Modelos lineales: Regresión, ANOVA y ANCOVA

Luis Cayuela

Octubre de 2015

Área de Biodiversidad y Conservación, Universidad Rey Juan Carlos, Departamental 1 – DI. 231, c/ Tulipán s/n. E-28933 Móstoles (Madrid), España. E-mail: luis.cayuela@urjc.es.

Modelos lineales: Regresión, ANOVA y ANCOVA (versión 1.5)

Publicado por: Luis Cayuela

Para cualquier comentario o sugerencia por favor remitirse al autor de la obra.

Índice

1. Conceptos estadísticos básicos

¿Qué es una regresión? ¿Y un ANOVA? ¿Cuál es la principal diferencia entre ambos? ¿Qué supuestos estadísticos debemos asumir cuando llevemos a cabo este tipo de análisis? Estas y otras preguntas son críticas en la aplicación de modelos lineales a la resolución de problemas estadísticos. Por ello, la primera parte de esta sesión la dedicaremos a aclarar dichos conceptos.

El análisis de regresión se usa para explicar o modelar la relación entre una variable continua Y, llamada variable respuesta o variable dependiente, y una o más variables continuas X1,.....,Xp, llamadas variables explicativas o independientes. Cuando $p = 1$, se denomina regresión simple y cuando $p > 1$ se denomina regresión múltiple. Cuando hay más de una variable respuesta Y, entonces el análisis se denomina regresión múltiple multivariada. Cuando las Y son totalmente independientes entre sí, entonces hacer una regresión múltiple multivariada sería el equivalente a realizar tantas regresiones múltiples univariadas como Y's haya.

Si la(s) variable(s) explicativas son categóricas en vez de continuas entonces nos enfrentamos ante un caso típico de análisis de la varianza o ANOVA (ADEVA en español). Al igual que antes, si $p = 1$, el análisis se denomina ANOVA unifactorial, mientras que si $p > 1$ el análisis se denomina ANOVA multifactorial. Si en vez de una variable respuesta continua tenemos dos o más Y, entonces el análisis se denomina ANOVA multivariado (MANOVA) de uno o varios factores. Este tipo de análisis también queda fuera del ámbito de esta sesión.

Por último, es posible que en el mismo análisis aparezcan tanto variables explicativas continuas como categóricas, y en este caso el análisis pasaría a denominarse análisis de la covarianza o ANCOVA. Aquí ya no haríamos distinción entre único o múltiple ya que este análisis se compone siempre de, al menos, dos variables explicativas (una continua y una categórica).

A pesar de la abundancia de terminología, todos estos modelos caen dentro de la categoría de modelos lineales. En esta sesión nos centraremos únicamente en las técnicas univariadas (regresión, ANOVA y ANCOVA). En R todos los análisis univariados de este tipo se ajustan utilizando una única función, la función lm(), ya que la forma de ajustar cualquiera de estos modelos es idéntica, independientemente de que tengamos una o más variables explicativas y de que éstas sean continuas o categóricas.

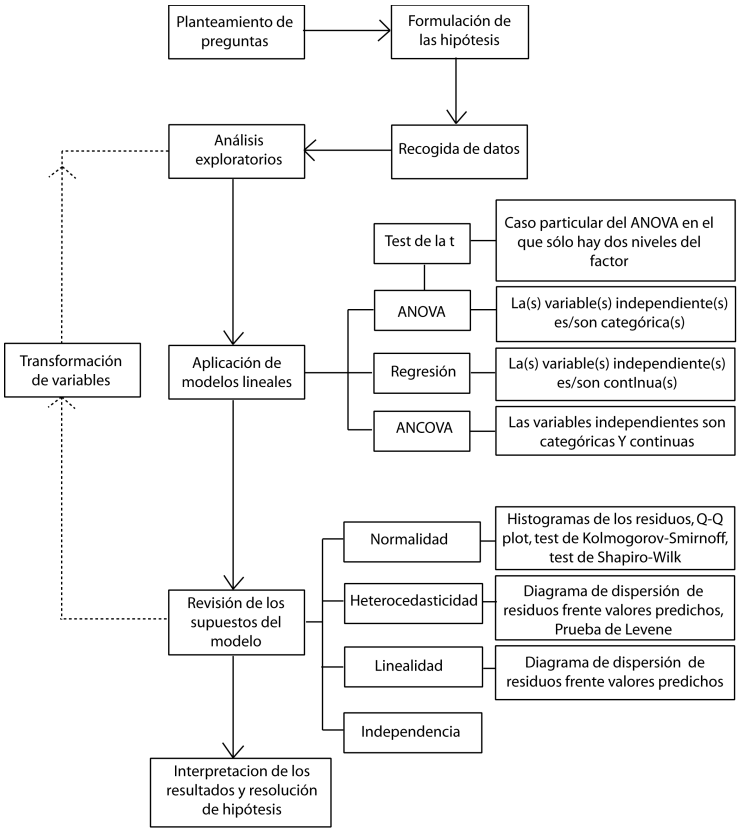


Figura 1: Esquema conceptual de los pasos que deben seguirse a la hora de ajustar un modelo lineal univariante.

Sin entrar en muchos detalles, cabe recordar que los modelos lineales se basan en una serie de supuestos, algunos de los cuales pueden y deben comprobarse una vez ajustado el modelo. Estos son:

1. **Independencia.** Los sujetos muestrales y, por tanto, los residuos del modelo, son independientes entre sí.
2. **Linealidad.** La respuesta de Y frente a X es lineal.
3. **Normalidad.** Los residuos del modelo son normales, es decir, siguen una distribución de tipo gaussiana (campana de Gauss).
4. **Homocedasticidad.** La varianza residual tiene que ser constante.

2. Cosas importantes antes de empezar

La estadística comienza con un problema, continua con la recogida de datos, y termina con el análisis de los mismos, lo que conduce a unas conclusiones sobre

las hipótesis de partida. Es un error muy común enredarse en análisis muy complejos sin prestar atención a los objetivos que se persiguen, a la pregunta que se quiere contestar, o incluso a si los datos de los que se dispone son los apropiados para el análisis propuesto. Para formular el problema correctamente uno debe:

1. Comprender el problema de fondo y su contexto.
2. Comprender bien el objetivo u objetivos del estudio. Hay que tener cuidado con los análisis no dirigidos. Si buscas lo suficiente siempre encontrarás algún tipo de relación entre variables, pero puede que esta relación no sea más que una coincidencia.
3. Plantear el problema en términos estadísticos. Este es uno de los pasos más difíciles e implica la formulación de hipótesis y modelos. Una vez que el problema ha sido traducido al lenguaje de la estadística, la solución suele ser rutinaria.
4. Entender bien los datos. ¿Son datos observacionales o experimentales? ¿Hay valores faltantes? ¿Cómo están representadas las variables cualitativas? ¿Cuáles son las unidades de medida? ¿Hay algún error en los datos? Por todo ello, es importante revisar bien los datos y llevar a cabo algún análisis preliminar para detectar anomalías en los mismos.

3. Regresión simple

3.1. Como ajustar un modelo lineal en R

Una vez que tenemos el problema formulado y los datos recogidos, ajustar un modelo lineal es muy, muy sencillo en R. La función lm() nos permite ajustar el modelo especificado. La forma más común de especificar el modelo es utilizando el operador ~ para indicar que la respuesta Y es modelada por un predictor lineal definido por X1,...,Xn. Tomemos como ejemplo la base de datos cars, que contiene la velocidad de 50 coches (millas/hora) y la distancia (pies) que les lleva frenar (¡ojo! ¡son datos de los años 20!).

```
> data(cars)
> lm.cars<-lm(dist~speed, data=cars)
```

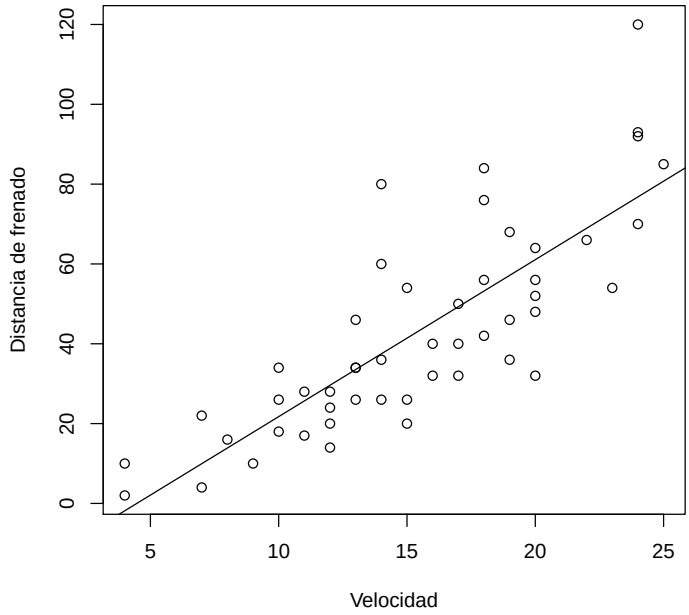
Ahora ya tenemos un objeto, llamado lm.cars, que contiene el modelo lineal ajustado, en dónde la distancia de frenado sería una función de la velocidad de los mismos. Si utilizamos la función str() veremos que este nuevo objeto tiene, en apariencia, una estructura muy compleja. Esto no debe asustarnos. El objeto creado contiene en realidad toda la información referente al modelo ajustado, como los coeficientes del modelo, la varianza explicada, los valores de los residuos, etc. Podemos acceder a esta información utilizando el operador $ de manera similar a cómo accedíamos a las variables de un arreglo de datos (p.e. lm.cars$fitted.values). Sin embargo, resulta mucho más fácil obtener los resultados del modelo utilizando la función summary().

```
> summary(lm.cars)

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,        Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Aquí podemos ver muchas de las cosas que nos interesan para responder a nuestra pregunta. En primer lugar tenemos los coeficientes del modelo ajustado y su significación (Pr(>|t|)). El modelo no sólo tiene un coeficiente que modela la relación lineal entre la variable respuesta (dist) y la variable explicativa (speed), sino que además tiene una constante, que es lo que R denomina Intercept o punto de corte con el eje Y, es decir el valor que toma Y cuando X = 0. Si este valor no es muy distinto de 0 entonces el Intercept suele no ser significativo[1]. En este caso, sí es significativo y toma un valor de -17.5791. Esto indicaría teóricamente que cuando la velocidad del coche es 0, su distancia de frenado es -17.5791 pies, si bien como todos sabemos, esta aseveración no tiene sentido alguno. El problema está en los supuestos de los modelos lineales, ya que la relación entre muchas variables es lineal sólo en un determinado rango de los valores de X y no puede extrapolarse más allá de estos valores, tal es el caso de nuestro ejemplo. Para representar gráficamente la recta de regresión, podemos usar la función gráfica de bajo nivel `abline()`.

[1] La significación es un valor que nos indica con que probabilidad la relación observada es distinta de la hipótesis nula (en este ejemplo la hipótesis nula sería que el punto de corte con el eje Y es cero) .

