

Modelos lineales en R: Regresión, ANOVA y ANCOVA

Luis Cayuela

6 de noviembre de 2009

EcoLab, Centro Andaluz de Medio Ambiente, Universidad de Granada – Junta de Andalucía, Avenida del Mediterráneo s/n, E-18006, Granada. E-mail: lcayuela@ugr.es.

Índice

1. Conceptos estadísticos básicos	3
1.1. Ejercicios	4
2. Cosas importantes antes de empezar	5
3. Como ajustar un modelo lineal en R	6
3.1. Un ejemplo de regresión simple	6
3.2. Un ejemplo de ANOVA	8
3.2.1. Ejercicios	11
3.3. Un ejemplo de ANCOVA	12
3.4. Interacción entre factores o factores y co-variables	14
3.4.1. Ejercicios	16
4. Evaluación de los supuestos del modelo: Exploración de los residuos	17
4.1. Ejercicios	18
5. Problemas de colinealidad: Reducción de variables	18

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 X_1, \dots, X_p , 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) respuesta 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.

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).

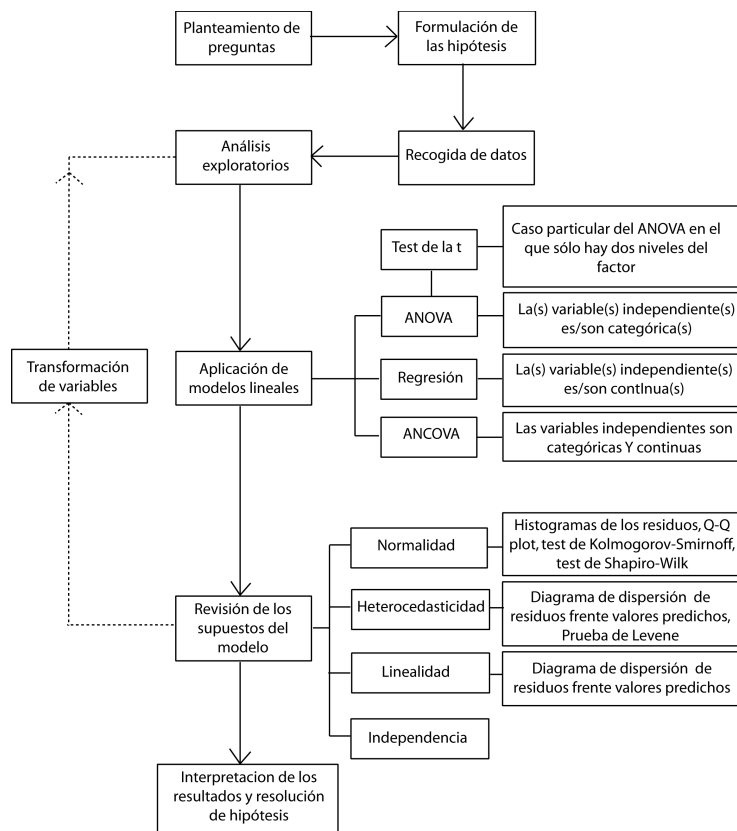


Figura 1: Esquema conceptual de los pasos que deben seguirse a la hora de ajustar un modelo lineal univariante.

4. **Homocedasticidad.** Las varianzas tienen que ser homogéneas en los distintos niveles del factor o en los diferentes intervalos de la variable respuesta.

1.1. Ejercicios

1. La siguiente tabla muestra las posibilidades univariadas y multivariadas, simples y múltiples para el ajuste de modelos lineales. Rellena con una x las casillas que correspondan para cada caso.

	Variable(s) respuesta		Variable(s) explicativa			
			Continua		Categórica	
Modelos lineales	$Y_j = 1$	$Y_j > 1$	$X_p = 1$	$X_p > 1$	$X_p = 1$	$X_p > 1$

ANÁLISIS UNIVARIADOS

Regresión simple	×		×			
Regresión múltiple						
ANOVA unifactorial						
ANOVA multifactorial						
ANCOVA						

ANÁLISIS MULTIVARIADOS

Regresión simple multivariada						
Regresión múltiple multivariada						
MANOVA unifactorial						
MANOVA multifactorial						
MANCOVA						

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. Como ajustar un modelo lineal en R

3.1. Un ejemplo de regresión simple

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 X_1, \dots, X_n . 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 donde 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: 1.490e-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 significativo¹. 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.

