}'Z_{n-r} = z_1z_1' + \ldots + z_{n-r}z_{n-r}'$; por lo que

$$E[R_0] = E[z_1z_1' + \ldots + z_{n-r}z_{n-r}'] = (n - r)\Sigma.$$

□

Teorema 1.3. *Sea $Y = XB + E$ el modelo lineal normal multivariante donde las filas de E son $\mathcal{N}_p(0, \Sigma)$ independientes y sea R_0 la matriz de residuos; se verifica que la distribución de R_0 es una Wishart $\mathcal{W}_p(\Sigma, n - r)$.*

Demostración. Por el teorema 1.2; $E[Z_{n-r}] = 0$, por lo que las filas de Z_{n-r} son $\mathcal{N}_p(0, \Sigma)$ independientes, por lo que $R_0 = Z_{n-r}'Z_{n-r}$ es una matriz $p \times p$ que sigue un distribución de Wishart □

1.3. Contrastes de hipótesis lineales

Sea una matriz H de rango t, tal que

$$H_0 : HB = 0$$

donde cada fila de H es una combinación lineal de las filas de X.

Como en el caso univariantes, bajo H_0 tendremos

$$Y = \tilde{X}\Theta + E$$

donde los parámetros de B restringidos a H_0 viene dado por la expresión

$$\hat{B}_H = \hat{B} - (X'X)^- H'(H(X'X)^- H')^{-1} H\hat{B}$$

y la matriz de residuos

$$R_1 = (Y - X\hat{B}_H)'(Y - X\hat{B}_H)$$

Teorema 1.4. *Sea $Y = XB + E$ el modelo lineal multivariante, donde las filas de E son $\mathcal{N}_p(0, \Sigma)$ independientes, R_0 la matriz de residuos, $H_0 : HB = 0$ una hipótesis lineal demostrable y R_1 la matriz de residuos bajo H_0, se verifica:*

1. *$R_0 \rightsquigarrow W_p(\Sigma, n - r)$*
2. *Si H_0 es cierta, la matrices R_1 y $R_1 - R_0$ siguen la distribución de Wishart*

$$R_1 \rightsquigarrow W_p(\Sigma, n - r') \quad R_1 - R_0 \rightsquigarrow W_p(\Sigma, t)$$

   con $t = rango(H)$, y $r' = r - t$.

3. *Si H_0 es cierto, las matrices R_0 y $R_1 - R_0$ son estocásticamente independientes.*

Demostración. Bajo H_0, el subespacio generado por las filas de H está contenido en el generado por las filas de X, por lo que podremos construis una base ortogonal de R^m

$$[u_1, \ldots, u_t, u_{t+1}, \ldots, u_r, u_{r+1}, \ldots, u_m]$$

tal que $[u_1, \ldots, u_t]$ generen H; $[u_1, \ldots, u_t, u_{t+1}, \ldots, u_r]$ generen X.

Consideremos entonces C de orden $m \times (r - t)$ generada por $[u_{t+1}, \ldots, u_r]$.Entonces $HC = 0$ y el modelo $Y = XB + E$ se convierte en $Y = \tilde{X}\Theta + E$, siendo $\tilde{X} = XC$ y $C\Theta = B$, pues $HB = HC\Theta = 0$. Así la matriz de diseño X se transforma en $\tilde{X} = XC$, donde las columnas de XC son una combinación lineal de las columnas de X.

Podemos construir una matriz ortogonal

$$T = [t_1 \ldots, t_{r'}, t_{r'+1}, \ldots, t_r, t_{r+1}, \ldots, t_n]$$

tal que las $r' = r - t$ primeras columnas generen XC y las r primeras generen X

$$C_{r'}(XC) = [t_1, \ldots, t_{r'}] \subset C_r(X) = [t_1, \ldots, t_r].$$

Siguiendo los mismo pasos que en el teorema 1.2, tenemos que

$$T'\hat{E} = \begin{bmatrix} 0 \\ Z_{n-r'} \end{bmatrix}$$

donde las $n - r'$ filas de $Z_{n-r'}$ son las $\mathcal{N}_p(0, \Sigma)$ independientes, por lo tanto

$$R_1 = (Y - \tilde{X}\hat{\Theta})'(Y - \tilde{X}\hat{\Theta}) = Z'_{n-r'}Z_{n-r'}$$

es Wishart $W_p(\Sigma, n - r')$.