```
> plot(cars$dist ~ cars$speed, xlab="Velocidad", ylab="Distancia de frenado")
> abline(lm.cars)
```



Más allá de la interpretación que hagamos de la constante, lo que interesaría más sería la significación de la variable explicativa speed, que en este caso concreto toma un valor muy bajo ($\mathsf{Pr(>|t|)} = \mathsf{1.49e\text{-}12}$). Esto significa que hay una probabilidad muy baja de que el coeficiente estimado de speed en el modelo lineal esté dentro de una distribución aleatoria de valores "nulos", es decir, de coeficientes obtenidos aleatoriamente pero que en realidad no son distintos de cero. Por tanto rechazaríamos la hipótesis nula de que este coeficiente es cero.

Por último, interesa ver el coeficiente de determinación del modelo o R^2. Este coeficiente indica la cantidad de variabilidad explicada por el modelo. Cuanto mayor sea este coeficiente más predecible es la variable respuesta en función de la variable o variables explicativas. El R^2 ajustado corrige el R^2 por el número de parámetros (variables explicativas) del modelo ya que, en general, cuantas más variables explicativas estén incluidas en el modelo, mayor es el R^2, independientemente de que dichas variables sean o no relevantes para el modelo. En nuestro modelo, el R^2 corregido es 0.6438, lo que significa que el 64 % de la variabilidad de la distancia de frenado se puede explicar por la velocidad a la que va el coche.

3.2. Fundamentos teóricos del cálculo e interpretación de los parámetros de la recta de regresión

El ajuste de un modelo de regresión comienza por el planteamiento de una hipótesis sobre causa y efecto: el valor de la variable X causa, directa o indirectamente, el valor de la variable Y. En algunos casos, la direccionalidad de la causa y el efecto es muy clara -nosotros hipotetizamos que la distancia de frenado de un coche depende de su velocidad y no al revés. En otros casos, la dirección de la causa y el efecto no es tan obvia -¿controlan los predadores la abundancia de las presas, o es la abundancia de las presas la que controla la abundancia de los predadores?

Una vez que hemos tomado una decisión sobre la dirección de la causa y el efecto, el siguiente paso es describir la relación como una función matemática:

$$Y = f(X)$$

En otras palabras, aplicaremos una función a cada valor de la variable X (el input) para generar un valor correspondiente de Y (el output). Hay muchas y muy complejas formas de describir matemáticamente la relación entre dos variables, pero una de las más sencillas es que Y sea una función linear de X:

$$Y = \beta_0 + \beta_1 X$$

Esta función dice que "tomes el valor de la variable X, lo multipliques por β_1, y se lo añadas a β_0. El resultado de esto es el valor de la variable Y". Esta ecuación describe la gráfica de una línea recta (ver la gráfica en el apartado 3.1). El modelo tiene dos parámetros β_0 y β_1, que se denominan intercepto y pendiente respectivamente. El intercepto (β_0) es el valor de la función cuando X=0. El intercepto se mide en las mismas unidades que la variable Y. La pendiente (β_1) mide el cambio en la variable Y por cada unidad de cambio en la variable X. La pendiente es por tanto un ratio y se mide en unidades $\triangle Y/\triangle X$. Si se conoce el valor de la pendiente y el intercepto, se puede calcular cualquier valor de Y para cualquier valor conocido de X.

3.2.1. Ajustar los datos a un modelo lineal

Los datos en un análisis de regresión consisten en una serie de observaciones pareadas. Cada observación incluye un valor para la variable X y un valor para la correspondiente variable Y, que tienen que ser medidos necesariamente sobre la misma muestra (réplica). En nuestro ejemplo, estos datos están recogidos en el arreglo de datos `cars`. El subíndice `i` indica el número de la réplica o muestra. Si hay un total de `n` réplicas en nuestros datos, el subíndice `i` puede tomar cualquier valor desde `i = 1` a `n`. El modelo que ajustaremos será entonces el siguiente:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Los dos parámetros β_0 y β_1 son desconocidos. Pero hay también otro parámetro desconocido, ε_i, que representa el término error. Mientras que β_0 y β_1 son constantes en el modelo, ε_i es una variable aleatoria que sigue una distribución normal. Esta distribución tiene un valor esperado (media) de 0, y una varianza equivalente a σ^2, que puede ser conocida o desconocida. Si todos nuestros datos caen perfectamente a lo largo de una única línea recta, entonces la $\sigma^2 = 0$, y será una cuestión fácil conectar todos los puntos y medir el intercepto (β_0) y la pendiente (β_1) de esa recta directamente de la línea. Sin embargo, la mayoría de los datos ecológicos exhiben un cierto grado de variación, y nuestros datos aparecerán dispersos formando una nube en lugar de una línea recta perfecta. Cuanto mayor sea el valor de σ^2, mayor será el ruido o error de los datos en torno a la recta de regresión.

Si observamos la figura del ejemplo anterior, vemos que hay una clara relación entre la distancia de frenado de un coche y su velocidad, pero los puntos no caen a lo largo de una línea recta perfecta. ¿Dónde deberíamos colocar la recta de regresión? Intuitivamente, parece que **la línea de la recta de regresión debería de pasar por el centro de la nube de datos**, definida por los puntos $(\overline{X}, \overline{Y})$. Para nuestro ejemplo, el centro correspondería a los puntos:

```
> meanX <- mean(cars$speed)
> meanY <- mean(cars$dist)
> meanX; meanY
```

```
[1] 15.4
```

```
[1] 42.98
```

Ahora podemos rotar la línea en torno a este punto central hasta que llegemos al mejor ajuste posible. Pero ¿cómo definimos el "mejor ajuste posible"? Para entender ésto, vamos a definir primero los residuos cuadrados d_i^2, como la diferencia entre el valor observado de Y (Y_i) y el valor Y predicho por la ecuación de regresión ($\hat{Y}_i$). Los residuos cuadrados d_i^2 se calculan de la siguiente forma:

$$d_i^2 = (Y_i - \hat{Y}_i)^2$$

Se calcula el cuadrado de los residuos porque estamos interesados en la magnitud, y no en el signo, de la diferencia entre el valor observado y el valor predicho. Para cualquier valor observado de Y, podríamos hacer pasar la recta de regresión por ese punto, de tal manera que minimizáramos su residuo ($d_i = 0$). Pero la línea de la recta de regresión tiene que ajustarse a todos los datos de forma colectiva, por lo que habrá que tener en cuenta la suma de todos los residuos, que es lo que se conoce como la **suma de cuadrados residual**, abreviado como **RSS** (del inglés *residual sum of squares*).

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

La línea de regresión que mejor se ajuste a los datos será aquella que minimice la suma de cuadrados residual (RSS). Minimizando la suma de cuadrados residual, aseguramos que la recta de regresión resulte en la menor diferencia entre cada valor observado de Y_i y cada valor $\hat{Y}_i$ predicho por el modelo de regresión.

Pero esto sigue sin explicar cómo elegimos la recta de regresión que mejor se ajusta. Podríamos hacer que la línea de regresión pase por el punto central $(\overline{X}, \overline{Y})$, y luego girarla hasta que encontremos una pendiente y un intercepto que minimice la suma de cuadrados residual. Esto implicaría numerosas estimaciones de la pendiente y el intercepto. Por suerte, hay una forma más fácil de estimar estos parámetros, pero antes vamos a explicar brevemente qué es la varianza y la covarianza.

3.2.2. Varianzas y covarianzas

La suma de cuadrados de una variable Y (SS_Y) es una medida de cuanta variabilidad existe en esa variable o, dicho de otra forma, de cuanto se desvía cada una de las observaciones hechas sobre la media de las observaciones.

$$SS_Y = \sum_{i=1}^{n} (Y_i - \overline{Y}_i)^2$$

Si dividimos esta suma por (n-1) obtenemos la formula de la **varianza** (s_Y^2):

$$s_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y}_i)^2$$

Si en lugar de tener una única variable Y, tenemos dos variables X e Y, en lugar de la suma de cuadrados de una variable, podemos definir la suma de sus productos (SS_{XY}) de la siguiente forma:

$$SS_{XY} = \sum_{i=1}^{n} (X_i - \overline{X}_i)(Y_i - \overline{Y}_i)$$

Y la **covarianza** de la muestra (s_{XY}):

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_i)(Y_i - \overline{Y}_i)$$

La varianza, al ser una suma de cuadrados, siempre es un número positivo. Sin embargo, esto no es necesariamente cierto para la covarianza. Si valores altos de X se asocian con valores altos de Y, entonces la suma de sus productos generará una covarianza grande. Si por el contrario no existe una relación clara entre X e Y, ocurrirá que algunos valores altos de X estarán asociados con valores pequeños o incluso negativos de Y. Esto generará al final una colección muy heterogénea de términos de covarianza, algunos con símbolo positivo y otros con símbolo negativo. La suma de todos estos términos estará muy próxima a cero.

Vamos a calcular la varianza y la covarianza para nuestro ejemplo anterior:

```
> n <- dim(cars)[1]
> SSy <- sum((cars$dist - meanY)^2) # Suma de cuadrados de Y
> s.y2 <- SSy/(n-1) # Varianza
> SSxy <- sum((cars$dist - meanY)*(cars$speed - meanX)) # Suma de productos de X e Y
> s.xy <- SSxy/(n-1) # Covarianza
> s.y2; s.xy
```

```
[1] 664.0608
```

```
[1] 109.9469
```

La mayor parte de los términos de la covarianza son positivos.