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 ($\Pr(>|t|) = 1.49e-12$). 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 (ver **Figura 2**). Por tanto rechazaríamos la hipótesis nula de que este coeficiente es cero y aceptamos su “influencia” sobre la variable respuesta *dist*.

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.

¹La 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) .

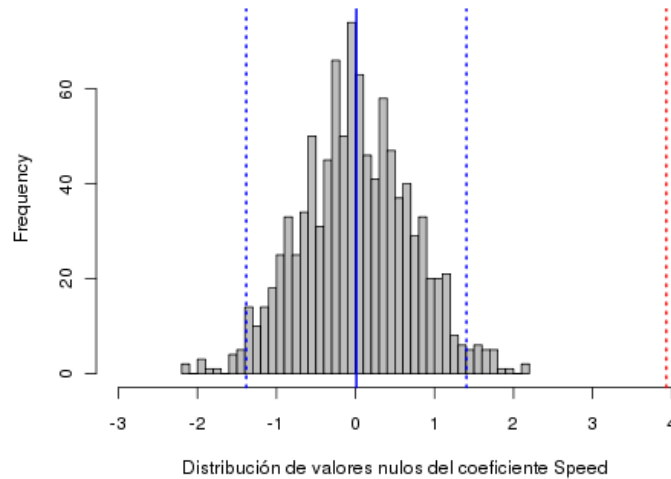


Figura 2: Distribución aleatoria de valores nulos del coeficiente estimado `speed` en un modelo lineal. La línea azul continua marca la media y las líneas azules discontinuas marcan ± 2 desviaciones estándar. Estas dos líneas acotan el 95 % de los valores de la distribución. La línea roja discontinua marca el coeficiente estimado en el modelo lineal. Por tanto, la probabilidad de que este valor esté dentro de la distribución de valores nulos es muy muy pequeña y esto permite rechazar la hipótesis nula de que el coeficiente es igual a 0.

3.2. Un ejemplo de 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 <i>n</i>	0	1	0

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

$$Y = C0 + C1 \cdot X1 + C2 \cdot X2$$

o dicho de otra forma:

$$\text{Distancia} = C0 + C1 \cdot \text{Velocidad media} + C12 \cdot \text{Velocidad alta}$$

Dónde velocidad media y velocidad alta tomarían valores 0 a 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 $C0$ 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 ($18,200 + 26,500 \cdot X1$) para aquellos coches que van una velocidad media, y 65,466 pies para aquellos coches que van a una velocidad alta ($18,200 + 47,267 \cdot X2$). 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 16854.6   8427.3   25.253 3.564e-08 ***
Residuals 47 15684.3    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 0.00030 -
Alta  1.8e-08 0.00511

```

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.

3.2.1. 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.

3.3. Un ejemplo de 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_2^2 .

