Índice general

Tema 4

Metodología del análisis estadístico con R

4.2. Fórmulas y modelos

4.2.1. Modelos estadísticos y modelos lineales

All models are wrong, but some are useful. [George E.P. Box]

En Estadística se formulan modelos estadísticos con la finalidad de describir (y/o predecir) el comportamiento de un cierto proceso. Se trata de modelos con componentes estocásticas que representan la incertidumbre, debida entre otras cosas a no disponer de la suficiente información sobre las variables que influyen en el fenómeno en estudio. La inferencia estadística proporciona herramientas para ajustar y evaluar la validez de los modelos estadísticos a partir de los datos observados.
En este tema nos centramos en la definición y tratamiento en R de modelos estadísticos donde una variable, denominada variable de respuesta, se pretender describir o explicar en términos de un conjunto de variables explicativas (o predictoras)[1]. Dos ejemplos de este tipo, fundamentales en Estadística, son los modelos de regresión y el análisis de la varianza (ANOVA). Ambos casos particulares de los denominados modelos lineales cuya formulación teórica se muestra a continuación:

Dadas n observaciones independientes de una variable aleatoria Y, $\{Y_1, \ldots, Y_n\}$, se dice que siguen un modelo lineal si

$$Y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{im}\beta_m + \epsilon_i, \quad i = 1, \ldots, n \tag{4.1}$$

donde $\beta_1, \ldots, \beta_m$ son parámetros (poblacionales) desconocidos, x_{ij} son valores conocidos, cada uno de los cuales representa situaciones experimentales distintas, y ϵ_i son errores aleatorios. En forma matricial el modelo anterior se escribe como

[1] Una referencia adecuada para extender estos apuntes es el libro: Faraway, J.J. (2004). *Linear Models with R.* Chapman & Hall/CRC Texts in Statistical Science. Una versión abreviada está disponible en `https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf`.

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}}_{\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}}.$$

La matriz $\mathbf{X}$ se denomina matriz del modelo y su rango constituye el rango del modelo lineal.

El modelo anterior se denomina modelo lineal de Gauss-Markov cuando se verifican las condiciones de Gauss-Markov:

$$\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0} \quad \mathbb{V}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n$$

siendo $\mathbf{0}$ un vector de ceros y $\mathbf{I}_n$ la matriz identidad de dimensión n.

El modelo (4.1) es general y admite como casos particulares algunos modelos básicos de la Estadística como son:

- Modelo de regresión lineal simple:

  $$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (i = 1, \ldots, n)$$

  En este caso la matriz del modelo (también llamada matriz de regresión) tiene dimensión $n \times 2$ y se escribe como:

  $$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

  Los parámetros del modelo son el término constante u ordenada en el origen β_0 y la pendiente β_1.

- Modelo de regresión lineal múltiple:

  $$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ik}\beta_k + \epsilon_i$$

  En este caso la matriz de regresión tiene dimensión $n \times (k+1)$ siendo k el número de variables independientes y se escribe como

  $$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,k} \end{pmatrix}$$

  y los parámetros β_j $(j = 1, \ldots, k)$ se denominan coeficientes de regresión o efectos de las variables explicativas.

- Regresión polinomial simple:

$$y_i = \beta_0 + \beta_1 x_i + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \epsilon_i$$

con matriz de regresión

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{pmatrix}$$

- Modelo de análisis de la varianza (ANOVA) de una vía. El objetivo es estudiar el efecto de un supuesto factor de variación sobre una variable de respuesta[2]. Si el factor tiene k niveles el modelo se escribe como:

$$y_{ij} = \mu_i + \epsilon_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad (i = 1, \ldots, k; j = 1, \ldots, n_i)$$

donde μ_i es la media de la variable de respuesta para el grupo i-ésimo, que se descompone como un factor común a todos los grupos, μ, más un factor específico de grupo, α_i. Aquí n_i representa el número de observaciones tomadas de dicho grupo, con lo que el total de observaciones es $n = \sum_{i=1}^{k} n_i$. El vector $\mathbf{Y}$, la matriz $\mathbf{X}$ (denominada en este contexto matriz de diseño) y el vector de parámetros en este caso tienen la forma:

$$\mathbf{Y} = \begin{pmatrix} Y_{1,1} \\ \vdots \\ Y_{1,n_1} \\ \hline Y_{2,1} \\ \vdots \\ Y_{2,n_2} \\ \hline \vdots \\ \hline Y_{k,1} \\ \vdots \\ Y_{k,n_k} \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ \hline 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \\ \hline \vdots & \vdots & \vdots & \vdots \\ \hline 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix}$$

