Práctica 3: Regresión lineal simple y anova unifactorial

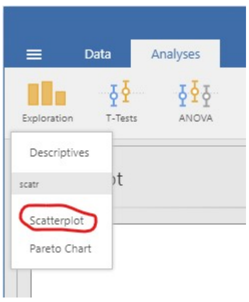
Esta práctica tiene dos partes. En la primera veremos cómo ajustar un modelo de regresión simple con jamovi y cómo una vez ajustado se llevan a cabo contrastes y se calculan intervalos para los parámetros del modelo. La segunda parte se dedica al modelo de análisis de la varianza unifactorial.

1. Regresión lineal simple

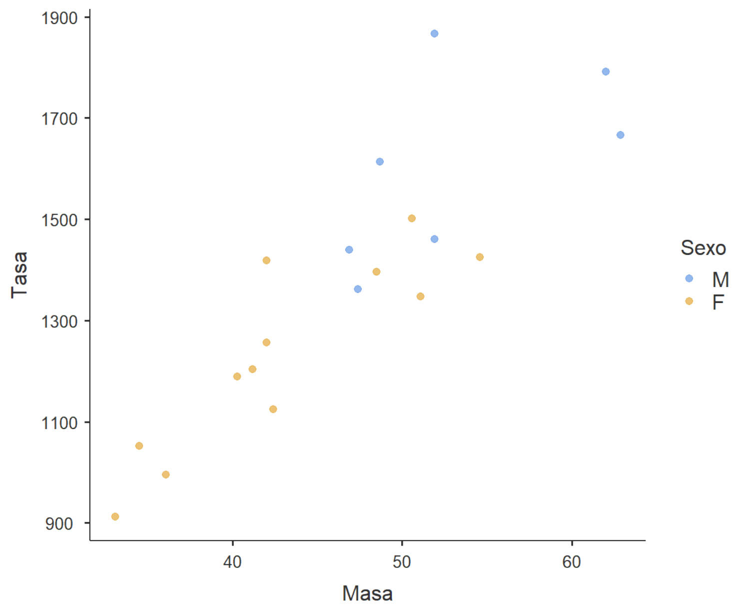
Utilizaremos los datos del fichero `metabolismo.omv`. Las variables corresponden a la tasa metabólica y la masa corporal magra de una muestra de 7 hombres y 12 mujeres. El objetivo en esta práctica es cuantificar la relación existente entre la masa corporal (que tomaremos como variable regresora) y la tasa metabólica (que tomaremos como variable respuesta). Por lo tanto, vamos a ajustar el modelo:

$$\texttt{Tasa} = \beta_0 + \beta_1 \texttt{Masa} + \epsilon.$$

Antes de obtener los resultados numéricos conviene representar el diagrama de dispersión de los datos para ver si podemos suponer que se cumplen algunas de las hipótesis habituales. Para confirmar otras hipótesis es necesario analizar los residuos del modelo. Recordemos que para representar los diagramas de dispersión es necesario instalar el módulo *scatr* y usar la opción



Representamos la masa en el eje de abscisas y la tasa en el de ordenadas. Parece que la relación es apropiadamente lineal aunque ligeramente heterocedástica:



Para obtener los estimadores de mínimos cuadrados de la pendiente y el término independiente y otros resultados relacionados con el modelo, la opción que tenemos que seleccionar es:

`Analyses` $\hookrightarrow$ `Regression` $\hookrightarrow$ `Linear regression...`

1. Como variable dependiente se debe elegir la variable respuesta. En el ejemplo, la variable `Tasa`.

2. Como variable independiente (*covariate*) se selecciona la variable regresora que en nuestro caso es `Masa`.

Se debe obtener el siguiente resultado:

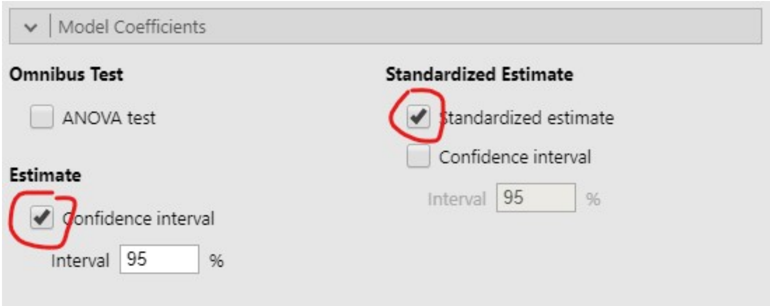
Model Fit Measures

Model	R	R^2
1	0.865	0.748

Model Coefficients

Predictor	Estimate	SE	t	p
Intercept	113.2	179.59	0.630	0.537
Masa	26.9	3.79	7.099	< .001

Para obtener más información sobre el ajuste del modelo, podemos marcar las opciones siguientes:



Con ello, se añadirán a la tabla anterior los intervalos de confianza para los coeficientes del modelo y los estimadores calculados a partir de las variables previamente estandarizadas:

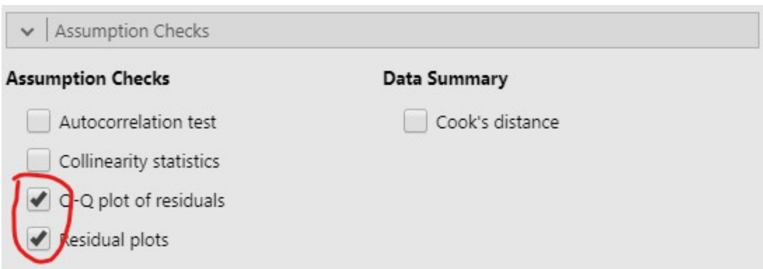
Model Fit Measures

Model	R	R^2
1	0.865	0.748

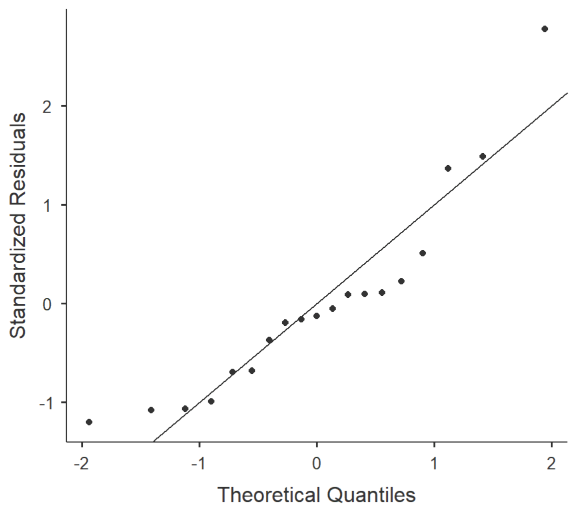
Model Coefficients

			95% Confidence Interval				
Predictor	Estimate	SE	Lower	Upper	t	p	Stand. Estimate
Intercept	113.2	179.59	−265.7	492.1	0.630	0.537	
Masa	26.9	3.79	18.9	34.9	7.099	< .001	0.865

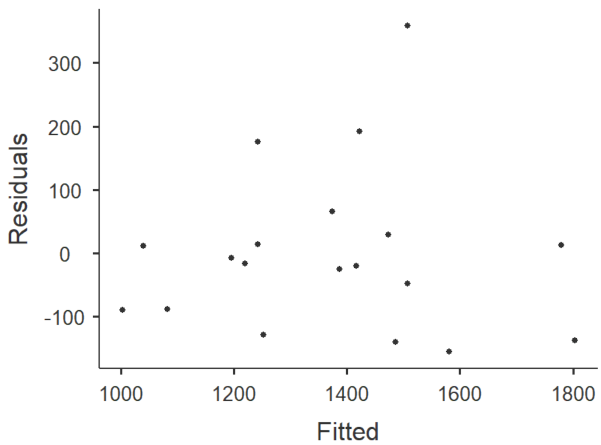
Finalmente, puede ser conveniente analizar algunos aspectos de los residuos para comprobar si las hipótesis habituales del modelo de regresión son aceptables. Para ello, marcamos las siguientes opciones



Este es el *QQ-plot* resultante en este caso. Corresponde a puntos razonablemente alineados con la excepción de algún dato ligeramente atípico. Parece aceptable la hipótesis de normalidad.