```
> (cars$dist - meanY)*(cars$speed - meanX)
```

```
 [1] 467.172 375.972 327.432 176.232 199.652 211.072 134.892  91.692  48.492
[10] 114.312  65.912  98.532  78.132  64.532  50.932  40.752  21.552  21.552
[19]  -7.248  23.772   9.772 -23.828 -51.828   9.192   6.792  -4.408  -6.588
[28]  -1.788 -17.568  -4.768  11.232  -2.548  33.852  85.852 106.652 -25.128
[37]  10.872  90.072 -50.508  23.092  41.492  59.892  96.692 151.932  83.752
[46] 232.372 421.572 430.172 662.372 403.392
```

Intuitivamente, esto debería de estar relacionado con la pendiente de la recta de regresión, ya que describe la relación (positiva o negativa) entre la variación en la variable X y la variación en la variable Y.

3.2.3. Estimadores por mínimos cuadrados

Habiendo definido qué es la covarianza, podemos ahora estimar los parámetros de la recta de regresión que minimizan la suma de cuadrados residual.

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2} = \frac{SS_{XY}}{SS_x}$$

dónde la suma de cuadrados de X (SS_X) es:

$$SS_X = \sum_{i=1}^{n}(X_i - \overline{X})$$

Utilizaremos el símbolo $\hat{\beta}_1$ para designar nuestra estima de la pendiente, y para distinguirlo de β_1, que es el verdadero valor del parámetro[2]. Por tanto, la pendiente será la covarianza de X e Y, escalada por la varianza de X. Como el denominador (n-1) es idéntico para los cálculos de s_{XY} y s_X^2, la pendiente puede expresarse también como el ratio entre la suma de productos (SS_{XY}) y la suma de cuadrados de X (SS_x). Para nuestros datos anteriores, tendríamos la siguiente estimación de la pendiente:

[2]Hay que tener en cuenta que β_1 sólo tiene un valor verdadero en el contexto de la estadística clásica (frecuentista). En un análisis Bayesiano, los parámetros mismos son vistos como una muestra aleatoria de una distribución de posibles parámetros.

```
> s.x2 <- sum((cars$speed - meanX)^2)/(n-1)
> B1 <- s.xy/s.x2
> B1

[1] 3.932409
```

Y, como observamos, se trata del mismo valor que obteníamos cuando usábamos la función `lm()`. Este valor indicaría que por cada incremento unitario en la velocidad (expresada en millas/hora), tendríamos un incremento estimado de la distancia de frenado de 3.93 pies.

Para calcular el intercepto en la ecuación sólo hay que tener en cuenta que la línea de regresión ha de pasar a través del centro de la nube de puntos, definida por $(\overline{X}, \overline{Y})$. Esto permite resolver la siguiente ecuación.

$$\hat{\beta_0} = \overline{Y} - \hat{\beta_1}\overline{X}$$

que para nuestro ejemplo, se calcularía en R de la siguiente forma:

```
> B0 <- meanY-(B1*meanX)
> B0

[1] -17.57909
```

El intercepto coincide exactamente con el valor estimado utilizando la función `lm()` en el apartado 3.1.

Todavía nos quedaría un último parámetro por estimar: **el término error** (ε_i). El error tiene una distribución normal con media 0 y varianza σ^2. ¿Cómo podemos estimar σ^2? Lo primero que hay que observar es que cuanto más pequeño sea σ^2, los datos estarán más próximos a la recta de regresión. Si $\sigma^2 = 0$ entonces no habrá desviación con respecto a las predicciones, es decir, que todos los datos caerán sobre la recta de regresión. Esta descripción es muy similar a la de la suma de cuadrados residuales (RSS), que mide la desviación cuadrada de cada observación con respecto al valor predicho por el modelo. Recordemos que la varianza de la muestra mide la desviación promedio de cada observación con respecto a la media. De forma similar, nuestra estima de la **varianza del término error (o varianza residual de la muestra)** es la desviación promedio de cada observación con respecto al valor predicho.

$$\hat{\sigma}^2 = \frac{RSS}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n}\left[Y_i - (\hat{\beta_0} + \hat{\beta_1}\hat{X}_i)\right]^2}{n-2}$$

La raíz cuadrada de la varianza del término error, $\hat{\sigma}$, es el **error estándar de la regresión**. Fíjate que en el denominador de la fórmula utilizamos (n-2) en vez de (n-1), como hacíamos antes en el caso de la varianza de la muestra. El denominador indica el número de grados de libertad, es decir, el número de piezas de información independientes utilizadas en el cálculo de la varianza. En este caso, ya hemos utilizado dos grados de libertad para estimar el intercepto y la pendiente de la recta de regresión. Para nuestro ejemplo, la varianza residual, la varianza residual de la muestra y el error estándar de la regresión se calcularía manualmente de la siguiente forma:

```
> RSS <- sum((cars$dist - (B0 + B1*cars$speed))^2)
> RMS <- RSS/(n-2)
> sterror <- RMS^0.5
> RMS

[1] 236.5317

> sterror

[1] 15.37959
```

El valor de la varianza residual de la muestra es lo que se denomina cuadrados medios residuales (RMS) en la tabla anova, que se obtendría en R con la función `anova()`. Y la varianza residual (total) es el equivalente a las sumas de cuadrados (RSS).

```
> anova(lm.cars)

Analysis of Variance Table

Response: dist
          Df Sum Sq Mean Sq F value   Pr(>F)
speed      1  21186 21185.5  89.567 1.49e-12 ***
Residuals 48  11354   236.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Como vemos, los números coinciden perfectamente.

3.2.4. Componentes de la varianza y el coeficiente de determinación

Una técnica fundamental en los análisis paramétricos es la de la partición de la suma de cuadrados en diferentes componentes. Empezando con los datos en bruto, considera que la suma de cuadrados de la variable Y (SS_Y) representa la variación total que estamos intentando particionar.

Uno de los componentes de esta variación total es el error aleatorio. Esta variación no puede ser atribuida a ninguna fuente específica y se estima a partir de la suma de cuadrados residual (RSS). La variación restante en Y_i no es aleatoria. Algunos valores de Y_i son altos porque están asociados con valores altos de X_i. La fuente de esta variación queda expresada en la relación de regresión $Y_i = \beta_0 + \beta_1 X_i$. De esta forma, conociendo la variación total (SS_Y) y la varianza residual (RSS) podemos calcular la varianza atribuida al modelo de regresión de la siguiente forma:

$$SS_{reg} = SS_Y - RSS$$

O expresado de otra forma, la varianza total es la suma de la varianza explicada por el modelo y la varianza residual.

$$SS_Y = SS_{reg} + RSS$$

Para nuestro ejemplo de la distancia de frenado, la varianza total es SSy y la varianza residual es res.var (las calculamos anteriormente). Luego la varianza explicada por el modelo sería:

```
> SSreg <- SSy - RSS
> SSreg
```

```
[1] 21185.46
```

Un índice natural que describe la importancia relativa de la regresión frente a la variación residual es el **coeficiente de determinación**, r^2:

$$r^2 = \frac{SS_{reg}}{SS_Y} = \frac{SS_{reg}}{SS_{reg}+RSS}$$

El coeficiente de determinación nos dice que proporción de la variación en la variable Y puede ser atribuida a la variación en la variable X por medio de la recta de regresión. Esta proporción va de 0.0 a 1.0. Cuanto mayor sea este valor mejor será el ajuste de los datos a la recta de regresión. Para nuestro ejemplo anterior, r^2 se calcularía de la siguiente forma:

```
> r2 <- SSreg/SSy
```

Es importante recordar que la relación causal entre X e Y es una hipótesis que el investigador propone de forma explícita. El coeficiente de determinación, por muy alto que sea, no confirma una relación causa-efecto entre dos variables.

Un estadístico asociado al coeficiente de determinación es el **coeficiente de correlación**, r, que se calcula como la raiz cuadrada del coeficiente de determinación. El signo de r indica cómo es la relación entre X e Y, si positiva o negativa.

3.2.5. Test de hipótesis

Hasta el momento hemos aprendido como fijar una línea recta para datos continuos de X e Y, y cómo utilizar el criterio de mínimos cuadrados para estimar la pendiente, el intercepto, y la varianza de la línea de regresión. El siguiente paso es testar hipótesis sobre la línea de regresión ajustada. Recordemos que los cálculos de mínimos cuadrados nos proporcionan estimaciones $(\hat{\beta_0}, \hat{\beta_1}, \hat{\sigma}^2)$ de los verdaderos valores de los parámetros $(\beta_0, \beta_1, \sigma^2)$. Como hay incertidumbre sobre estas estimaciones, vamos a querer testar si algunas de las estimas de estos parámetros difieren significativamente de cero.