El problema del análisis de la varianza de una vía se puede reducir a un contraste de igualdad de medias del tipo:

$$\begin{aligned} H_0 &: \quad \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1 &: \quad \mu_i \neq \mu_l \quad \text{para algún } i \neq l \end{aligned}$$

[2]El nombre abreviado ANOVA viene del inglés *Analysis of Variance* y se utiliza porque la idea es descomponer la variabilidad total de la variable de respuesta en una parte debida al factor de clasificación y otra de error.

donde la hipótesis nula es equivalente a $\alpha_1 = \cdots = \alpha_k = 0$.

El análisis de la varianza de una vía se puede generalizar a más vías considerando más factores de clasificación que a su vez pueden interaccionar entre sí. Este tipo de modelos forman parte de lo que se denominan *diseños experimentales* o factoriales.

4.2.2. Definición de un modelo estadístico en R

Desde el punto de vista del lenguaje, en tratamiento en R de este tipo de modelos es muy similar.

Para definir un modelo estadístico en R se suelen emplear fórmulas[3] del tipo:

```
respuesta ~ modelo
```

donde `modelo` especifica la expresión que describe la `respuesta`. Algunos ejemplos pueden ser:

`y ~ x` Modelo de regresión lineal simple, $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.
Otra fórmula equivalente sería `y ~ 1 + x`.

`y ~ x - 1` Regresión lineal simple pasando por el origen de coordenadas ($\beta_0 = 0$).
Otra fórmula equivalente sería `y ~ 0 + x`.

`y ~ x + I(x ^ 2)` Regresión polinomial de grado 2, $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$.
Una versión usando polinomios ortogonales es `y ~ poly(x,2)`.

`y ~ A` Análisis de la varianza de una vía, $Y_{ij} = \mu_i + \epsilon_{ij}$. Aquí `A` define los k grupos (por ejemplo un factor con k niveles).

`y ~ A * B` Diseño experimental con dos factores de clasificación, `A` y `B`.
Otra fórmula equivalente sería `y ~ A + B + A:B`.

`y ~ (A + B + C)^ 2` Diseño experimental con tres factores de clasificación, `A`, `B` y `C`, pero solo se consideran interacciones de orden 2.
Otra fórmula equivalente sería `y ~ A*B*C - A:B:C`.

`y ~ A * x` Modelos de regresión lineal simple separados para cada nivel de `A`.
Otra fórmula equivalente sería `y ~  A/x`.

Las fórmulas anteriores se pueden escribir de forma general como:

```
respuesta ~ op_1 term_1 op_2 term_2 op_3 term_3 ···
```

donde

[3] También para generar algunos gráficos como por ejemplo los diagramas de cajas múltiples con `boxplot`.

- `response` es un vector (o una matriz) con las observaciones de la(s) variable(s) de respuesta;
- `op_1`, `op_2`, ..., son operadores de fórmula (con un significado especial en este contexto, ver Tabla 4.1);
- `term_1`, `term_2`, ..., son alguno de: vector, matriz, factor, o una expresión en términos de factores, vectores o matrices conectados por operadores.

Operador	Descripción
`+` , `-`	incluye, excluye efectos principales
`1`	término constante (por defecto se incluye siempre)
`*` , `:`	efectos principales más interacciones `a*b = a+b+a:b`
`/` , `\%in\%`	efectos anidados `a/b = a + b \%in\% a = a + a:b`
`^n`	interacciones hasta nivel n `(a+b)^2 = a+b+a:b`
`I()`	función identidad `y ~ I(x^2)` en lugar de `y ~ x^2 = y ~ x` (la interacción `x:x=x`)
`poly()`	polinomios ortogonales

Tabla 4.1: Operadores para escribir fórmulas en R.