También se puede escribir:

$$T'(Y - \tilde{X}\hat{\Theta}) = \begin{bmatrix} 0 \\ Z_{n-r'} \end{bmatrix} = \begin{bmatrix} 0 \\ Z_t \\ Z_{n-r} \end{bmatrix}$$

donde las $t = r - r'$ filas de Z_t son independientes de las $n - r$ filas de Z_{n-r}. Entonces $R_1 = Z'_tZ_t + Z'_{n-r}Z_{n-r}$, es decir:

$$R_1 - R_0 = Z'_tZ_t$$

donde $R_1 - R_0$ sigue una Wishart $W_p(\Sigma, n - r')$ e independiente de R_0.

□

Bajo H_0, se verifica que R_0 y $R_1 - R_0$ son Wishart independientes y

$$\Lambda = \frac{|R_0|}{|(R_1 - R_0) + R_0|} = \frac{|R_0|}{|R_1|} \rightsquigarrow \Lambda(p, n - r, t),$$

con $0 \leq \Lambda \leq 1$ y con una distribución de Wilks.

Por tanto se acepta H_0 cuando Λ no es significativo y rechazaremos H_0 para Λ pequeño y significativo.

De todo esto resulta la siguiente tabla ANOVA

	g.l.	matriz de Wishart	lambda de Wilks
Desviación hipótesis	t	$R_1 - R_0$	$\Lambda = \frac{\lvert R_0 \rvert}{\lvert R_1 \rvert}$
Residuo	$n - r$	R_0	

Si $\Lambda < \Lambda_\alpha$ se rechaza H_0 con $P(\Lambda(p, n - r, t) < \Lambda_\alpha) = \alpha$.

1.4. MANOVA de un factor

El modelo del diseño con un único factor (causa de variabilidad) es de la forma:

$$y_{ih} = \mu + \alpha_i + e_{ih} \quad i = 1, \ldots, k;\ h = 1, \ldots, n_i$$

donde μ es un vector de medias general, α_i es el efecto del nivel i del factor, y_{ih} es la observación h en la población i, correspondiente a la misma situación experimental del análisis canónico de poblaciones, con $n = n_1 + \ldots + n_k$. La hipótesis nula consiste en afirmar que las α_i observaciones son iguales a cero. Tenemos que

$$W = R_0, \quad B = R_1 - R_0, \quad T = R_1 = B + W$$

son las matrices de dispersión dentro de grupos; entre grupos y total:

	$g.l.$	matriz de Wishart	lambda de Wilks
entre grupos	$k-1$	B	$\Lambda = \dfrac{\lvert W\rvert}{\lvert T\rvert}$
dentro de grupos	$n-k$	W	$\rightsquigarrow \Lambda(p, n-k, k-1)$
Total	$n-1$	T	

1.5. MANOVA de dos factores

Si suponemos que las $n = a \times b$ observaciones multivariantes dependen de los factores fila y columna, con a y b niveles respectivamente, el modelo será:

$$y_{ih} = \mu + \alpha_i + \beta_j + e_{ij} \quad \imath = 1, \ldots, a;\ j = 1, \ldots, b$$

donde μ es la media general, α_i es el efecto aditivo del nivel i del factor fila, β_j es el efecto aditivo del nivel j del factor columna. Como generalización del caso univariante, intervienen las matrices $A = (a_{uv}); B = (b_{uv}); T = (t_{uv})$ y $R_0 = (r_{uv})$ con elementos

$$a_{uv} = b \sum_i (y_{i.u} - \bar{y}_u)(y_{i.v} - \bar{y}_v)$$

$$b_{uv} = a \sum_j (y_{.ju} - \bar{y}_u)(y_{.jv} - \bar{y}_v)$$

$$r_{uv} = \sum_{ij} (y_{iju} - y_{i.u} - y_{.ju} + \bar{y}_u)(y_{iju} - y_{i.v} - y_{.jv} + \bar{y}_v)$$

$$t_{uv} = \sum_{ij} (y_{iju} - \bar{y}_u)(y_{iju} - \bar{y}_v) \quad u, v = 1, \ldots, p.$$

siendo, para cada variable Y_u, $\bar{y}_u$ la media general, $y_{.jv}$ la media fijado el nivel j de cada factor columna...