En particular, el supuesto que subyace a la relación causa-efecto de nuestras variables X e Y está contenido en el parámetro de la pendiente. La magnitud de β_1 mide la fuerza de la respuesta de Y a cambios en X. Nuestra hipótesis nula es que β_1 no es diferente de cero. Si no podemos rechazar la hipótesis nula, entonces no tenemos evidencias para establecer una relación entre las variables X e Y. Las hipótesis nula y alternativa se formularían de la siguiente forma:

$$\beta_1 = 0 \text{ (Hipótesis nula)}$$

$$\beta_1 \neq 0 \text{ (Hipótesis alternativa)}$$

Para comprobar la hipótesis nula se deben de organizar los datos en la tabla del análisis de la varianza (ANOVA). Aunque una tabla ANOVA se asocia de forma natural con el análisis de la varianza (sección4), la partición de la suma de cuadrados es común al ANOVA, a la regresión y al ANCOVA, además de a otros modelos lineales generalizados.

La tabla ANOVA tiene una serie de columnas que resumen la partición de la suma de cuadrados, como ya hemos ido viendo a lo largo de esta sección. En las filas aparecerán las diferentes fuentes de variación. Si el modelo tiene una única variable explicativa, entonces aparecerán dos filas: X y residual. Si hubiera más variables explicativas, entonces habrá tantas filas como variables haya en el modelo más la habitual de la varianza residual.

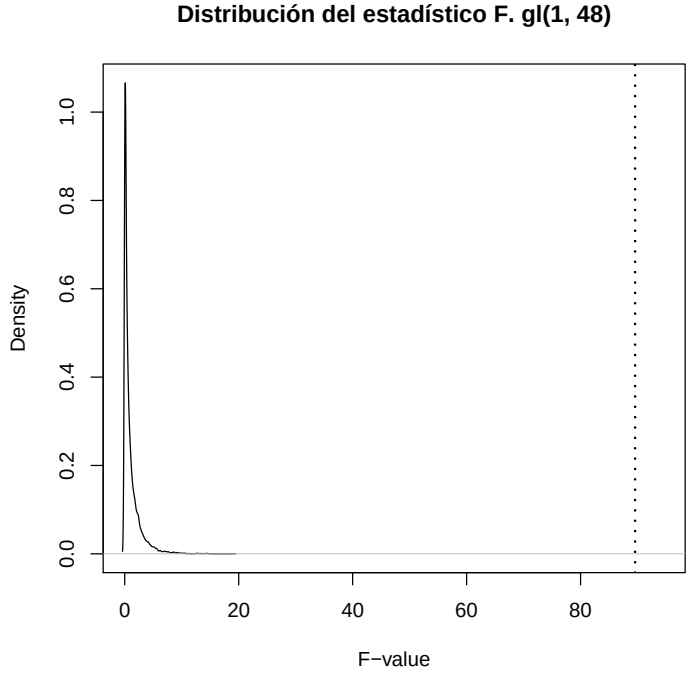
En lo que respecta a la comprobación de la hipótesis nula establecida anteriormente, ésta se lleva a cabo utilizando un estadístico denominado F (**F-ratio**). Éste se calcula diviendo los cuadrados medios del modelo por los cuadrados medios residuales, o lo que es lo mismo:

$$F - ratio = \frac{SS_{reg}/1}{RSS/(n-2)}$$

El F-ratio se compara con una distribución del estadístico F generada bajo el supuesto de que $\beta_1 = 0$. Esta distribución se genera conociendo los grados de libertad en el denominador y en el numerador. Si nuestro F-ratio queda probabilísticamente muy alejada de la distribución del estadístico F, entonces podremos decir con cierta seguridad que rechazamos la hipótesis nula, con lo que $\beta_1 \neq 0$. El **p-valor**, que se genera a partir del F-ratio conociendo la función de distribución del estadístico F, es por tanto la probabilidad de obtener un test estadístico (F-ratio) tan extremo como el observado, asumiendo que la hipótesis nula es cierta. Si el p-valor es de 0.8, quiere decir que 8 de cada 10 veces obtendremos por azar un F-ratio igual al que hemos obtenido a partir de las relaciones observadas entre X e Y. ¿Cómo saber cuando esta probabilidad es suficientemente pequeña como para rechazar la hipótesis nula? Pues bien, esto tenemos que definirlo a priori y es lo que se conoce como **nivel de significación**, α. Normalmente $\alpha = 0{,}05$. Si p-valor $<$ α entonces rechazaremos la hipótesis nula. Si por el contrario el p-valor $= \alpha$, aceptaremos la hipótesis nula, por lo que no tendremos evidencia suficiente para decir que $\beta_1 \neq 0$.

Vamos a calcular el F-ratio y ver dónde estaría situado dentro de una distribución del estadístico F asumiendo la hipótesis nula.

```
> F.ratio <- (SSreg/1)/(RSS/(n-2))
> plot(density(rf(n=10000, df1=1, df2=n-2)), xlim=c(0,F.ratio+5),
+ main="", xlab="F-value")
> title("Distribución del estadístico F. gl(1, 48)")
> abline(v=F.ratio, lwd=2, lty=3)
```

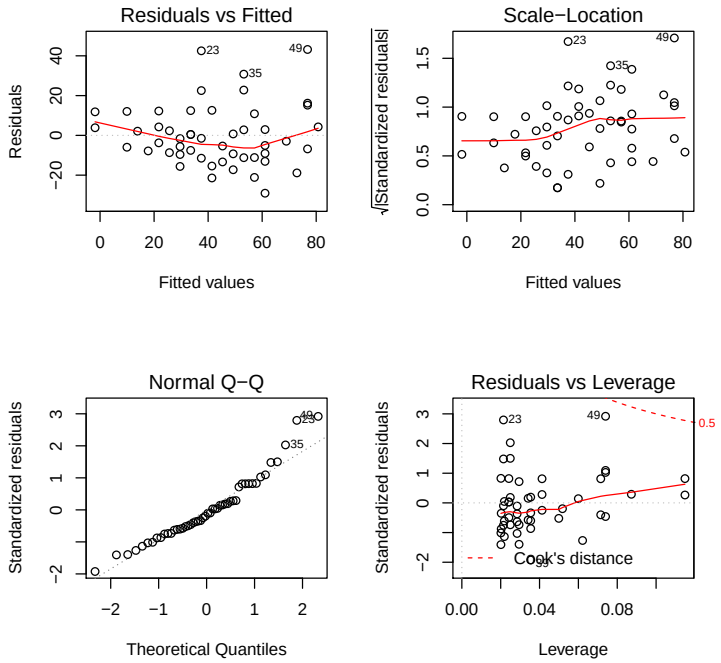


3.3. Evaluación de los supuestos del modelo: Exploración de los residuos

Una parte muy importante de la construcción de modelos estadísticos paramétricos es la comprobación de los supuestos del modelo. En concreto, nos interesa comprobar las hipótesis de normalidad y homocedasticidad (homogeneidad de varianzas).

La función plot() dibuja los gráficos de los residuos cuando el argumento principal es un objeto del tipo lm.

```
> par(mfcol=c(2,2))
> plot(lm.cars)
```



En los gráficos de los residuos vemos que los datos no son del todo normales ya que se desvían ligeramente de la diagonal en el Q-Q plot. También parece que los datos son ligeramente heterocedásticos, como indica el gráfico de residuos frente a valores predichos. Para comprobar estadísticamente (más que visualmente) si los residuos son normales podemos utilizar el test de Shapiro-Wilk (función shapiro.test()). Este test comprueba la hipótesis nula de que los datos son normales. Si rechazamos la hipótesis nula (p-valor < 0.05) podemos por tanto asumir que nuestro modelo NO es normal.

```
> shapiro.test(residuals(lm.cars))

        Shapiro-Wilk normality test

data:  residuals(lm.cars)
W = 0.94509, p-value = 0.02152
```

Por lo que podríamos asumir que nuestro modelo no es normal, además de la heterocedasticidad que se manifiesta en el gráfico de residuos frente a valores predichos. Habría que pensar por tanto en la posibilidad de transformar variables o utilizar algún otro tipo de modelo (modelos lineales generalizados, modelos no lineales, modelos aditivos generalizados, modelos no paramétricos).

Podemos también comprobar la hipótesis de normalidad con el test "RESET". Este test comprueba si X e Y se relacionan de forma lineal o, si por el contrario, existe una relación no lineal entre ellas definida por potencias de la variable respuesta, la variable explicativa o el primer componente principal de X. La hipótesis nula es que se relacionan de modo lineal. Si el p-valor es muy bajo (< 0.05) se rechaza la hipótesis nula, lo que indicaría algún tipo de relación no lineal. Para comprobar esta hipótesis podemos usar la función resettest() del paquete lmtest, que habrá que instalar previamente.

```
> library(lmtest)
> resettest(lm.cars)

        RESET test

data:  lm.cars
RESET = 1.5554, df1 = 2, df2 = 46, p-value = 0.222
```

En principio podemos asumir que hay linealidad.

3.4. Ejercicios

1. El archivo gala (accesible como archivo de datos de R dentro del paquete `faraway`) contiene información sobre la riqueza de especies en 30 islas del archipiélago de las Galápagos. Queremos saber si hay un efecto de las variables área de la isla (`Area`), elevación máxima de la isla (`Elevation`) y distancia a la isla más próxima (`Nearest`) sobre la riqueza de especies (`Species`).
   Se aconseja seguir los siguientes pasos:

   - Instalar y cargar el paquete `faraway`.
   - Representar gráficas exploratorias de la variable respuesta (`Species`) con respecto a cada una de las variables explicativas.
   - Ajustar el modelo lineal.
   - Interpretar los resultados del modelo.
   - Comprobar los supuestos del modelo.