4.2.3. Ajuste de un modelo lineal

El vector de parámetros $\boldsymbol{\beta}$ del modelo lineal se estima por mínimos cuadrados o máxima verosimilitud. En el caso de errores $\boldsymbol{\epsilon}$ con distribución Normal ambos métodos son equivalentes, siendo el estimador mínimo-cuadrático y máximo-verosímil de $\boldsymbol{\beta}$ el definido por la expresión:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Al proceso de estimación de los parámetros no referimos comúnmente como "ajuste del modelo". En R la función básica para esto es la función `lm` cuya sintaxis puede ser simplemente:

```
lm(formula, data)
```

donde `data` es un objeto de tipo data frame que incluye las variables usadas en la fórmula[4].

[4]Es posible omitir dicho data frame en la evaluación de la función pero en dicho caso es necesario que las variables usadas en la fórmula estén en el espacio de trabajo (o en general en la lista de búsqueda de R).

Es posible además especificar entre otros los siguientes argumentos opcionales: `subset`, para especificar un subconjunto de los datos para el ajuste; `weights`, para ajustar el modelo usando un criterio de mínimos cuadrados ponderados; y `offset`, que permite especificar una componente del modelo conocida a priori, en cuyo caso se restará a la respuesta.

La función `lm` devuelve un objeto de tipo lista de la clase `lm`, con al menos las siguientes componentes:

- `coefficients`: vector de coeficientes del modelo ajustado, $\widehat{\boldsymbol{\beta}}$.
- `fitted.values`: vector con los valores ajustados, $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$.
- `residuals`: vector con los residuos del ajuste $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}}$.
- `rank`: rango del modelo (rango de $\mathbf{X}$).
- `df.residual`: los grados de libertad de los residuos.

4.2.4. Funciones genéricas para extraer información adicional del ajuste

Un objeto de tipo `lm` contiene diversa información del ajuste que puede mostrarse, representarse gráficamente, así como extenderse a través de varias funciones genéricas como las que se resumen en la Tabla 4.2. El uso y la utilidad de estas funciones lo ilustramos a continuación (y en la sesión de prácticas) a través de ejemplos.

Función	Descripción
`fitted`	valores ajustados
`coef`	coeficientes estimados (y errores estándar)
`confint`	intervalos de confianza para los coeficientes
`residuals`	residuos
`summary`	resumen detallado del modelo estimado
`predict`	calcula predicciones para nuevos datos
`anova`	tablas ANOVA (y comparación de modelos)
`vcov`	matriz de covarianzas de los parámetros estimados
`plot`	gráficos de diagnóstico
`termplot`	gráfico de efectos parciales
`step`	selección de modelos organizados jerárquicamente

Tabla 4.2: Funciones genéricas para extraer información de un ajuste con `lm`.

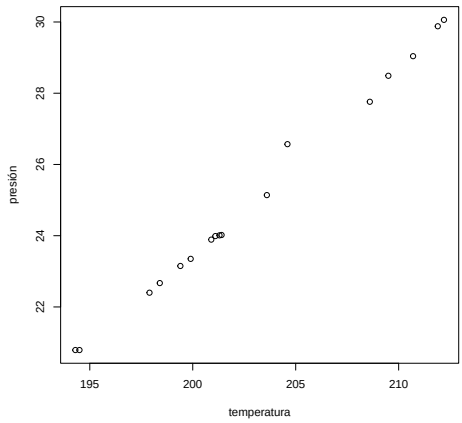
4.2.5. Ejemplos

Ejemplo 1: Regresión lineal simple

Entre 1840 y 1850 el físico escocés James D. Forbes desarrolló una serie de experimentos con el fin de estimar la altitud sobre el nivel del mar a partir de medidas sobre el punto de ebullición del agua. Él sabía que la altitud podía determinarse a partir de la presión atmosférica medida a través de un barómetro, con presiones menores correspondientes a elevadas altitudes. En aquella época los barómetros eran instrumentos muy frágiles y Forbes pensó que sería posible reemplazar las medidas de la presión con medidas de la temperatura de ebullición del agua. Forbes recogió datos en 17 lugares de los Alpes y los montes de Escocia. En cada lugar midió con un barómetro la presión en pulgadas de mercurio (`pres`) y con un termómetro la temperatura de ebullición del agua en grados Farenheit (`bp`).