El siguiente gráfico corresponde a residuos frente a valores ajustados. No aparece ningún patrón claro, con la excepción de la ligera heterocedasticidad que ya habíamos observado.



Ejercicio 1

1. Identifica en los resultados anteriores el valor del coeficiente de correlación entre la tasa metabólica y la masa corporal.

2. Fíjate en la tabla que contiene los coeficientes estimados con las variables estandarizadas. ¿Por qué crees que el estimador del término independiente no aparece? ¿Por qué coincide el estimador de la pendiente con el coeficiente de correlación entre ambas variables?

3. Calcula un intervalo de confianza de nivel 99 % para la pendiente de la recta de regresión.

4. Queremos contrastar si los datos aportan suficiente evidencia para afirmar que la pendiente de la recta de regresión es positiva. ¿Qué contraste de hipótesis tenemos que llevar

a cabo?

(a) $\begin{cases} H_0 : \beta_1 = 0, \\ H_1 : \beta_1 \neq 0. \end{cases}$ (b) $\begin{cases} H_0 : \beta_1 \leq 0, \\ H_1 : \beta_1 > 0. \end{cases}$ (c) $\begin{cases} H_0 : \beta_1 \geq 0, \\ H_1 : \beta_1 < 0. \end{cases}$

¿Cuál es la conclusión del contraste para los niveles de significación usuales?

5. Queremos contrastar si los datos aportan suficiente evidencia para afirmar que la pendiente de la recta de regresión no es igual a 25. ¿Qué contraste de hipótesis tenemos que llevar a cabo?

   (a) $\begin{cases} H_0 : \beta_1 = 25, \\ H_1 : \beta_1 \neq 25. \end{cases}$ (b) $\begin{cases} H_0 : \beta_1 \leq 25, \\ H_1 : \beta_1 > 25. \end{cases}$ (c) $\begin{cases} H_0 : \beta_1 \geq 25, \\ H_1 : \beta_1 < 25. \end{cases}$

   ¿Permiten los datos anteriores afirmar a nivel $\alpha = 0{,}05$ que la pendiente de la recta no es igual a 25?

6. Si un individuo tiene una masa corporal igual a 50, utiliza el modelo ajustado para predecir su tasa metabólica.

7. Ajusta un modelo utilizando sólo los datos de los hombres, y otro utilizando sólo los datos de las mujeres.

   ¿Qué correlaciones hay en cada uno de los grupos entre las variables tasa metabólica y la masa corporal? ¿Qué grupo se ajusta mejor al modelo lineal?

   ¿Cuáles son las pendientes de la recta de mínimos cuadrados en cada uno de los grupos?

2. Análisis de la varianza unifactorial

En este apartado utilizaremos el fichero `dientes.omv`. Este fichero contiene 60 observaciones y 2 variables. La variable respuesta es la longitud de los dientes de varios conejos de indias (`longitud`) que han recibido tres dosis diferentes de vitamina C (`dosis`). El objetivo consiste en estudiar si la dosis de vitamina C tiene algún efecto sobre la longitud de los dientes.

Como paso previo a realizar los cálculos es necesario comprobar que la variable que va a hacer el papel de factor es de tipo ordinal o nominal. En este caso concreto hay que cambiar el tipo de la variable `dosis` puesto que en el fichero disponible aparece como variable continua.