4. Análisis de la varianza (ANOVA)

Supongamos ahora que nuestra variable explicativa no es cuantitativa sino categórica, con tres niveles: velocidad baja, velocidad media y velocidad alta.

```
> speed.cat<-cut(cars$speed, breaks=c(0, 12, 18, 26))
> levels(speed.cat)<-c("Baja", "Media", "Alta")
```

La pregunta sigue siendo la misma ¿Depende la distancia de frenado de la velocidad del coche? Lo que cambia aquí es la naturaleza de la variable explicativa y por ello el análisis se denomina análisis de la varianza en vez de análisis de regresión, aunque en esencia, ambos procedimientos son prácticamente iguales. De hecho, la función que utilizaremos para ajustar un modelo ANOVA es la misma función que se utiliza para ajustar un modelo de regresión: la función lm().

```
> lm.cars2<-lm(cars$dist~speed.cat)
> summary(lm.cars2)

Call:
lm(formula = cars$dist ~ speed.cat)

Residuals:
    Min      1Q  Median      3Q     Max
-33.467 -12.392  -1.833   8.925  54.533

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     18.200      4.717   3.859 0.000347 ***
speed.catMedia  26.500      6.240   4.247 0.000101 ***
speed.catAlta   47.267      6.670   7.086 6.05e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.27 on 47 degrees of freedom
Multiple R-squared:  0.518,        Adjusted R-squared:  0.4975
F-statistic: 25.25 on 2 and 47 DF,  p-value: 3.564e-08
```

¿Cómo se interpretan aquí los resultados? Para entender ésto, hay primero que entender cómo se ajusta el modelo en el caso de tener variables explicativas categóricas. Cuando una de las variables explicativas es categórica, el modelo entiende que hay tantos coeficientes en el modelo como niveles del factor -1. Es decir, que si el factor tiene tres niveles, el modelo tendrá dos parámetros más el punto de corte con el eje Y o Intercept. Este último recogería el valor que toma la variable respuesta cuando los dos niveles del factor para los cuales se ha estimado un coeficiente son cero, es decir, que representaría el tercer nivel del factor, no representado de manera explícita en el modelo. Por tanto, una variable categórica con tres niveles representa en realidad a tres variables explicativas que toman valores 0 ò 1. A este tipo de variables se les denomina variables *dummy*.

	Velocidad baja	**Velocidad media**	**Velocidad alta**
Coche 1	0	1	0
Coche 2	0	0	1
Coche 3	1	0	0
Coche 4	0	0	1
Coche 5	1	0	0
⋮	⋮	⋮	⋮
Coche n	0	1	0

En este caso concreto el modelo que formulamos sería de la siguiente forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

o dicho forma más específica:

$$Distancia = \beta_0 + \beta_1 Velocida.media + \beta_2 Velocidad.alta$$

Dónde velocidad media y velocidad alta tomarían valores 0 o 1 respectivamente. Por tanto, un coche que tenga una velocidad de 25 millas por hora (¡¡¡alta en los años 20!!!) tomaría un valor X1 = 0 y un valor X2 = 1, mientras que un coche con una velocidad de 8 millas por hora (velocidad baja) tomaría un valor de X1 = 0 y X2 = 0, por lo que quedaría representado en el modelo por el β_0 o Intercept.

En nuestro ejemplo, la significación alta (Pr(>|t|) < 0.05) del punto de corte y de los dos coeficientes del modelo indican que los tres niveles del factor son importantes para determinar la velocidad de frenado de un coche. Los valores estimados según el modelo serían de 18,200 pies de distancia de frenado para aquellos coches que van una velocidad baja, 44,700 pies (18,200 + 26,500*X1) para aquellos coches que van una velocidad media, y 65,466 pies para aquellos coches que van a una velocidad alta (18,200 + 47,267*X2). Podemos ver estos valores con la función fitted.values().

```
> fitted.values(lm.cars2)
```

```
       1        2        3        4        5        6        7        8
18.20000 18.20000 18.20000 18.20000 18.20000 18.20000 18.20000 18.20000
       9       10       11       12       13       14       15       16
18.20000 18.20000 18.20000 18.20000 18.20000 18.20000 18.20000 44.70000
      17       18       19       20       21       22       23       24
44.70000 44.70000 44.70000 44.70000 44.70000 44.70000 44.70000 44.70000
      25       26       27       28       29       30       31       32
44.70000 44.70000 44.70000 44.70000 44.70000 44.70000 44.70000 44.70000
      33       34       35       36       37       38       39       40
44.70000 44.70000 44.70000 65.46667 65.46667 65.46667 65.46667 65.46667
```

```
      41       42       43       44       45       46       47       48
65.46667 65.46667 65.46667 65.46667 65.46667 65.46667 65.46667 65.46667
      49       50
65.46667 65.46667
```

El coeficiente de determinación del modelo (R^2) es, en este caso, menor que en el caso anterior y, el modelo en su conjunto explicaría un 49,75 % de la variabilidad de la variable respuesta (distancia de frenado).

Otra manera de representar los resultados es considerando la significación del factor en su conjunto. Un factor es significativo si la variable respuesta en al menos uno de sus niveles es significativamente distinta del resto de los niveles. La manera de representar estos datos es a través de la tabla ANOVA, en dónde se muestra el factor como una variable única en vez de considerar los niveles del factor como variables *dummy*. Para ello se puede utilizar la función anova().

```
> anova(lm.cars2)

Analysis of Variance Table

Response: cars$dist
          Df Sum Sq Mean Sq F value    Pr(>F)
speed.cat  2  16855  8427.3  25.253 3.564e-08 ***
Residuals 47  15684   333.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

En esta tabla tenemos la significación de la variable explicativa speed.cat y la suma de cuadrados, que se utilizan para calcular el coeficiente de determinación y la variabilidad explicada por cada una de las variables en el caso de tener más de un predictor. Las funciones anova() y summary() se deben de utilizar de manera complementaria para interpretar mejor los resultados del modelo.

En el caso del ANOVA podemos además estar interesados en cómo son de distintos los niveles del factor comparados dos a dos. En este caso, sabemos que el nivel Velocidad media es significativamente superior al nivel Velocidad baja, ya que el coeficiente estimado para el último es positivo y además significativo, lo que indica que es mayor que el punto de corte o Intercept, que representa al nivel del factor Velocidad baja. Lo mismo podemos decir con respecto al nivel Velocidad alta con respecto al nivel Velocidad baja. Pero ¿son significativamente distintos entre sí los niveles del factor Velocidad media y Velocidad alta? Para comprobar ésto, se pueden utilizar el test de Bonferroni, aunque hay otros muchos tests que se pueden aplicar igualmente. El test de Bonferroni compara los niveles del factor dos a dos y ajusta el nivel de significación para disminuir el error de tipo I (rechazar hipótesis nula siendo falsa). La función pairwise.t.test() implementa este test.

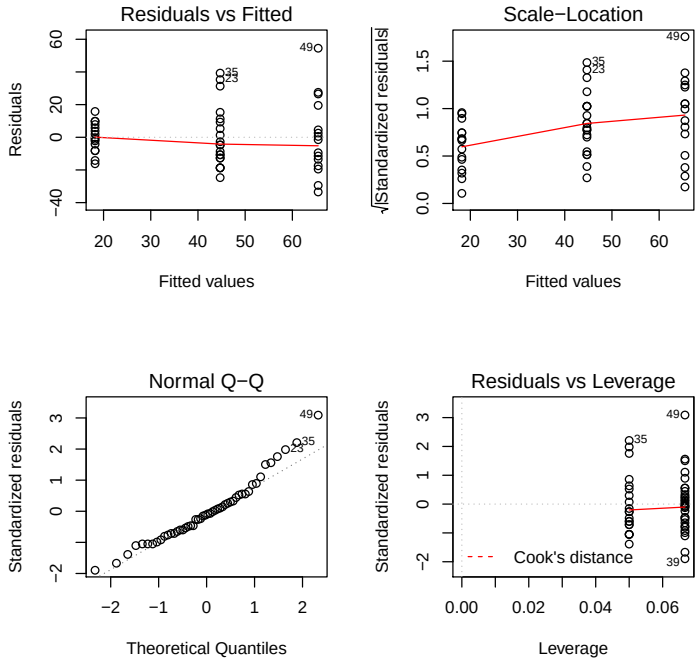
```
> pairwise.t.test(cars$dist, speed.cat, p.adjust = "bonferroni")
```

```
        Pairwise comparisons using t tests with pooled SD

data:  cars$dist and speed.cat

      Baja    Media
Media 0.0003  -
Alta  1.8e-08 0.0051

P value adjustment method: bonferroni
```

Lo que indica que, efectivamente, todos los niveles del factor son significativamente distintos ($p\text{-}valor < 0.05$) entre sí.

Faltaría, por último, evaluar los supuestos del modelo. Para ello analizaremos, como hicimos anteriormente, los gráficos de los residuos.

```
> par(mfcol=c(2,2))
> plot(lm.cars2)
```



En los gráficos de los residuos vemos fundamentalmente problemas de heterocedasticidad. Además de comprobar estadísticamente si los residuos son normales con el test de Shapiro-Wilk (función shapiro.test()), comprobaremos la hipótesis concreta de homogeneidad de varianzas con el test de Levene (función `levene.test()` del paquete `car`, que deberemos de instalar si no lo hemos hecho antes).

```
> shapiro.test(residuals(lm.cars2))

        Shapiro-Wilk normality test

data:  residuals(lm.cars2)
W = 0.95641, p-value = 0.06288

> install.packages("car", dep=T)

> library(car)
> leveneTest(dist~speed.cat, data=cars)

Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2   3.202 0.0497 *
      47
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Con lo que vemos que nuestros datos son normales, pero no homocedásticos.

4.1. Cambio del nivel de referencia en los contrastes de los niveles del factor

Cuando se hace un ANOVA, R interpreta cual es el nivel de referencia con el que comparar los coeficientes estimados para el resto de los niveles del factor. Éste es siempre el primer nivel del factor en orden alfabético. Si escribimos:

```
> levels(speed.cat)

[1] "Baja"  "Media" "Alta"
```

Vemos que “Baja” es el primer nivel en orden alfabético, y éste será el que use R como nivel de referencia. Esto sólo tiene importancia a la hora de interpretar los valores de los coeficientes estimados para los otros niveles, que si son positivos querrá decir que incrementan la respuesta con respecto al nivel “Baja” y si son negativos se interpretará como que disminuyen la respuesta con respecto a este mismo nivel.