Los datos recogidos por Forbes están disponibles en el data frame `forbes` del paquete `MASS`. Como primer paso construimos un diagrama de dispersión de los datos para explorar la relación que existe entre las variables:

```
> library(MASS)
> plot(forbes$bp,forbes$pres, xlab = 'temperatura', ylab = 'presión')
```



Del gráfico vemos que parece existir una fuerte relación lineal. Nuestro objetivo es ajustar un modelo lineal que permita describir la presión en función de la temperatura, esto es,

$$\text{presión} = \beta_0 + \beta_1 \times \text{temperatura} + \epsilon$$

donde ϵ representa los errores del modelo verificando las condiciones del modelo lineal Normal de Gauss-Markov.

Usamos la función `lm` para ajustar el modelo que escribimos como `pres~bp`:

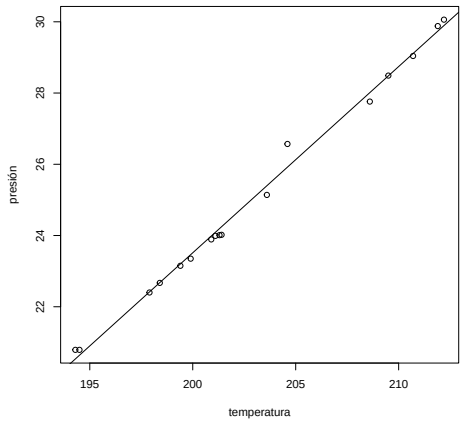
```
> lm(pres~bp,data=forbes)


Call:
lm(formula = pres ~ bp, data = forbes)

Coefficients:
(Intercept)           bp
   -81.0637       0.5229
```

La función muestra los coeficientes estimados $\widehat{\beta}_0 = -81.0637$ y $\widehat{\beta}_1 = 0.5229$. Para ver todos los valores que devuelve dicha función asignamos su valor a un objeto y visualizamos su contenido:

```
> fit<-lm(pres~bp,data=forbes)
> typeof(fit); class(fit)

[1] "list"
[1] "lm"

> names(fit)

 [1] "coefficients"  "residuals"     "effects"       "rank"          "fitted.values"
 [6] "assign"        "qr"            "df.residual"   "xlevels"       "call"
[11] "terms"         "model"
```

El ajuste podemos representarlo sobre el diagrama de dispersión anterior usando la función `abline`:

```
> plot(forbes$bp,forbes$pres, xlab = 'temperatura', ylab = 'presión')
> abline(fit)
```



Para evaluar la bondad del ajuste usamos la función `summary`:

```
> summary(fit)


Call:
lm(formula = pres ~ bp, data = forbes)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25717 -0.11246 -0.05102  0.14283  0.64994

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -81.06373    2.05182  -39.51   <2e-16 ***
bp            0.52289    0.01011   51.74   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2328 on 15 degrees of freedom
Multiple R-squared:  0.9944,Adjusted R-squared:  0.9941
F-statistic:  2677 on 1 and 15 DF,  p-value: < 2.2e-16
```

Entre otros resultados localizamos el valor del coeficiente de determinación $R^2 = 0.9944$, y el error estándar residual $\widehat{\sigma}_R = 0.2328$. El primero nos indica que el modelo ajusta bastante bien a los datos.
También podemos ver los resultados de los contrastes de significación de la pendiente y de la ordenada en el origen. Estos contrastes se realizan bajo el supuesto de que se cumplen las condiciones de Gauss-Markov además de que los errores del modelo siguen una distribución Normal[5].
Para la pendiente se ha formulado el problema $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$. Los resultados muestran el cálculo del contraste como sigue:

- $\widehat{\beta}_1 = 0.5229$ y su error estándar $s.e.(\widehat{\beta}_1) = 0.0101$
- Estadístico de contraste: $t = \frac{\widehat{\beta}_1 - 0}{s.e.(\widehat{\beta}_1)} = 51.7408$
- P-valor$\approx$ 0, lo que indica que se debe rechazar H_0, y por tanto la temperatura contribuye de manera significativa a explicar la presión.