Se verifica que

$$T = A + B + R_0.$$

Si las α o las β son nulas, entonces $R_1 = R_0 + A$ o $R_1 = R_0 + B$ respectivamente. Por lo que, indicando que $q = (a-1)(b-1)$, para contrastar la hipótesis de que no influye el factor fila o el factor columna, en ninguna de las variables, obtendremos la tabla:

	$g.l.$	matriz de Wishart	lambda de Wilks
filas	$a-1$	A	$\lvert R_0\rvert \,/\, \lvert R_0 + A\rvert \rightsquigarrow \Lambda(p, q, a-1)$
columnas	$b-1$	B	$\lvert R_0\rvert \,/\, \lvert R_0 + B\rvert \rightsquigarrow \Lambda(p, q, b-1)$
residuo	q	R_0	
Total	$ab-1$	T	

1.6. MANOVA de dos factores con interacción

En el diseño de dos factores con interacción suponemos que las $n = a \times b \times c$ observaciones multivariantes dependen de dos factores fila y columna, con a y b niveles, y que hay c observaciones (replicas) para cada una de las $a \times b$ combinaciones de los niveles. El modelo lineal es:

$$y_{ijh} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijh} \quad i = 1, \ldots, a; j = 1, \ldots, b; h = 1, \ldots, c$$

donde μ es la media general, α_i es el efecto aditivo del nivel i del factor fila, β_j es el efecto aditivo del nivel j del factor columna, γ_{ij} es la interacción que mide la desviación de la aditivilidad del efecto de los factores, e $y_{ijh} = (y_{ijh1}, \ldots, y_{ijhp})'$ es la répilca multivariante h de las variables observables. También, al igual que el caso univariante, intervienen las matrices $A = (a_{uv}); B = (b_{uv}); AB = (c_{uv}), T = (t_{uv})$ y $R_0 = (r_{uv})$ donde

$$a_{uv} = bc \sum_i (y_{i..u} - \bar{y}_u)(y_{i..v} - \bar{y}_v)$$

$$b_{uv} = ac \sum_j (y_{.j.u} - \bar{y}_u)(y_{.j.v} - \bar{y}_v)$$

$$c_{uv} = c \sum_{ij} (y_{ij.u} - y_{i..u} - y_{.j.v} + \bar{y}_u)(y_{ij.v} - y_{i..v} - y_{.j.v} + \bar{y}_v)$$

$$r_{uv} = \sum_{ijh} (y_{ijhu} - y_{i..u})(y_{ijhv} - y_{i..v})$$

$$t_{uv} = \sum_{ij} (y_{iju} - \bar{y}_u)(y_{iju} - \bar{y}_v) \quad u, v = 1, \ldots, p.$$

que verifican que:

$$T = A + B + AB + R_0$$

Teniendo que $q = (a-1)(b-1)$, $r = ab(c-1)$, para contrastar las hipótesis de que los factores filas, columna o las interacciones no influyan en ninguna de las variables, tendremos la tabla:

	$g.l.$	matriz de Wishart	lambda de Wilks
filas	$a-1$	A	$\lvert R_0\rvert \,/\, \lvert R_0+A\rvert \rightsquigarrow \Lambda(p,r,a-1)$
columnas	$b-1$	B	$\lvert R_0\rvert \,/\, \lvert R_0+B\rvert \rightsquigarrow \Lambda(p,r,b-1)$
interacción	q	AB	$\lvert R_0\rvert \,/\, \lvert R_0+AB\rvert \rightsquigarrow \Lambda(p,r,q)$
residuo	r	R_0	
Total	$abc-1$	T	

1.7. Otros criterios

Sean $\lambda_1, \geq \lambda_2 \geq \ldots \lambda_p$ los valores propios de R_0 respecto de R_1, es decir las raíces de la ecuación $det(R_0 - \lambda R_1) = 0$. Podemos expresar el criterio de Wilks como

$$\Lambda = \frac{|R_0|}{|R_1|} = \lambda_1 \times \ldots \times \lambda_p$$

Teniendo en cuenta que si λ es la razón de verosimilitud en el test de hipótesis, entonces $\lambda = \Lambda^{n/2}$.