En ocasiones, nos puede interesar utilizar un nivel de referencia que no es el que selecciona R por defecto ¿cómo cambiamos ésto? Muy fácil. La función relevel() nos permite hacerlo de la siguiente forma:

```
> speed.cat <- relevel(speed.cat, ref="Media")
> levels(speed.cat)

[1] "Media" "Baja"  "Alta"
```

4.2. Ejercicios

1. El archivo InsectSprays (accesible como archivo de datos de R) contiene información sobre 72 parcelas experimentales que han sido sometidas a 6 tipos de insecticidas distintos. La variable respuesta es número de insectos recogidos en trampas de insectos tras aplicar el tratamiento (count). La variable explicativa es el tipo de tratamiento aplicado (spray). ¿Qué sprays son más efectivos?
   Se aconseja seguir los siguientes pasos:

   - Representar los datos (count) en función del tipo de spray (gráfico de cajas).
   - Ajustar el modelo lineal.
   - Realizar comparaciones múltiples de los niveles del factor dos a dos.
   - Interpretar los resultados.
   - Comprobar los supuestos del modelo.

5. Análisis de la covarianza (ANCOVA)

Una vez entendidos los fundamentos de la regresión simple y el ANOVA unifactorial, la interpretación de modelos con más variables explicativas es simplemente una extensión de lo visto hasta el momento, incluso en el caso de que se combinen variables explicativas continuas y categóricas. Tal es el caso del ANCOVA o análisis de la covarianza.

Tomemos como ejemplo un experimento realizado con la planta herbácea *Echinochloa crus-galli* en Norteamérica (Potvin *et al.* 1990) en dónde se pretende ver el efecto que distintas variables tienen sobre la captación de CO_2 por parte de esta planta. En concreto, se pretende investigar si plantas sometidas a distintas concentraciones de CO_2 (conc) captan o no la misma cantidad de este compuesto (uptake) y, además, interesa ver qué efecto tienen dos tratamientos distintos (enfriamiento de la planta por la noche vs. no enfriamiento) a los que se somete la planta (Treatment) sobre su capacidad de fijación de CO_2. Estos datos están contenidos en el archivo de datos CO_2[3].

```
> str(CO2)
```

Las hipótesis nulas que vamos a comprobar son, en principio, dos:

H_{0_A}: No hay una relación significativa entre la captación de CO_2 por parte de la planta y la concentración atmosférica de este compuesto (la pendiente es nula).

[3] Aunque los datos originales fueron tomados sobre un diseño de medidas repetidas (Potvin *et al.* 1990), para este ejemplo asumiremos que las muestras representan a individuos distintos y son, por tanto, independientes.

$H0_{0_B}$: No hay diferencias en la captación de CO_2 entre plantas sometidas a distintos tratamientos.

El modelo teórico que se plantea sería por tanto el siguiente:

uptake ~ conc + Treatment

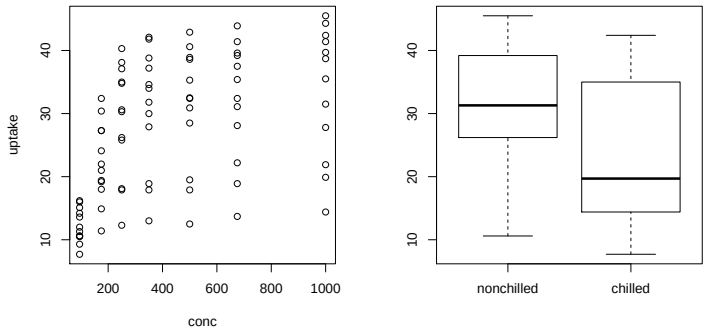
Pero el modelo estadístico subyacente sería este otro:

uptake ~ C0 + C1*conc + C2*Treatment_2

dónde C0, C1 y C2 serían los coeficientes del modelo y el efecto del Tratamiento 1 quedaría representado en el término C0.

Antes de empezar es recomendable explorar los datos.

```
> par(mfrow=c(1,2))
> plot(uptake ~ conc, data = CO2)
> boxplot(uptake ~ Treatment, data = CO2)
```



A primera vista parece que existe una relación positiva, aunque no del todo clara, entre la fijación de CO_2 y la concentración atmosférica de dicho compuesto. También parece que hay alguna diferencia entre los dos tratamientos. El siguiente paso es llevar a cabo un análisis de la covarianza para ver si estas diferencias que se observan a primera vista son estadísticamente significativas o no lo son. Una vez más, utilizaremos la función lm().

```
> CO2.model<-lm(uptake~Treatment+conc, data=CO2)
```

Para obtener información adicional sobre los coeficientes del modelo, así como el R^2, utilizaremos el comando summary().

```
> summary(CO2.model)
```

```
Call:
lm(formula = uptake ~ Treatment + conc, data = CO2)

Residuals:
    Min      1Q  Median      3Q     Max
-19.401  -7.066  -1.168   7.573  17.597

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      22.930052   1.989746  11.524  < 2e-16 ***
Treatmentchilled -6.859524   1.944840  -3.527 0.000695 ***
conc              0.017731   0.003306   5.364 7.55e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.912 on 81 degrees of freedom
Multiple R-squared:  0.3372,        Adjusted R-squared:  0.3208
F-statistic:  20.6 on 2 and 81 DF,  p-value: 5.837e-08
```

Para obtener la tabla ANOVA con la suma de cuadrados, los F, y los niveles de significación del factor o factores, utilizaremos el comando anova().

```
> anova(CO2.model)

Analysis of Variance Table

Response: uptake
          Df Sum Sq Mean Sq F value    Pr(>F)
Treatment  1  988.1  988.11  12.440 0.0006952 ***
conc       1 2285.0 2284.99  28.767  7.55e-07 ***
Residuals 81 6433.9   79.43
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

¿Cómo interpretamos estos resultados? Al igual que ocurría con el ANOVA, se estiman tantos coeficientes para el factor como niveles – 1. El nivel del factor que no se estima queda incluido en el punto de corte del modelo (Intercept). Los niveles de significación nos indican que el coeficiente estimado para uno de los tratamientos (Treatmentchilled) es significativamente menor que cero. El Intercept también es significativo, lo que indica que el otro tratamiento (Treatmentnonchilled) es significativamente distinto de cero y, en este caso, tiene un efecto positivo sobre la fijación de CO_2 (Estimate = 22.019163). Podemos utilizar el gráfico de cajas (boxplot) para ayudarnos a interpretar estos resultados.

Lo segundo que vemos es que el modelo en su conjunto es significativo (p-value: 5.837e-08) y que explica cerca del 32 % de la variabilidad en la fijación de CO_2 de la planta (adjusted R-squared: 0.3208).

Como en este caso el factor sólo tiene dos niveles, no hace falta hacer comparaciones múltiples. Al ser significativo el efecto del factor ya sabemos

que uno será mayor que el otro. Los coeficientes estimados para el modelo nos dan esta información, como ya hemos visto.

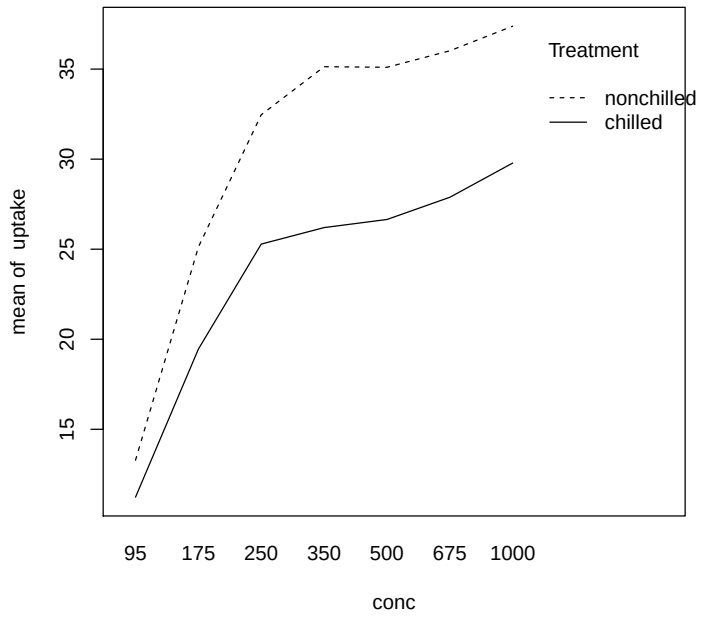
5.1. Homogeneidad de pendientes

En el caso del ANCOVA, es necesario cumplir un supuesto más además de los supuestos estadísticos ya vistos para la regresión y el ANOVA: **la homogeneidad de pendientes**. Las pendientes de las rectas de regresión entre X e Y dentro de cada uno de los niveles del factor tienen que ser paralelas para poder estimar con precisión los efectos principales del factor.

La hipótesis nula H_0 de que las pendientes entre grupos son iguales, $\beta_1 = \beta_2 = ... = \beta_n$, se puede testar estadísticamente examinando si la interacción entre el factor y la variable continua es igual a 0, es decir, si no existe interacción. Una interacción entre un factor y una variable continua se interpreta como un cambio en la pendiente de la recta de regresión entre la variable respuesta y la covariable en los distintos niveles del factor. Para el ejemplo anterior, un término de interacción en el modelo significaría que la respuesta de captación de CO_2 de la planta frente a las concentraciones atmosféricas de CO_2 depende del tipo de tratamiento al que han sido sometidas. Un caso extremo de esta interacción sería, por ejemplo, que mientras las plantas sometidas al tratamiento `nonchilled` reaccionan positivamente a las concentraciones de CO2 atmosférico, las plantas sometidas al tratamiento `chilled` reaccionan negativamente a las mismas. En cambio, si no hay interacción, esto indica que las pendientes son iguales, y por tanto, los efectos principales del factor (estimados sobre el intercepto de las rectas de regresión) son creibles independientemente del valor de la covariable. **Cuando no hay interacción, siempre hay que reajustar el modelo eliminando este término**. De otra manera, la interacción (sea significativa o no) podría afectar los cálculos del estadístico F y los p-valores para los efectos principales (factor y covariable).

Una manera de explorar la homogeneidad de pendientes visualmente es utilizando la función gráfica `interaction.plot()`.