Para la ordenada en el origen se ha formulado el problema $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$, y los resultados mostrados son:

[5]Esto habrá que comprobarlo pero lo dejamos para la sesión de prácticas.

- $\widehat{\beta}_0 = -81.0637$ y su error estándar $s.e.(\widehat{\beta}_0) = 2.0518$
- Estadístico de contraste: $t = \frac{\widehat{\beta}_0 - 0}{s.e.(\widehat{\beta}_0)} - 39.5082$
- P-valor≈ 0, lo que indica que se debe rechazar H_0, y por tanto incluir un término constante en la ecuación del modelo parece adecuado.

Entre los resultados también se muestra la solución al contraste de regresión basado en la descomposición de la variabilidad. El valor del estadístico de contraste es $F = 2677.1$. Se trata de un valor muy elevado que lleva a rechazar la hipótesis nula (observa que el p-valor correspondiente es aproximadamente 0) que en este caso coincide con la del contraste de significación de la pendiente que hemos descrito antes.
Los cálculos intermedios del contraste de regresión se pueden recoger en la denominada tabla ANOVA para la regresión, que podemos obtener con la función `anova`:

```
> anova(fit)

Analysis of Variance Table

Response: pres
          Df  Sum Sq Mean Sq F value    Pr(>F)
bp         1 145.125 145.125  2677.1 < 2.2e-16 ***
Residuals 15   0.813   0.054
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En la tabla ANOVA que nos da R se muestran las dos de las fuentes de variación del problema, VNE=0.813, y VE=145.125, con sus respectivos grados de libertad (15 y 1). Con ellas se calcula el valor del estadístico de contraste es $F = 2677.1$.

Usando de nuevo la hipótesis de normalidad de los errores se calculan intervalos de confianza para la pendiente, β_1, y la ordenada en el origen, β_0, usando la función `confint`:

```
> confint(fit)

                2.5 %      97.5 %
(Intercept) -85.437080 -76.6903740
bp            0.501352   0.5444328
```

El resultado muestra los límites inferiores y superiores de cada intervalo. Para la pendiente se obtiene $(0.5014; 0.5444)$, y para la ordenada $(-85.4371; -76.6904)$. Por defecto nos muestra intervalos con nivel de confianza $1 - \alpha = 0.95$, pero se puede cambiar si se desea usando el argumento `level`.

A partir del modelo ajustado podemos plantearnos también hacer predicciones que era el objetivo del estudio de Forbes. A continuación usamos la función `predict` para calcular un intervalo de confianza para la presión que se esperaría (por término medio) para una localización en la que la temperatura de ebullición del agua es de 200:

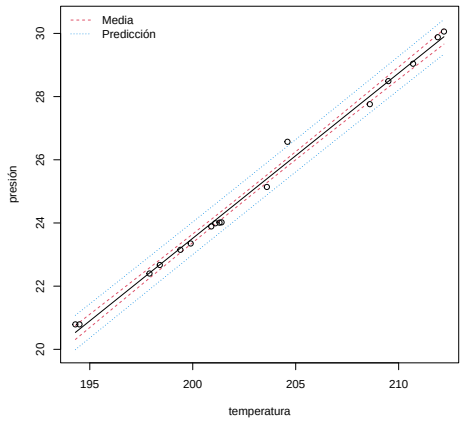
```
> predict(fit,newdata=data.frame(bp=200),interval='confidence',se.fit=TRUE)

$fit
       fit      lwr      upr
1 23.51475 23.37862 23.65089

$se.fit
[1] 0.06386993

$df
[1] 15

$residual.scale
[1] 0.2328294
```

La presión esperada es de 23.5147 con un error estándar de 0.0639, lo que conduce al intervalo (23.3786; 23.6509) con una confianza del 95 %. Observa que el intervalo es bastante estrecho.
Si en lugar de un intervalo para la presión media esperada buscamos un intervalo de predicción entonces el resultado sería:

```
> predict(fit,newdata=data.frame(bp=200),interval='prediction')

       fit      lwr      upr
1 23.51475 23.00016 24.02935
```

Podemos finalmente representar las bandas de confianza para la media de la presión a partir del modelo y las bandas de predicción (al 95 %). Para ello usamos de nuevo la función `predict` pero hacemos variar el punto donde se realiza la predicción dentro del rango de valores observados de la temperatura. El código que permite hacer esto y representar las bandas obtenidas se muestra a continuación junto con el resultado.