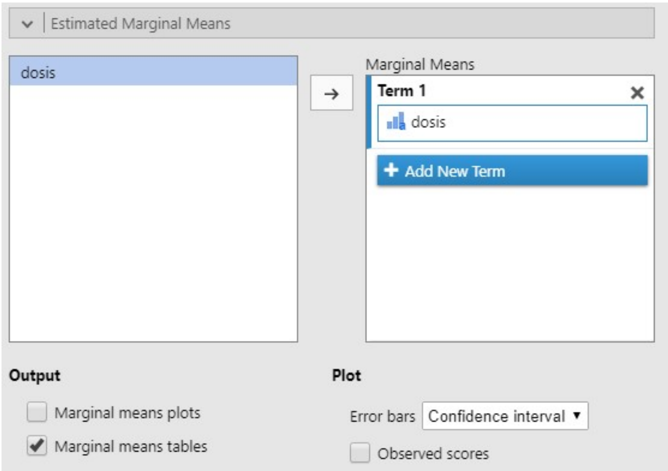
Para obtener la tabla ANOVA hay que elegir

`Analyses ↪ ANOVA ↪ ANOVA`

1. Como variable *dependiente* se debe elegir la variable respuesta. En el ejemplo, la variable `longitud`.

2. Se selecciona el factor que, en nuestro caso, es la variable `dosis`.

3. Para obtener las medias estimadas para cada grupo y los correspondientes intervalos de confianza se seleccionan las opciones siguientes:



Se debe obtener el siguiente resultado:

ANOVA

ANOVA

	Sum of Squares	df	Mean Square	F	p
dosis	2426	2	1213.2	67.4	< .001
Residuals	1026	57	18.0		

[3]

Estimated Marginal Means

dosis

Estimated Marginal Means - dosis

			95% Confidence Interval	
dosis	Mean	SE	Lower	Upper
0.5	10.6	0.949	8.71	12.5
1.0	19.7	0.949	17.84	21.6
2.0	26.1	0.949	24.20	28.0

[4]

Ejercicio 2

1. Sea μ_i la longitud media poblacional de los dientes de los conejos de indias correspondientes al nivel i del factor. (Es decir, μ_1 es la longitud media cuando la dosis de vitamina C es 0,50, μ_2 es la longitud media cuando la dosis de vitamina C es 1,00 y μ_3 es la longitud media cuando la dosis de vitamina C es 2,00) ¿Podemos rechazar a nivel $\alpha = 0{,}05$ la igualdad $\mu_1 = \mu_2 = \mu_3$? ¿Y a nivel $\alpha = 0{,}01$?

2. ¿Cuál es el estimador de la longitud media μ_2 de una población de conejos de indias que recibiera la segunda dosis (correspondiente al nivel 1.00)? ¿Y el error típico de este estimador?

3. ¿Por qué coinciden los errores típicos de los estimadores de μ_1, μ_2 y μ_3?

4. Usando los valores de los estadísticos descriptivos y el formulario de contrastes, contrasta a nivel $\alpha = 0{,}05$ la hipótesis nula $H_0 : \mu_2 \leq 18$ frente a la alternativa $H_1 : \mu_2 > 18$.

5. Utiliza los datos del fichero `garganta.omv` que contiene tres variables correspondientes a 35 pacientes que han sido sometidos a cirugía: la variable `D` corresponde a la duración en minutos de la cirugía; la variable `T` corresponde al medio para garantizar la respiración (`T=0` máscara laríngea, `T=1` tubo traqueal) y la variable `Y` corresponde a si el paciente experimentó dolor de garganta al despertar (`Y=0` no, `Y=1` sí).

    (a) Compara si hay diferencias significativas ($\alpha = 0{,}05$) en la duración media de la cirugía entre los pacientes que experimentaron dolor y los que no. Utiliza para ello un modelo de análisis de la varianza unifactorial.

    (b) Responde a la pregunta anterior utilizando un contraste t de Student para dos muestras independientes. Compara los resultados de ambos procedimientos y enumera las relaciones que ves entre ellos.