```
> attach(CO2)
> interaction.plot(x.factor=conc, trace.factor=Treatment, response=uptake)
```



Para testar la homogeneidad de pendientes, se debe incluir el término interacción entre el factor y la covariable en el modelo estadístico. Para ello se pueden utilizar dos sintaxis distintas.

```
> CO2.model2<-lm(uptake~Treatment+conc+Treatment:conc, data=CO2)
> CO2.model2<-lm(uptake~Treatment*conc, data=CO2)
```

El operador **:** específica la interacción entre dos términos del modelo pero no se refiere al efecto de cada uno de los términos indivuales sobre la variable respuesta, mientras que el operador ***** se refiere tanto a los términos simples como a la interacción entre ellos. Ambas fórmulas son equivalentes.

Ahora obtenemos la tabla ANOVA con la suma de cuadrados, los F, y los niveles de significación del factor.

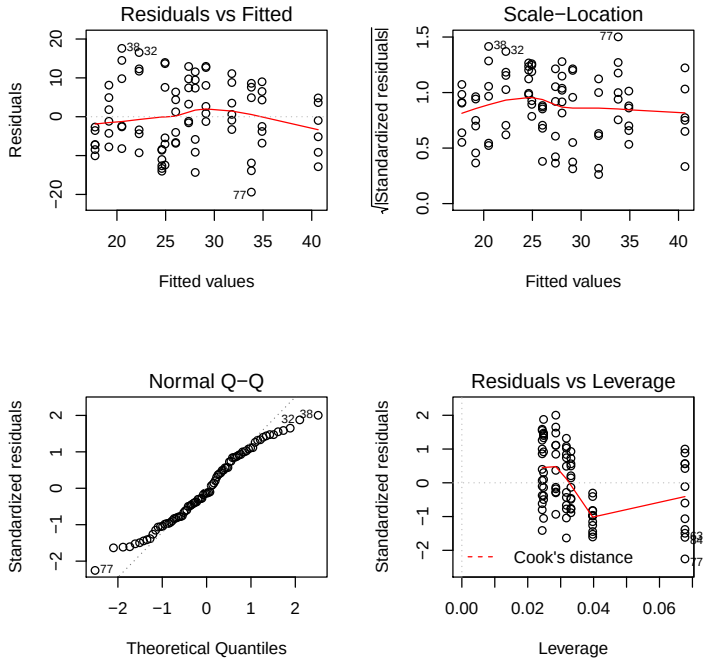
```
> anova(CO2.model2)

Analysis of Variance Table

Response: uptake
               Df Sum Sq Mean Sq F value    Pr(>F)
Treatment       1  988.1  988.11 12.3476 0.0007297 ***
```

```
conc            1 2285.0 2284.99 28.5535 8.377e-07 ***
Treatment:conc  1   31.9   31.87  0.3983 0.5297890
Residuals      80 6402.0   80.02
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Vemos que el factor y la covariable son significativos, pero la interacción entre ambos no lo es, como parecía indicar el gráfico de interacciones. Por lo tanto, debemos quedarnos con el modelo anterior, no sólo porque tiene menos parámetros y explica prácticamente la misma cantidad de variabilidad, sino también porque el modelo ANCOVA asume homogeneidad de pendientes e incluir un término interacción estaría violando este supuesto.

Por último, deberíamos comprobar el resto de los supuestos del modelo utilizando para ello los gráficos de los residuos (opcionalmente podríamos también testar las hipótesis concretas de normalidad y homocedasticidad).

```
> par(mfcol=c(2,2))
> plot(CO2.model)
```



5.2. ¿Qué ocurre si la interacción es significativa?

La homogeneidad de pendientes no debería ser considerada simplemente como un supuesto del modelo ANCOVA. Las interacciones entre factores y covariables normalmente representan efectos de considerable interés biológico.

Las diferencias entre las pendientes de las rectas de regresión indican que los tratamientos afectan la relación entre la variable respuesta Y y la covariable. Explicar esto puede ser tan interesante o más que explicar los efectos principales.

Cuando las pendientes son claramente heterogéneas (la interacción es significativa en el modelo) se pueden hacer varias cosas, dependiendo de cual sea la cuestión de interés para el investigador.

Si la interacción es lo que más nos interesa, nos vale con quedarnos con el modelo ANCOVA con interacción y concluir que la respuesta de la covariable es diferente entre grupos, pero sin concluir nada sobre los efectos principales.

Si nos interesa el efecto de la covariable entonces lo más fácil sería ajustar tantos modelos de regresión como niveles del factor haya. Se pueden proyectar todas las rectas de regresión en una única gráfica para ver las diferencias entre ellas. Otras opciones son discutidas en Quinn & Keough (2002).

5.3. Ejercicios

El arreglo de datos restauracion (http://tinyurl.com/restauracion) contiene información sobre un experimento de restauración llevado a cabo en taludes de carretera. El objetivo es ver si la exposición a la luz y la sequía estival afectan la producción de biomasa leñosa y, por tanto, a la capacidad del matorral para fijar el suelo en taludes de carretera. Para comprobar el efecto de estos dos factores se ha diseñado un experimento en dónde se han delimitado parcelas de 2 x 2 m en 91 taludes con características similares en cuanto a pendiente, exposición, clima, etc. El experimento se ha aleatorizado y se ha asignado a cada talud unas condiciones de exposición a la luz y cantidad de agua disponible (simulando la sequía estival) diferentes. En concreto, para el factor "exposición a la luz" se han definido a priori 3 niveles: nivel 1 (100 % de luz), nivel 2 (80 % de luz), nivel 3 (50 % de luz); y para el factor "sequia estival" se han definido dos niveles: sequía estival (condiciones permanentes de sequía durantes los meses de julio y agosto) y lluvia estival (riego una vez a la semana durante los meses de julio y agosto). Tras los meses de verano se ha cortado toda la vegetación leñosa, se ha secado y se ha pesado, teniendo así una estimación de la biomasa leñosa producida durante este periodo.

- ¿Tiene la luz un efecto sobre la producción de biomasa leñosa?
- ¿Tiene la disponibilidad de agua (sequía estival) un efecto sobre la producción de biomasa leñosa?
- ¿Existe una interacción entre ambos factores? De ser así ¿cómo se interpreta ésta? ¿qué gráficos puedes usar para ayudarte a interpretar esta interacción en caso de que sea significativa?

6. Problemas de colinealidad: Reducción de variables

Cuando tenemos modelos con un gran número de variables explicativas puede ocurrir que dichas variables sean redundantes o, lo que es lo mismo, que

muchas de estas variables estén correlacionadas entre sí. Al introducir variables correlacionadas en un modelo, el modelo se vuelve inestable. Por un lado, las estimaciones de los parámetros del modelo se vuelven imprecisas y los signos de los coeficientes pueden llegar incluso a ser opuestos a lo que la intuición nos sugiere. Por otro, se inflan los errores estándar de dichos coeficientes por lo que los test estadísticos pueden fallar a la hora de revelar la significación de estas variables.

Por tanto, siempre que tengamos varias variables explicativas (sobretodo cuando tenemos un gran número de ellas), es importante explorar la relación entre ellas previamente al ajuste del modelo estadístico.

Tomemos como ejemplo datos sobre las características climáticas predominantes en la región de origen de 54 especies del género *Acacia*. Dichas características podrían explicar el número de inflorescencias que desarrollan estas plantas, lo que a su vez podría determinar el carácter invasivo de las especies. Los datos están disponibles en `http://www.escet.urjc.es/biodiversos/R/acacia.txt`.

```
> acacia <- read.table(url("http://www.escet.urjc.es/biodiversos/R/acacia.txt"), header=T,
> names(acacia)

 [1] "Especie"                "Invasora"               "Inflor"
 [4] "Tm_anual"               "Tmax_mes_calido"        "Tmin_mes_frio"
 [7] "Rango_T.diurno"         "Rango_T_anual"          "P_anual"
[10] "P_mes_humedo"           "P_mes_seco"             "Estacionalidad_T"
[13] "Estacionalidad_P"       "Altitud"                "P_cuarto_seco"
[16] "Max_Tm_anual"           "Max_Tmax_mes_calido"    "Max_Tmin_mes_frio"
[19] "Max_Rango_T.diurno"     "Max_Rango_T_anual"      "Max_P_anual"
[22] "Max_P_mes_humedo"       "Max_P_mes_seco"         "Max_Estacionalidad_T"
[25] "Max_Estacionalidad_P"   "Max_Altitud"            "Max_P_cuarto_seco"
[28] "Min_Tm_anual"           "Min_Tmax_mes_calido"    "Min_Tmin_mes_frio"
[31] "Min_Rango_T.diurno"     "Min_Rango_T_anual"      "Min_P_anual"
[34] "Min_P_mes_humedo"       "Min_P_mes_seco"         "Min_Estacionalidad_T"
[37] "Min_Estacionalidad_P"   "Min_Altitud"            "Min_P_cuarto_seco"
[40] "Rango_Tm"               "Rango_Tmax_mes_calido"  "Rango_Tmin_mes_frio"
[43] "Rango_P_anual"          "Rango_P_mes_humedo"     "Rango_P_mes_seco"
[46] "Rango_Estacionalidad_T" "Rango_Estacionalidad_P" "Rango_Altitud"
```