```
> x0<-data.frame(bp=seq(min(forbes$bp),max(forbes$bp),length.out=20))
> pred.m<-predict(fit,newdata=x0,interval='confidence',se.fit=T)
> pred.p<-predict(fit,newdata=x0,interval='prediction',se.fit=T)
> matplot(x0$bp,cbind(pred.m$fit,pred.p$fit[,-1]),lty=c(1,2,2,3,3),
+   col=c(1,2,2,4,4),type='l',xlab='temperatura',ylab='presión',main='')
> legend('topleft',c('Media','Predicción'), lty=c(2,3),col=c(2,4),bt='n')
> points(forbes$bp,forbes$pres)
```



El estudio anterior se basa en el supuesto de que se verifican las hipótesis de Gauss-Markov, además de la normalidad de los errores del modelo. La verificación de estos supuesto constituye lo que se denomina el análisis de los residuos o diagnósticos del modelo, y se realiza sobre los residuos `fit$residuals`, que constituyen estimaciones de los errores del modelo $\boldsymbol{\epsilon}$. La función `plot` nos permite mostrar algunos gráficos de diagnóstico (`plot(fit)`). Como hemos indicado antes este aspecto lo describiremos con más detalle en la sesión de prácticas.

Ejemplo 2: Análisis de la varianza de una vía

Los siguientes datos corresponden a 24 tiempos de coagulación sanguínea en ratas[6]. Los 24 animales se asignaron aleatoriamente a 4 dietas diferentes y las muestras se tomaron al azar.

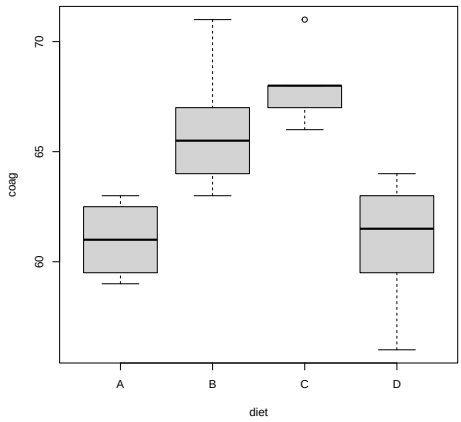
A	B	C	D
60	65	71	62
59	66	66	63
63	67	68	60
62	63	68	61
	64	67	64
	71	68	63
			56
			59

El objetivo es comprobar si existen diferencias significativa entre los tiempos de coagulación para las 4 dietas. Para ello formulamos un modelo ANOVA con el tiempo de coagulación como variable de respuesta y la dieta como factor de clasificación.

Los datos anteriores están disponibles en un data frame (`coagulation`) dentro del paquete *faraway*. Comenzamos cargando los datos y construyendo un diagrama de cajas múltiple que nos permita apreciar visualmente las posibles diferencias entre las dietas:

[6]Box, G.P, Hunter, W.G. y Hunter, J.S. (1978) *Statistics for Experimenters*. Wiley.

```
> library(faraway)
> data(coagulation)
> boxplot(coag~diet,data=coagulation)
```



Este gráfico parece mostrar diferencias entre los grupos. Por otro lado de este gráfico también podemos tener una primera impresión sobre la validez de las hipótesis del modelo, en concreto verificar que no se observa nada de los siguiente:

- Datos anómalos (*outliers*).
- Asimetría (que sería incompatible con el supuesto de normalidad).
- Varianzas desiguales (lo que correspondería a cajas de dimensiones notablemente diferentes).

En este caso no parece haber problemas en relación a ninguno de las tres aspectos anteriores, teniendo en cuenta el reducido número de observaciones en algunos grupos[7].

A continuación procedemos al ajuste del modelo. Para ello podemos hacerlo de nuevo usando la función `lm`:

```
> lm(coag~diet,data=coagulation)