Es fácil ver que si $0 \leq \lambda_i \leq 1$, y se llaman correlaciones canónicas generalizadas (al cuadrado) a $r_i^2 = 1 - \lambda_i$, $i = 1, \ldots, p$, entonces el criterio de Wilks en términos de correlaciones es:

$$\Lambda = \prod_{i=1}^{p}(1 - r_i^2)$$

Se demuestra que cualquier estadístico que sea invariante ante cambios de escala y de origen son función de los valores propios $\lambda_1, \geq \lambda_2 \geq \ldots \lambda_p$, así se pueden proponer los siguientes estadísticos:

- Traza de Hotelling:
$$tr[R_0^{-1}(R_1 - R_0)] = \sum_{i=1}^{p} \frac{1-\lambda_i}{\lambda_i} = \sum_{i=1}^{p} \frac{r_i^2}{1-r_i^2}$$
- Traza de Pillai:
$$tr[R_1^{-1}(R_1 - R_0)] = \sum_{i=1}^{p}(1-\lambda_i) = \sum_{i=1}^{p} r_i^2$$
- Raíz mayor de Roy:
$$\theta = 1 - \lambda_p = r_1^2$$

1.8. Aplicación en R

En esta práctica se utilizan los metodos descritos en el tema actual y los test del tema anterior basados en el estadístico T^2 de hotelling y todos ellos en base del ejemplo de Fisher para los tipos de iris. En primer lugar cargaremos el fichero de datos:

```
attach(iris)
 View(iris)
head(iris)

##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Se pueden obtener algún estadístico descriptivo básico, con los que ya podemos podemos observar las diferencias entre las variables:

```
summary(iris)

##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

Realizaremos en primer lugar un análisis MANOVA para comprobar la igualdad de los grupos según los diferentes tipos de especies:

```
iris.manova <- manova(cbind(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width) ~ Species)
iris.manova

## Call:
##    manova(cbind(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width) ~
```

```
##     Species)
##
## Terms:
##                   Species Residuals
## Sepal.Length      63.2121   38.9562
## Sepal.Width       11.3449   16.9620
## Petal.Length     437.1028   27.2226
## Petal.Width       80.4133    6.1566
## Deg. of Freedom         2       147
##
## Residual standard errors: 0.5147894 0.3396877 0.4303345 0.20465
## Estimated effects may be unbalanced
```

Ya con esta primera salida podemos ver que parece haber hay diferencias entre las variables según el tipo de planta. Si obtenemos los contrastes más comunes, tenemos:

```
summary(iris.manova, test="Wilks")

##            Df    Wilks approx F num Df den Df    Pr(>F)
## Species     2 0.023439   199.15      8    288 < 2.2e-16 ***
## Residuals 147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Haciendo los test de Roy; de Pillay y de Hotelling-Lawley

```
summary(iris.manova, test="Roy")

##            Df    Roy approx F num Df den Df    Pr(>F)
## Species     2 32.192     1167      4    145 < 2.2e-16 ***
## Residuals 147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(iris.manova, test="Hotelling-Lawley")

##            Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
## Species     2           32.477   580.53      8    286 < 2.2e-16 ***
## Residuals 147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(iris.manova, test="Pillai")

##            Df Pillai approx F num Df den Df    Pr(>F)
## Species     2 1.1919   53.466      8    290 < 2.2e-16 ***
## Residuals 147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Por lo que por todos los contrastes vemos que existen diferencias entre los vectores medias de los tres tipos de variedades.

Contrastes marginales de medias

Hemos contrastado la igualdad de medias conjuntas, en este caso haremos los contrastes marginales.

En el caso de la longitud de sépalos se rechaza que marginalmente sean iguales:

```
modelo1<-aov(Sepal.Length~ Species )
summary(modelo1)

##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  63.21  31.606   119.3 <2e-16 ***
## Residuals   147  38.96   0.265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para anchura de sépalos se rechaza la igualdad de medias:

```
modelo2<-aov(Sepal.Width~ Species )
summary(modelo2)

##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  11.35   5.672   49.16 <2e-16 ***
## Residuals   147  16.96   0.115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para la longitud de pétalos ocurre lo mismo modelo3¡-aov(Petal.Length Species) summary(modelo3) @ y finalmente para la anchura de los pétalos también:

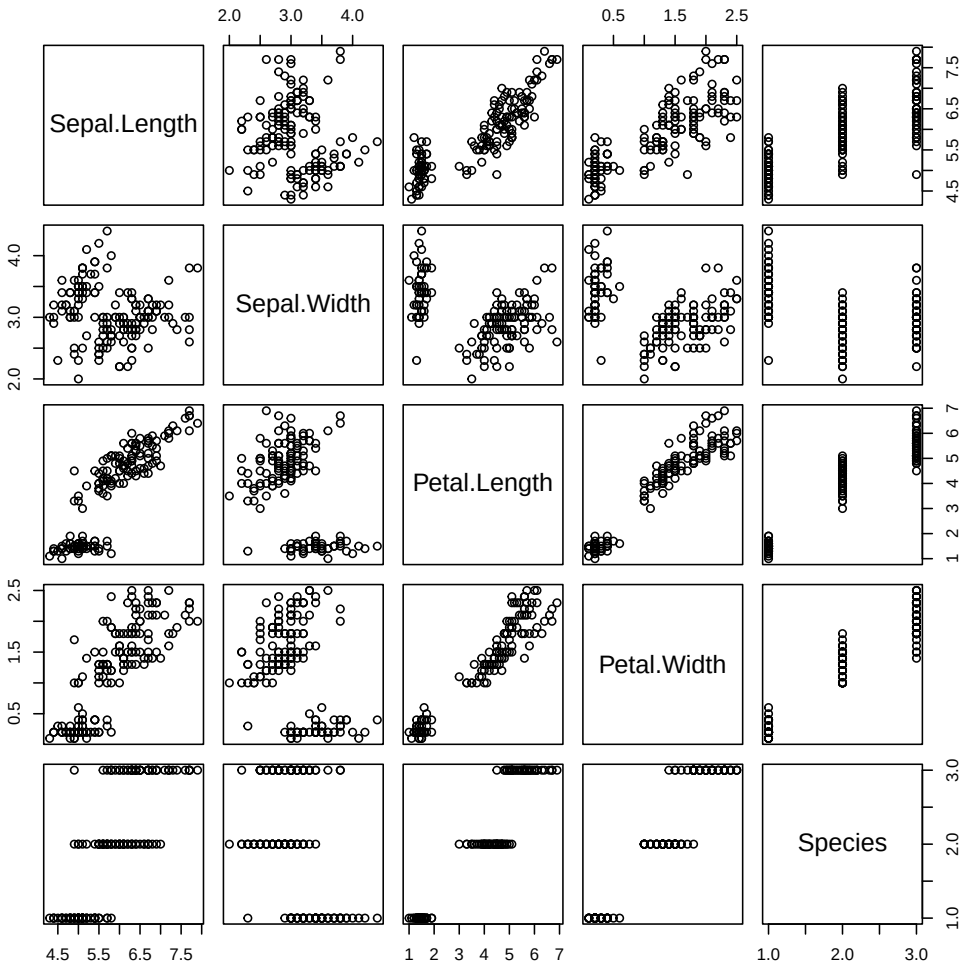
```
modelo4<-aov(Petal.Width~ Species )
summary(modelo4)

##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  80.41   40.21     960 <2e-16 ***
## Residuals   147   6.16    0.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.8.1. Representaciones gráficas

Con la orden