```
> str(CO2)
```

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

H_{0A} : 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).

H_{0B} : 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:

$$\text{uptake} \sim \text{conc} + \text{Treatment}$$

Pero el modelo estadístico subyacente sería este otro:

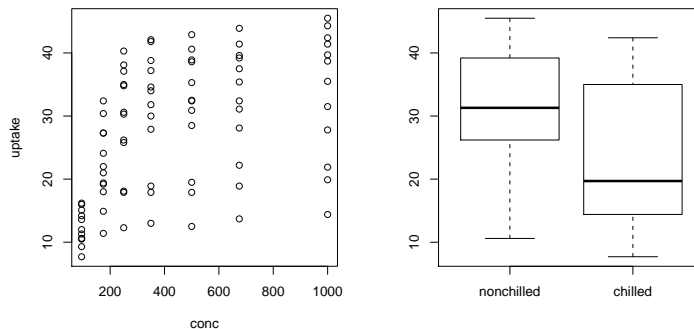
$$\text{uptake} \sim C0 + C1*\text{conc} + C2*\text{Treatment}_2$$

²Aunque 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.

dónde C_0 , C_1 y C_2 serían los coeficientes del modelo y el efecto del Tratamiento 1 quedaría representado en el término C_0 .

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    988.1  12.440 0.0006952 ***
conc      1 2285.0   2285.0  28.767 7.55e-07 ***
Residuals 81 6433.9    79.4
---
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).

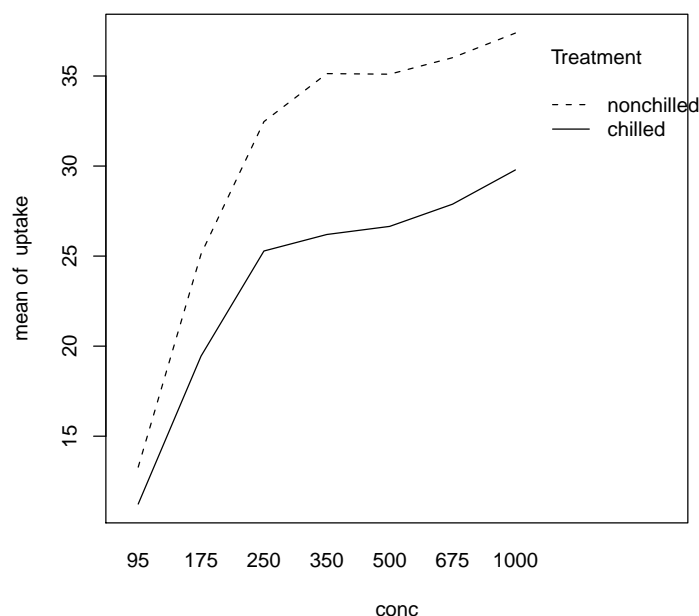
Como en este caso el factor sólo tiene dos niveles, no hace falta hacer comparaciones múltiples. Al ser significativo el efecto del factor ya sabemos que uno será mayor que el otro. Los coeficientes estimados para el modelo nos dan esta información, como ya hemos visto.

3.4. Interacción entre factores o factores y co-variables

Puede ocurrir que también estemos interesados en comprobar hipótesis específicas sobre la interacción entre distintos factores del modelo, o entre factores y co-variables. 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 CO_2 atmosférico, las plantas sometidas al tratamiento **chilled** reaccionan negativamente a las mismas. Una manera de explorar la interacción visualmente es utilizando la función gráfica `interaction.plot()`.

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



Para incluir el término interacción entre el factor y la covariable en el modelo estadístico, 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 `:` especifica la interacción entre dos términos del modelo pero no se refiere al efecto de cada uno de los términos individuales 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	988.1	12.3476	0.0007297 ***
conc	1	2285.0	2285.0	28.5535	8.377e-07 ***
Treatment:conc	1	31.9	31.9	0.3983	0.5297890
Residuals	80	6402.0	80.0		

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, podríamos quedarnos con el modelo anterior, que tiene menos parámetros y explica prácticamente la misma cantidad de varianza.

3.4.1. Ejercicios

1. El arreglo de datos `happy` del paquete `faraway` contiene información sobre la felicidad de 39 individuos (valorada del 1 al 10) y de distintas variables que podrían afectar al estado de felicidad. Queremos probar inicialmente un modelo que contenga todas las variables explicativas: `money`, `sex`, `love` y `work`. ¿Qué nos indica el modelo? ¿Qué variables son relevantes y en qué sentido lo son? Ahora construye un modelo que incluya las interacción entre `love` y `work`, `love` y `sex`, y `sex` y `work`. Los pasos que tienes que seguir son:
 - Cargar el paquete `faraway` (que ya debería de estar instalado en el ordenador).
 - Mira la estructura de la base de datos `happy`. Verás que las variables `sex`, `love` y `work` son numéricas cuando deberían ser factores. Convierte estas variables a factor usando el comando `as.factor()`. P.e. `happy$love <- as.factor(happy$love)`.
 - Genera gráficas exploratorias de la variable respuesta `happy` con cada una de las variables explicativas.
 - Ajusta un primer modelo `happy1` que incluya todos los términos independientes sin interacciones. Mira qué variables son significativas e interpreta su efecto sobre la variable respuesta.
 - Explora el efecto de la interacción entre pares de variables explicativas sobre la variable respuesta utilizando la función gráfica `interaction.plot()`.
 - Ajusta un segundo modelo `happy2` que contenga los términos independientes y sus interacciones dos a dos (sólo para los factores, es decir, sin tener en cuenta la interacción de `love`, `sex` y `work` con `money`). ¿Hay alguna interacción que sea importante? De ser así ¿cómo explicas dicha interacción?

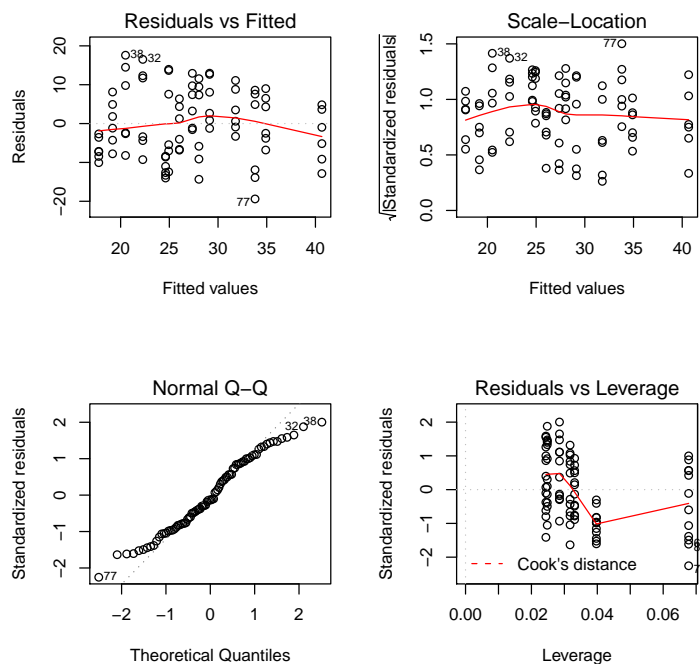
- Finalmente, reduce el modelo eliminando todos los términos que no sean significativos. Este modelo se llamará `happy3`.

4. 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(CO2.model)
```