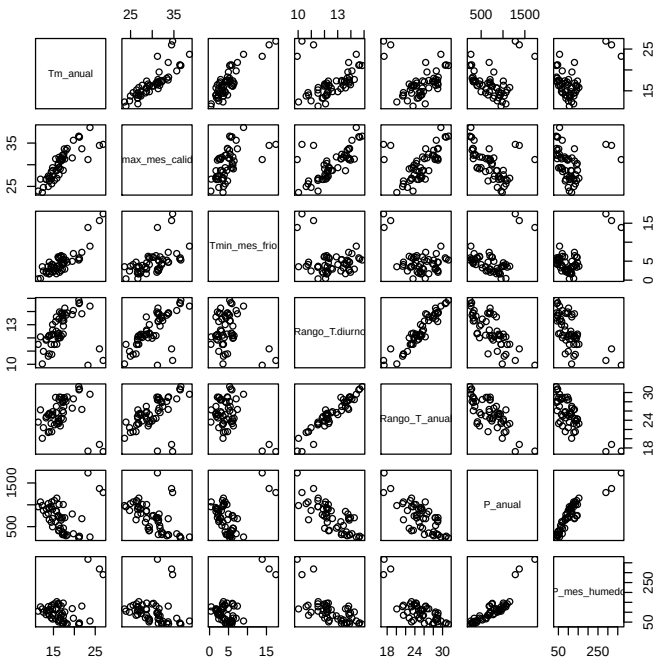
Imaginemos que queremos construir un modelo lineal en dónde el número de inflorescencias quede en función de las variables climáticas. Para ello, antes de construir el modelo deberemos comprobar la correlación entre las variables explicativas. Cómo hay un gran número de ellas (45), sólo vamos a explorar la correlación entre las 7 primeras a modo de ejemplo.

```
> acacia<-na.omit(acacia)
> round(cor(acacia[, c(4:10)]), 3)

               Tm_anual Tmax_mes_calido Tmin_mes_frio Rango_T.diurno
Tm_anual          1.000           0.844         0.855          0.292
```

```
Tmax_mes_calido     0.844           1.000         0.547          0.700
Tmin_mes_frio       0.855           0.547         1.000         -0.185
Rango_T.diurno      0.292           0.700        -0.185          1.000
Rango_T_anual       0.050           0.530        -0.420          0.946
P_anual            -0.068          -0.527         0.188         -0.769
P_mes_humedo        0.358          -0.125         0.604         -0.638
                Rango_T_anual P_anual P_mes_humedo
Tm_anual                0.050  -0.068        0.358
Tmax_mes_calido         0.530  -0.527       -0.125
Tmin_mes_frio          -0.420   0.188        0.604
Rango_T.diurno          0.946  -0.769       -0.638
Rango_T_anual           1.000  -0.759       -0.746
P_anual                -0.759   1.000        0.880
P_mes_humedo           -0.746   0.880        1.000
```

La función na.omit() la utilizamos para eliminar las filas que tengan datos faltantes (NA). También podríamos utilizar la función gráfica pairs().

```
> pairs(acacia[, c(4:10)])
```



¿Qué variables están más correlacionadas entre sí (p.e. $|r > 0.8|$)? Dado que existe correlación alta entre algunas variables, la solución podría ser hacer una selección de las variables que estén menos correlacionadas entre sí. En el caso de que haya mucha colinealidad entre las variables explicativas, o de que haya muchas variables explicativas, como en este caso, otra opción es hacer un

análisis de componentes principales (PCA) para reducir la dimensionalidad de los datos. El PCA resume en vectores ortogonales (es decir, independientes) la variabilidad representada por un conjunto de variables. El ejemplo más típico son las variables climáticas, en donde existe casi siempre una alta colinealidad. Un conjunto de 25 o 30 variables climáticas pueden resumirse en dos o tres ejes que representen ciertas características de los datos (por ejemplo, estacionalidad, temperatura) y que resuman una gran proporción de la variabilidad de las variables originales (a veces dos o tres ejes del PCA pueden resumir hasta un 80 % o un 90 % de la variabilidad de los datos originales).

7. Sumas de cuadrados de tipo I y III

Cuando tenemos modelos con más de una variable explicativa existen varias formas de calcular las sumas de cuadrados para cada una de ellas (es decir, la variación compartida de cada una de ellas con la variable respuesta). Las sumas de cuadrados más comúnmente utilizadas son las de tipo I y III.

Las **sumas de cuadrado de tipo I** se obtienen calculando la reducción en la suma de cuadrados residual a medida que vamos añadiendo términos al modelo de forma secuencial.

Las sumas de **cuadrados de tipo III** se denominan sumas de cuadrados marginales. Este tipo de sumas de cuadrados calculan la reducción en la suma de cuadrados residual para un efecto tras haber ajustado todos los demás efectos en el modelo. Para estimar la suma de cuadrados para cada uno de los coeficientes del modelo se hace lo siguiente: al modelo completo (una vez que hemos estimado los coeficientes del mismo) se le quita una variable y se estima la suma de cuadrados de esta variable calculando la diferencia entre la suma de cuadrados explicada del modelo completo y la suma de cuadrados explicada del modelo al que se le ha extraido dicha variable.

Es importante que tengamos en cuenta que los coeficientes estimados del modelo utilizando una suma de cuadrados de tipo I y III no cambian, lo que cambia es la variabilidad explicada por cada uno de ellos y su significación.

7.1. ¿Cuándo usar una u otra?

Existe un intenso debate entre los estadísticos sobre qué tipo de suma de cuadrados se debe de utilizar. En principio, si nuestro modelo lineal sólo contiene variables continuas (regresión), el tipo de suma de cuadrados que utilicemos no es relevante siempre y cuando no exista co-linealidad (ver sección 6) entre nuestras variables explicativas. Si existe colinealidad, aunque sea pequeña, entonces deberemos preguntarnos si existe una cierta jerarquización de los efectos sobre la variable respuesta. Por ejemplo, si tenemos un modelo en dónde queremos comprobar el efecto de la temperatura media anual y de la intensidad de uso antrópico sobre la abundancia de una especie (variable respuesta) sería lógico pensar que la variable climática va a tener un efecto regional y más general que la variable de uso antrópico, que tendría un efecto más local, por lo que el uso de una suma de cuadrados de tipo I con un orden

de entrada definido primero por las variables que tienen un efecto más regional (temperatura media anual) y luego por las variables de efecto más local (uso antrópico) tiene sentido.

Si tenemos un diseño de tipo ANOVA o ANCOVA, entonces la cosa no está tan clara y es aquí en dónde el debate se vuelve más intenso. Algunos libros dicen que si el diseño es balanceado (mismo número de casos en cada nivel del factor) entonces se debe de utilizar una suma de cuadrados de tipo I. También usaremos una suma de cuadrados de tipo I si existe un efecto bloque o una cierta anidación en el orden de entrada de las variables en el modelo, lo cual suele ser bastante frecuente en modelos ecológicos. Esto es útil porque nos permite controlar la variabilidad de determinadas variables (bloques) antes de testar las hipótesis de interés. Recordemos que para el tipo I, las sumas de cuadrados para cada uno de los efectos en el modelo pueden cambiar si cambiamos el orden de entrada de las variables en el modelo.

La suma de cuadrados de tipo III se debe de utilizar cuando no asumamos un efecto anidado, tanto para diseños balanceados como para diseños no balanceados.

Es muy importante, por tanto, pensar bien sobre las hipótesis que estamos comprobando y sobre su posible anidación en el modelo, aunque el tema de la anidación de variables también se puede (y debe) tratar con otro tipo de herramientas como son los modelos lineales mixtos.

Dicho esto, hay que tener en cuenta que, por defecto, R siempre va a calcular sumas de cuadrados de tipo I cuando usemos cualquier tipo de modelo lineal o modelo lineal generalizado, mientras que las sumas de cuadrados implementadas por defecto en otros software estadísticos como SPSS o Statistica es la de tipo III. Por tanto, si repetimos el mismo análisis con R y SPSS o Statistica y no cambiamos el tipo de suma de cuadrados, es posible que obtengamos distintos resultados.

7.2. Especificar diferentes sumas de cuadrados en R

Como ya hemos dicho, las sumas de cuadrados que R implementa por defecto son las de tipo I. Si queremos utilizar otras sumas de cuadrados podemos utilizar la función `Anova()` del paquete `car` de John Fox.

```
> install.packages("car", dep=T)
```

```
> library(car)
> lm.acacia <- lm(Inflor~ Tmin_mes_frio + Tm_anual, data=acacia)
> anova(lm.acacia)
```

```
Analysis of Variance Table

Response: Inflor
              Df Sum Sq Mean Sq F value  Pr(>F)
Tmin_mes_frio  1  231.0 230.999  4.0509 0.04955 *
```

```
Tm_anual       1    0.0   0.000  0.0000 0.99915
Residuals     50 2851.2  57.024
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> Anova(lm.acacia, type="III")

Anova Table (Type III tests)

Response: Inflor
               Sum Sq Df F value Pr(>F)
(Intercept)     85.95  1  1.5073 0.2253
Tmin_mes_frio   62.13  1  1.0895 0.3016
Tm_anual         0.00  1  0.0000 0.9992
Residuals     2851.19 50
```

8. Referencias

- Crawley, M.J. (2007). The R Book. Wiley.
- Engqvist, L. (2005). The mistreatment of covariate interaction terms in linear model analyses of behavioural and evolutionary ecology studies. Animal Behaviour 70: 967-971.
- Faraway, J.J. (2005). Linear models with R. Chapman & Hall/CRC Press, Florida, USA.
- Quinn, G.P. & Keough, M.J. (2002). Experimental design and data analysis for biologists. Cambridge University Press, Cambridge.
- Zuur, A.F., Ieno, E.N. & Smith, G.M. (2007). Analysing ecological data. Springer, New York.