Call:
lm(formula = coag ~ diet, data = coagulation)

Coefficients:
(Intercept)         dietB         dietC         dietD
  6.100e+01     5.000e+00     7.000e+00     2.991e-15
```

[7] Observa que en este caso las diferencias en variabilidad pueden explicarse por los reducidos tamaños muestrales junto con que hay valores repetidos.

Observamos que nos devuelve los efectos del modelo con término constante, esto es, $\mu = 61$, $\alpha_B = 5$, $\alpha_C = 7$, $\alpha_D = 2.9914279 \times 10^{-15}$ [8]. Otra posibilidad sería estimar directamente los parámetros $\mu_i = \mu + \alpha_i$ para lo cual escribiríamos:

```
> lm(coag~diet-1,data=coagulation)


Call:
lm(formula = coag ~ diet - 1, data = coagulation)

Coefficients:
dietA  dietB  dietC  dietD
   61     66     68     61
```

Una vez ajustado el modelo (con cualquiera de las dos opciones anteriores) resolvemos el problema de contraste:

$$\begin{aligned} H_0 &: \quad \mu_A = \mu_B = \mu_C = \mu_D \\ H_1 &: \quad \mu_i \neq \mu_l \quad \text{para algún } i \neq l \end{aligned}$$

para lo que utilizamos la función `anova` evaluada en el objeto resultante del ajuste:

```
> fit<-lm(coag~diet,data=coagulation)
> anova(fit)

Analysis of Variance Table

Response: coag
          Df Sum Sq Mean Sq F value    Pr(>F)
diet       3    228    76.0  13.571 4.658e-05 ***
Residuals 20    112     5.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El p-valor 4.658471×10^{-5} nos indica que podemos rechazar claramente la hipótesis nula lo que supondría que la dieta tiene un efecto significativo en el tiempo de coagulación.

Otra forma de ajustar el modelo ANOVA y resolver el problema es usando la función `aov` en lugar de `lm`. Su uso y resultado en este caso sería:

[8] Con esta parametrización del modelo se impone que $\sum n_i \alpha_i = 0$ por lo que solo es necesario estimar tres de los α's.

```
> fit2<-aov(coag~diet,data=coagulation)
> summary(fit2)

            Df Sum Sq Mean Sq F value   Pr(>F)
diet         3    228    76.0   13.57 4.66e-05 ***
Residuals   20    112     5.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cuando encontramos diferencias significativas entre los factores es interesante realizar un análisis posterior que permite descubrir la raíz de estas diferencias. Una primera idea sería hacer una comparación de los grupos dos a dos usando la función `pairwise.t.test`:

```
> pairwise.t.test(coagulation$coag,coagulation$diet)


Pairwise comparisons using t tests with pooled SD

data:  coagulation$coag and coagulation$diet

  A       B       C
B 0.01141 -       -
C 0.00090 0.31755 -
D 1.00000 0.00345 0.00014

P value adjustment method: holm
```

Un estudio más adecuado sería a través de las comparaciones múltiples de Tukey que podemos obtener usando la función `TukeyHSD`:

```
> TukeyHSD(aov(coag~diet, coagulation))

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = coag ~ diet, data = coagulation)

$diet
    diff         lwr       upr     p adj
B-A    5   0.7245544  9.275446 0.0183283
C-A    7   2.7245544 11.275446 0.0009577
D-A    0  -4.0560438  4.056044 1.0000000
```

```
C-B    2  -1.8240748  5.824075 0.4766005
D-B   -5  -8.5770944 -1.422906 0.0044114
D-C   -7 -10.5770944 -3.422906 0.0001268
```

Como resultado nos muestra intervalos de confianza para las misma diferencias $\alpha_i - \alpha_j$, junto con una corrección del p-valor adecuada para el problema de comparaciones múltiples formulado.

Para finalizar tenemos que de nuevo hacer un comentario importante y es que los análisis anteriores se han hecho bajo el supuesto de que se cumplen las hipótesis del modelo. Evaluar si se verifican dichas hipótesis supondría un análisis de residuos similar al caso de los modelos de regresión. Este aspecto de nuevo no lo recogemos aquí.