```
plot(iris)
```

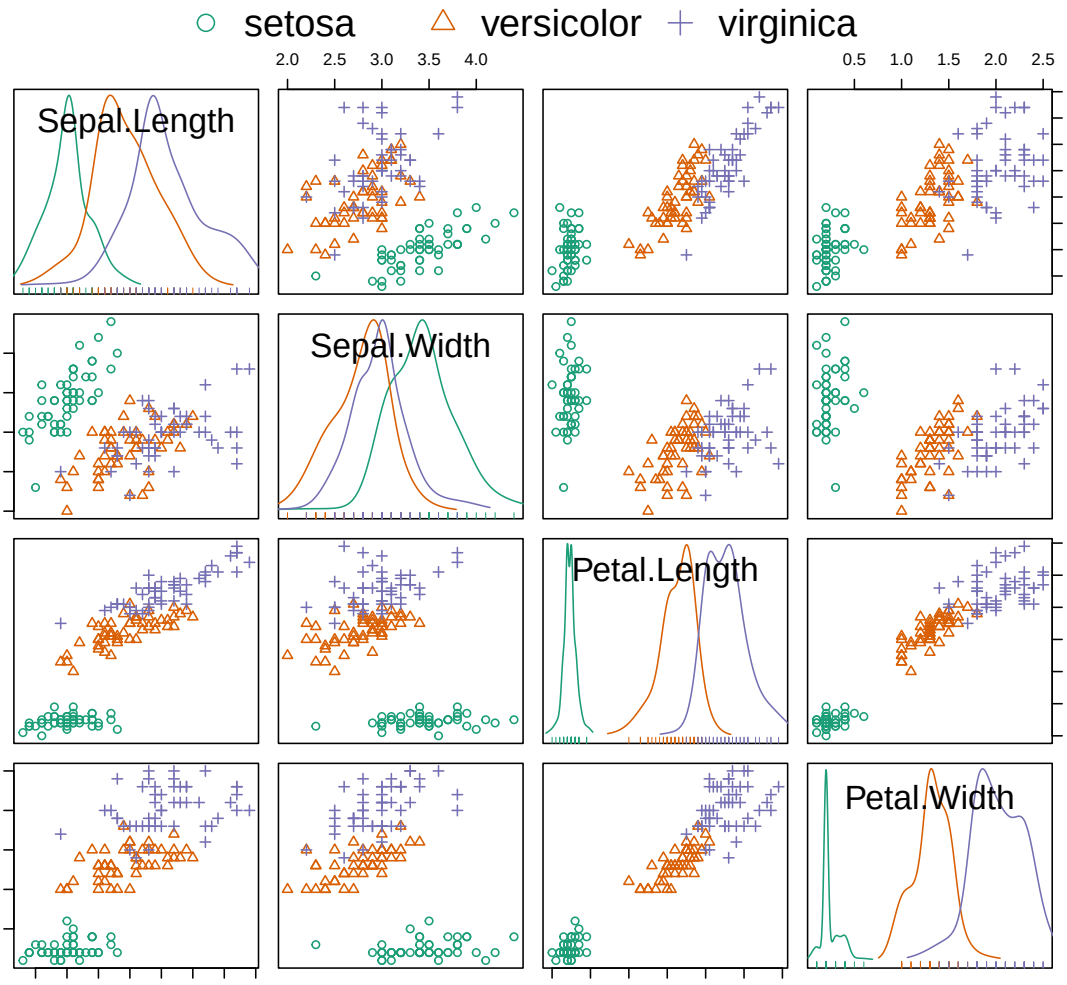


Podemos obtener una representación sencilla de los datos

Podemos obtener representaciones más avanzadas usando las ordenes:

```
library(RColorBrewer)
colors <- brewer.pal(nlevels(iris$Species),"Dark2")
library(car)

scatterplotMatrix(iris[1:4],groups=iris[[5]],
        smooth=FALSE,regLine=FALSE,      legend=FALSE,oma=c(0,0,8,0),
        col=colors)
legend("top",legend=levels(iris[[5]]),pch=1:3,col=colors,horiz=TRUE,bty="n",
        cex=1.5,xpd=TRUE)
```



1.8.2. Normalidad

Una de las hipótesis del modelo es que la muestra siga una distribución normal multivariante. Para demostrarlo podemos:

```
library(MVN)

mvn(data =iris[1:4] , univariateTest = "SW", desc = T)

## $multivariateNormality
##            Test       HZ p value MVN
## 1 Henze-Zirkler 2.336394       0  NO
##
## $univariateNormality
##           Test     Variable Statistic   p value Normality
## 1 Shapiro-Wilk Sepal.Length    0.9761  0.0102      NO
## 2 Shapiro-Wilk Sepal.Width     0.9849  0.1012      YES
## 3 Shapiro-Wilk Petal.Length    0.8763  <0.001      NO
## 4 Shapiro-Wilk Petal.Width     0.9018  <0.001      NO
##
## $Descriptives
##                n     Mean   Std.Dev Median Min Max 25th 75th       Skew
## Sepal.Length 150 5.843333 0.8280661   5.80 4.3 7.9  5.1  6.4  0.3086407
## Sepal.Width  150 3.057333 0.4358663   3.00 2.0 4.4  2.8  3.3  0.3126147
## Petal.Length 150 3.758000 1.7652982   4.35 1.0 6.9  1.6  5.1 -0.2694109
## Petal.Width  150 1.199333 0.7622377   1.30 0.1 2.5  0.3  1.8 -0.1009166
##                Kurtosis
## Sepal.Length -0.6058125
## Sepal.Width   0.1387047
## Petal.Length -1.4168574
## Petal.Width  -1.3581792
```

Podemos observar que las variables de forma conjunta no son normales y de forma marginal solo la anchura de los sépalos. Si condicionamos a las variedades, tendremos:

```
mvn(data =iris[1:50,1:4] , univariateTest = "SW", desc = T)

## $multivariateNormality
##            Test        HZ    p value MVN
## 1 Henze-Zirkler 0.9488453 0.04995356  NO
##
## $univariateNormality
##           Test     Variable Statistic   p value Normality
## 1 Shapiro-Wilk Sepal.Length    0.9777  0.4595      YES
## 2 Shapiro-Wilk Sepal.Width     0.9717  0.2715      YES
```

```
## 3 Shapiro-Wilk Petal.Length    0.9550  0.0548      YES
## 4 Shapiro-Wilk Petal.Width     0.7998  <0.001      NO
##
## $Descriptives
##               n  Mean   Std.Dev Median Min Max 25th  75th       Skew   Kurtosis
## Sepal.Length 50 5.006 0.3524897    5.0 4.3 5.8  4.8 5.200 0.11297784 -0.4508724
## Sepal.Width  50 3.428 0.3790644    3.4 2.3 4.4  3.2 3.675 0.03872946  0.5959507
## Petal.Length 50 1.462 0.1736640    1.5 1.0 1.9  1.4 1.575 0.10009538  0.6539303
## Petal.Width  50 0.246 0.1053856    0.2 0.1 0.6  0.2 0.300 1.17963278  1.2587179
```

Para la variedad setosa no se da la normalidad multivariante frente a la versicolor se da la normalidad multivariante

```
mvn(data =iris[51:100,1:4] , univariateTest = "SW", desc = T)

## $multivariateNormality
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.8388009 0.2261991 YES
##
## $univariateNormality
##           Test     Variable Statistic   p value Normality
## 1 Shapiro-Wilk Sepal.Length    0.9778    0.4647    YES
## 2 Shapiro-Wilk Sepal.Width     0.9741    0.3380    YES
## 3 Shapiro-Wilk Petal.Length    0.9660    0.1585    YES
## 4 Shapiro-Wilk Petal.Width     0.9476    0.0273    NO
##
## $Descriptives
##               n  Mean   Std.Dev Median Min Max  25th 75th        Skew
## Sepal.Length 50 5.936 0.5161711   5.90 4.9 7.0 5.600  6.3  0.09913926
## Sepal.Width  50 2.770 0.3137983   2.80 2.0 3.4 2.525  3.0 -0.34136443
## Petal.Length 50 4.260 0.4699110   4.35 3.0 5.1 4.000  4.6 -0.57060243
## Petal.Width  50 1.326 0.1977527   1.30 1.0 1.8 1.200  1.5 -0.02933377
##                Kurtosis
## Sepal.Length -0.6939138
## Sepal.Width  -0.5493203
## Petal.Length -0.1902555
## Petal.Width  -0.5873144
```

Para la variedad virgínica se da la normalidad multivariante y marginal en todas las variables.

```r
mvn(data =iris[101:150, 1:4] , univariateTest = "SW", desc = T)
```

```
## $multivariateNormality
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.7570095 0.4970237 YES
##
## $univariateNormality
##           Test     Variable Statistic   p value Normality
## 1 Shapiro-Wilk Sepal.Length    0.9712    0.2583    YES
## 2 Shapiro-Wilk Sepal.Width     0.9674    0.1809    YES
## 3 Shapiro-Wilk Petal.Length    0.9622    0.1098    YES
## 4 Shapiro-Wilk Petal.Width     0.9598    0.0870    YES
##
## $Descriptives
##                n  Mean   Std.Dev Median Min Max  25th  75th       Skew
## Sepal.Length 50 6.588 0.6358796   6.50 4.9 7.9 6.225 6.900  0.1110286
## Sepal.Width  50 2.974 0.3224966   3.00 2.2 3.8 2.800 3.175  0.3442849
## Petal.Length 50 5.552 0.5518947   5.55 4.5 6.9 5.100 5.875  0.5169175
## Petal.Width  50 2.026 0.2746501   2.00 1.4 2.5 1.800 2.300 -0.1218119
##                Kurtosis
## Sepal.Length -0.2032597
## Sepal.Width   0.3803832
## Petal.Length -0.3651161
## Petal.Width  -0.7539586
```

1.8.3. Igualdad de las matrices de varianzas covarianzas

Si obtenemos las matrices de varianzas covarianzas, tendremos:

```r
by(iris[,-5], Species, var)
```

```
## Species: setosa
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length   0.12424898 0.099216327  0.016355102 0.010330612
## Sepal.Width    0.09921633 0.143689796  0.011697959 0.009297959
## Petal.Length   0.01635510 0.011697959  0.030159184 0.006069388
## Petal.Width    0.01033061 0.009297959  0.006069388 0.011106122
## ------------------------------------------------------------
## Species: versicolor
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length   0.26643265  0.08518367   0.18289796  0.05577959
## Sepal.Width    0.08518367  0.09846939   0.08265306  0.04120408
## Petal.Length   0.18289796  0.08265306   0.22081633  0.07310204
## Petal.Width    0.05577959  0.04120408   0.07310204  0.03910612
## ------------------------------------------------------------
## Species: virginica
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length   0.40434286  0.09376327   0.30328980  0.04909388
## Sepal.Width    0.09376327  0.10400408   0.07137959  0.04762857
## Petal.Length   0.30328980  0.07137959   0.30458776  0.04882449
## Petal.Width    0.04909388  0.04762857   0.04882449  0.07543265
```

Si realizamos el contraste de igualdad de matrices de varianzas, tendremos:

```
library(biotools)

boxM(iris[,-5],Species)

##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  iris[, -5]
## Chi-Sq (approx.) = 140.94, df = 20, p-value < 2.2e-16
```

Por lo que las matrices de varianzas covarianzas son estadisticamente diferentes.

Se puede contrastar también la hipótesis de igualdad de varianza de forma marginal:

```
library(car)
leveneTest(Sepal.Length,Species )

## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value   Pr(>F)
## group   2  6.3527 0.002259 **
##       147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

leveneTest(Sepal.Width,Species )
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  0.5902 0.5555
##       147
```

```
leveneTest(Petal.Length,Species )
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   2   19.48 3.129e-08 ***
##       147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(Petal.Width,Species )
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   2  19.892 2.261e-08 ***
##       147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.8.4. Modelos aditivos y multiplicativos

Siguiendo el mismo esquema, podemos hacer los modelos MANOVA con interacciones. Para ellos vamos a generar una variable aleatoria que se llamará Zona, la cual tomará dos valores, Zona 1 y Zona 2 y la añadiremos al fichero de datos IRIS (cada alumno podrá tener valores diferentes al ser generada de forma aleatoria):

```
zona<-sample(1:2,150,replace=T)
iris2<-(cbind(iris,zona))
```

Si ahora hacemos los modelos aditivos, tendremos:

```
iris2.manova <- manova(cbind(Sepal.Length, Sepal.Width, Petal.Length,
                             + Petal.Width) ~ Species+zona)
summary(iris2.manova)
```

```
##            Df  Pillai approx F num Df den Df Pr(>F)
```

```
## Species     2 1.19234   53.147      8    288 <2e-16 ***
## zona        1 0.02376    0.870      4    143 0.4835
## Residuals 146
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vemos que la especie no es significativa frente a zona que si lo es. Si hacemos el modelo de efectos multiplicativos

```
iris3.manova <- manova(cbind(Sepal.Length, Sepal.Width, Petal.Length,
                             +  Petal.Width) ~ Species*zona)
summary(iris3.manova)

##               Df  Pillai approx F num Df den Df Pr(>F)
## Species        2 1.19579   52.785      8    284 <2e-16 ***
## zona           1 0.02449    0.885      4    141 0.4748
## Species:zona   2 0.04967    0.904      8    284 0.5134
## Residuals    144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

resulta que la interacción entre ambas si es significativa.