En los gráficos de los residuos vemos que los datos no son del todo normales ya que se desvían 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 asegurarnos de que estas dos hipótesis o supuestos del modelo se cumplen podemos realizar unos test estadísticos. El test de Shapiro-Wilk (función `shapiro.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. El test de Levene (función `levene.test()` del paquete `cars`) comprueba la hipótesis nula de que la varianza en los distintos niveles del factor es homogénea.

```
> shapiro.test(residuals(CO2.model))

      Shapiro-Wilk normality test

data:  residuals(CO2.model)
W = 0.9727, p-value = 0.07073

> install.packages("car")

> library(car)
> levene.test(uptake ~ Treatment, data = CO2)
```

```
Levene's Test for Homogeneity of Variance
  Df F value Pr(>F)
group 1  1.2999 0.2576
      82
```

Por lo que podríamos asumir que nuestro modelo es normal y homocedástico y no habría necesidad de transformar las variables o buscar un modelo alternativo.

4.1. Ejercicios

1. Comprueba si el modelo `happy3` que generastes en el ejercicio anterior se ajusta a los supuestos de normalidad, homocedasticidad y linealidad.

5. 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 (sobretudo 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://tinyurl.com/yz446yz>.

```
> acacia <- read.table(url("http://tinyurl.com/yz446yz"), header = T,
+   sep = "\t", dec = ",")
> 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. Como 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)
```

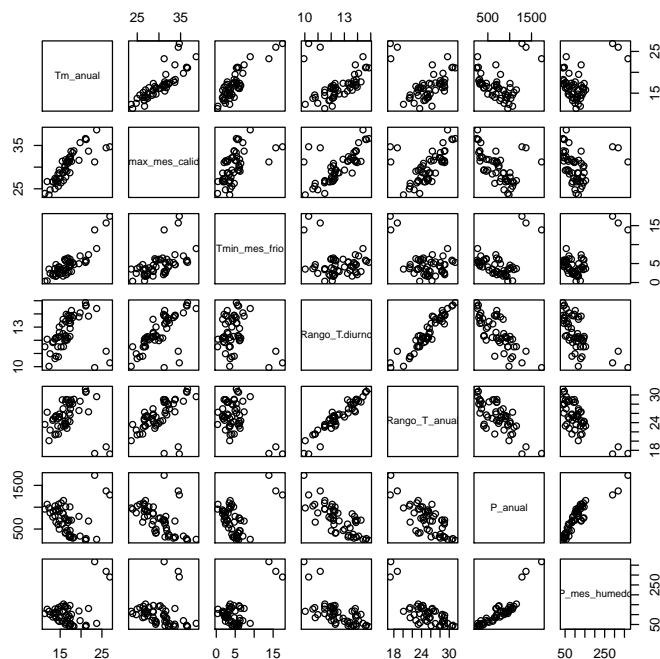
	Tm_anual	Tmax_mes_calido	Tmin_mes_frio	Rango_T.diurno
Tm_anual	1.000	0.844	0.855	0.292
Tmax_mes_calido	0.844	1.000	0.547	0.700
Tmin_mes_frio	0.855	0.547	1.000	-0.185
Rango_T.diurno	0.292	0.700	-0.185	1.000
Rango_T_anual	0.050	0.530	-0.420	0.946
P_anual	-0.068	-0.527	0.188	-0.769
P_mes_humedo	0.358	-0.125	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). El PCA se verá en más detalle en la última sesión de este curso.